Automatic speech and language processing for precision medicine in Huntington's disease

Rachid Riad

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'École Normale Supérieure

Automatic speech and language processing for precision medicine in Huntington's disease

Soutenue par
Rachid RIAD
Le 1er Avril 2022

École doctorale n°540
Lettres, Arts, Sciences humaines et sociales

Spécialité
Sciences cognitives, neurosciences, psychologie

Composition du jury :

Corinne Fredouille Professeur des Universités Avignon Université, LIA	*Présidente du Jury*
Emily Mower Provost Associate Professor University of Michigan	*Rapporteur*
Anoopum Gupta Associate professor Massachusetts General Hospital	*Examinateur*
Joaquim Ferreira Professor University of Lisbon	*Examinateur*
Anne-Catherine Bachoud-Lévi UPEC, ENS, INSERM, IMRB	*Directrice de thèse*
Emmanuel Dupoux EHESS, ENS, INRIA	*Directeur de thèse*

1794
ENS

PSL

Abstract

Neurodegenerative diseases are a major social issue and public health priority worldwide. Huntington Disease (HD) is a rare disease of genetic origin that causes cognitive, behavioural and motor disorders due to brain lesions, in particular in the striatum. People with the genetic mutation of HD have a pre-symptomatic phase of several decades during which they have no neurological disorder before the symptomatic phase occurs. The symptoms of this disease have many implications in the life activities of the patient, with a gradual loss of autonomy, until the death of the patient. This makes HD a potential model of neurodegenerative diseases that could lead to the development of new clinical monitoring tools. The current medical monitoring in HD is expensive and requires the patient to travel regularly to the hospital, generating a significant human and financial burden. The purpose of this thesis is to develop and validate new computational methods for automatically monitoring Huntington's Disease individuals, thanks to the analysis of their spoken language productions. Spoken language production invokes various cognitive, social and motor skills, and its realisation is influenced by the mental state of the individual. Our hypothesis is that through the inspection of the produced speech and its content we can assess these different skills and states. To this date, the analysis of spoken language disorders in HD is only performed in a few clinical departments and specialised research teams, at a small scale without classic clinical validation. In addition, the potential of spoken language markers to predict the different symptoms in HD have not been explored.

Therefore in this thesis, we designed a comprehensive spoken language battery, along with a complete annotation protocol that is parsable by a computer program. This battery measures different parameters to obtain a wide clinical picture of spoken language in HD, that varies the linguistic target, the cognitive load, the emotional content, the topics and the materials of the discourse. To speed up the annotations protocol, we designed and developed open-source software to manage linguistic annotation campaigns. This allowed us to collect what is, to the best of our knowledge, the largest database of fine-grained annotated spoken language productions in HD, with 125 annotated interviews of 3 groups of individuals: healthy controls, premanifest individuals carrying the gene that causes HD and manifest HD at different stages. Besides, we also formalized and implemented the tracks of

communication introduced by H. Clark, which allow analyzing the use of spoken language in spontaneous exchanges for HD individuals. Then, to speed up and automate the annotation process, we developed and validated machine learning methods to recognise turn-takings and identify these tracks of communication directly from speech. Finally, thanks to this new database, we assessed the capabilities of spoken language markers to predict the different symptoms in HD. We especially found out that rhythm and articulatory markers extracted from tasks with a cognitive load can predict accurately the global, motor, functional and cognitive components of the disease. We additionally found significant correlations between silence statistics and the volume of the striatum, the neuro-anatomical hallmark of the disease progress. In spontaneous productions, we found that the ratio of tracks of communication was different between HD individuals and other groups. The primary track was diminished, the timing ratio of secondary presentation (filled pauses) also decreased and the timing of incidental elements (ex: vocal noises, audible respiration) greatly increased. We also proposed new methodologies to examine the emotional speech production in HD. Finally, we found out that the manifest individuals with HD have both vocal and linguistic impairments during emotional speech production.

Abstract (French)

Les maladies neurodégénératives sont un enjeu social majeur et une priorité de santé publique dans le monde entier. La maladie de Huntington (MH) est une maladie rare d'origine génétique qui provoque des troubles cognitifs, comportementaux et moteurs dus à des lésions cérébrales, notamment au niveau du striatum. Les personnes porteuses de la mutation génétique de la MH ont une phase pré-symptomatique de plusieurs décennies au cours de laquelle elles n'ont aucun trouble neurologique avant que la phase symptomatique n'apparaisse. Les symptômes de cette maladie ont de nombreuses implications dans le quotidien, avec une perte progressive d'autonomie jusqu'au décès du patient. Cela fait de la MH un modèle potentiel de maladies neurodégénératives qui pourrait conduire au développement de nouveaux outils de suivi clinique. Le suivi médical actuel pour la MH est onéreux et oblige le patient à se rendre régulièrement à l'hôpital, générant une charge humaine et financière importante. L'objectif de cette thèse est de développer et de valider de nouvelles méthodes computationnelles pour le suivi automatique des individus avec la MH, grâce à l'analyse de leurs productions langagières. En effet, la production du langage oral fait appel à diverses compétences cognitives, sociales et motrices, et sa

réalisation est influencée par l'état mental et neurologique. Notre hypothèse est qu'à travers l'inspection de la parole et de son contenu nous pouvons évaluer ces différentes compétences et traits. À ce jour, l'analyse des troubles de la parole et du langage pour la MH n'est pratiquée que dans quelques services cliniques et équipes spécialisées, à petite échelle. De plus, la capacité des marqueurs issus du langage oral à prédire les différents symptômes de la MH n'a pas été explorée.

Par conséquent, dans cette thèse, nous avons conçu une batterie complète de tâches qui testent plusieurs niveaux de production de la parole, ainsi qu'un protocole d'annotations complet qui reste analysable par un programme informatique. Cette batterie a été conçu pour obtenir un tableau clinique complet du langage parlé en MH, qui fait varier la cible linguistique, la charge cognitive, le contenu émotionnel, les sujets du discours. Pour accélérer le processus d'annotations, nous avons conçu et développé un logiciel open source pour gérer les campagnes d'annotations. Nous avons pu ainsi collecter la plus grande base de données de productions langagières, à ce jour, avec 125 entretiens annotés pour 3 groupes d'individus: des témoins sains, des individus porteurs du gène qui cause la MH mais sans symptômes cliniques et des individus symptomatiques avec la MH à différents stades. Par ailleurs, nous avons également formalisé et implémenté les voies de communication introduites par H. Clark, qui permettent d'analyser la parole dans des échanges spontanés. Ensuite, pour accélérer et automatiser les annotations, nous avons développé et validé des algorithmes d'apprentissage profond pour reconnaître les tours de parole lors des entretiens et reconnaître les voies de communication directement à partir de l'audio. Enfin, grâce à cette nouvelle base de données, nous avons évalué les capacités des marqueurs issus de la parole à prédire les différents symptômes de la MH. Nous avons notamment découvert que les marqueurs rythmiques et articulatoires, lors des tâches à charge cognitive plus élevée, pouvaient prédire les composantes globales, motrices, fonctionnelles et cognitives de la maladie et corrélaient avec le volume du striatum, la marque neurale de l'évolution de la maladie. Nous avons également proposé de nouvelles méthodologies pour examiner la production de la parole émotionnelle dans la MH. Nous avons ainsi découvert que les individus avec des symptoms cliniques de la MH ont à la fois des troubles vocaux et linguistiques lors de la production de la parole émotionnelle.

Acknowledgements

The good metaphor for this thesis is a "brain that is producing spoken language". I present the work of many and it was made only possible by many more. Talking engages so many parts of the brain, and my thesis would not have been possible without all of these people.

I would like to thank my PhD advisors, the *frontal cortex* of this thesis, Emmanuel Dupoux and Anne-Catherine Bachoud-Lévi, for their patience, unwavering support during all the steps of this thesis journey, and the scientific rigor they provided. It is in this part of the brain that composition of ideas happen and where long-term plans are made. I want to thank Anne-Catherine for opening me the door to Neurology, sharing her enthusiasm for language research and make me never forget to put patients at the centre. Besides, I want to especially thank Emmanuel for sharing his taste for research, cognitive science, modelling, language and interdisciplinary research, helping other. Emmanuel always gave me the chance to not feeling being judged when I did not know something, wrote something completely false, failed the submission of a PDF into a conference platform, . . . (this list is long).

I want to thank also the *cerebellum* of this thesis, Hadrien Titeux and Xuan Nga Cao, the part of the brain that accounts for most of neurons (80%) and that facilitates and make possible *all* the functions of the brain, especially help to time and coordinate speech production. Hadrien and Xuan Nga have been incredible mental-support and research team mates all along this thesis; and the work presented here is also theirs.

Then, I want to thank the *temporal lobe* of this thesis, all the people that know so many things and that always took the time to share with me their knowledge, offer me their time and support at any time: in Neurology, Marine Lunven, Laurent Cleret de Langavant, Laurie Lemoine; and in Machine Learning and Signal Processing, Julien Karadayi, Neil Zeghidour, Thomas Schatz. Special thanks to Marine and Laurie to always be there for me, the drinks, and all the fun in Bressanone.

I also want to thank the sensory parts of this thesis, the *ears* and the *eyes* of this thesis, the neuropsychologists that help me for the protocol, Justine Montillot, Jennifer Hamet Bagnou, Agnes Sliwinski, Nicolas Fraisse, and the many many speech pathologists that took part in the project.

Talking alone is way less fun than talking with other people, that is why I want to thank the scientific *interlocutors* of this thesis, Emily Mower Provost, Anoopum Gupta, Joaquim Ferreira, Corinne Fredouille, for their comments and the participation in the thesis committee.

I also want to thank the inner circles of scientific interlocutors, Frank Rudzizc who welcomed me in Canada, Matthew Perez from UMich, and also all the PhD students of CoML and NPI, Juliette Millet, Ronan Riochet, Robin Algayres, Mathieu Rita, Rahma Chaabouni, Marvin Lavechin, Maureen de Seyssel and Amin Gharbi. I admire their scientific curiosity and their own research subjects, and they helped me so much in so many ways: through friendly drinks, writing clubs, listening me venting, etc ... Besides I want to think the *glia* of this thesis, the other members of the CoML team, Catherine Urban, Nicolas Hamilakis, Mathieu Bernard, Manel Khentout, Marianne Metais the people that make possible everything in the lab and always eager to help. I also want to thank the interns who took part in our projects, Corentin Dancette, Cécile DiFolco, Elias Aouad, Charlotte Gallezot, Reda Arab, Remy Nguyen. I think they did not realize how much I learned from them, and pushed me to be scientifically so rigorous.

Talking is the cement of our *social world*, this feeling for belonging is what makes us the most human. That is why I want to thank my friends for all the fun, support, and joy they bring every time we gather: my friends from High School, Antoine, Paul, Bertrand, Jean-Vincent, my friends from 'prepa', Antoine, Léa, Joana, Diego, Georges, Vincent, Johan and Sylvain, and also my friends from school and San Francisco, Louis, Etienne, Victor, Martin, Corentin, Léo T, Félix, Ludovic, Loic, Maxime, Mathieu, Thibaut, Stéphanie, Julien, Stéphanie, Carole, Elise, Domitille, Gabriel, Adrien, Hernan, Louise, Léo B, Gweltaz. I don't know how they manage to support my 'talk-arguing' all the time.

The question of language acquisition is definitely not solved, but I definitely want to thank who made who I am, the people who built and still builds the *best environment* for growing up, my family. I want to deeply thank my parents and my brother for their caring, all the love they share with me, all the support to believe in myself, that nothing is impossible thanks to them. I also want to thank Pauline, whom with I shared all these years, all the joy, and for all the support throughout the years. The true heroes of academia are the partners with stable jobs. I cannot thank you enough Pauline.

Finally, I will never forget the *topic* of the conversation of this thesis, participants affected with Huntington's Disease and their relatives. They all always welcomed

me during their visits at the hospital, always engaged to participate in research, and they remind us every day why medical research is so important. Thank you.

List of acronyms

HD: Huntington's Disease, or individuals with at least 36 CAG repeats on the mutant Huntingtin gene, with clinical symptoms.

HTT: portion of the gene 4 coding for the Huntingtin protein.

CAG: Cytosine – Adenine – Guanine

mHTT: mutant Huntingtin protein.

preHD: Individuals with at least 36 CAG repeats on the mutant Huntingtin gene, without clinical symptoms.

HC: Healthy controls, neurotypical individuals.

UHDRS: Unified Huntington's Disease Rating Scale.

TMS: Total Motor Score.

PBA: Problem Behaviours Assessment

TFC: Total Functional Capacity.

IS: Independence scale.

SDMT: Symbol Digital Modality Test.

SW: Word component of the Stroop test.

SC: Color component of the Stroop test.

SI: Interference component of the Stroop test.

SDMT: Symbol Digital Modality Test.

cUHDRS: composite score the UHDRS.

PD: Parkinson's Disease.

NDD: Neurodegenerative disease.

BasalVoice: Database with HD speech collected in this thesis.

SpontaneousCHAT: Annotation protocol introduced in this thesis, derived from the CHAT protocol.

DFT: Discrete Fourier Transform.

STFT: Short-time Fourier Transform.

MPS: Modulation Power Spectrum.

ASR: Automatic Speech Recognition.

MFCC: Mel-Frequency Cepstral Coefficients.

HNR: Harmonics-to-Noise Ratio.

ML: Machine Learning.

PCA: Principal Component Analysis.

VSA: Vowel Space Area.

VAI: Vowel Articulation Index.

MPT: Maximum Phonation Time.

MPT_{VB}: Maximum Phonation Time until first voice break.

DDK: Diadochokinesis.

SVM: Support-Vector-Machine

NB: These acronyms are always first defined in the text before being used as abbreviations.

Table of contents

List of figures

List of tables

1 Introduction

> *The clinic, the laboratory, the ward are all designed to restrain and focus behavior, if not indeed to exclude it altogether. They are for a systematic and scientific neurology, reduced to fixed tests and tasks, not for an open, naturalistic neurology. For this one must see the patient unselfconscious, unobserved, in the real world, wholly given over to the spur and play of every impulse, and one must oneself, the observer, be unobserved.*
>
> **— Oliver Sacks**
> The man who mistook his wife for a hat

Neurodegenerative diseases (NDDs) are a major social issue and a public health priority worldwide (World-Health-Organization, 2007). These diseases are incurable and provokes the progressive degeneration and/or the death of nerve cells. In Europe, they reach about 10 million people at a cost of US $ 263 billion a year (DiLuca & Olesen, 2014). NDDs are responsible for cognitive, psycho-behavioural and motor disorders. These disorders contribute to a loss of autonomy for the patient who becomes dependent on those around him and is often forced to live in a clinical institution. The progressive nature of the symptoms in NDDs requires neurological, psychiatric and cognitive expertise, and frequent physical visits to the hospital. This frequent follow-up represent a human and financial burden that make it difficult to evaluate the efficiency of new therapies. The burden of NDDs continue to increase as population is growing and aging, as the prevalence of NDDs steeply increases with age. Governments will face increasing demands for evaluation, treatments and therapeutics (Feigin et al., 2019).

Among NDDs, Huntington's Disease (HD) is often considered as a *model* of NDDs that could lead to the development of new clinical monitoring tools. Indeed, HD provokes a large spectrum of behaviors and the deterministic genetic origin of the disease allows to follow-up individuals even before the motor clinical onset. HD is a rare NDD that is caused by a inherited cytosine-adenine-guanine (CAG) repeat

expansion in the huntingtin gene (HTT) (Gusella et al., 1983).The associated mutant HTT protein (mHTT) that results from the CAG repeat leads to neuronal cell death. In particular, the medium spiny neurons in the striatum are particularly vulnerable to the presence of the mHTT, even though significant neuronal dysfunction also occurs in the cerebral cortex. HD is characterized by a large spectrum of motor, psychiatric and cognitive symptoms leading to a progressive functional impairment. The associated mutant HTT protein (mHTT) that results from the CAG repeat leads to neuronal cell death. In particular, the medium spiny neurons in the striatum are particularly vulnerable to the presence of the mHTT, even though significant neuronal dysfunction also occurs in the cerebral cortex. HD occurs clinically in middle-aged patients (around 40 years of age) and progressively worsens in 15 to 20 years after motor clinical onset, until death. People with the genetic mutation of HD have a pre-symptomatic phase of several decades during which they have no neurological disorder, before the symptomatic phase occurs, especially before the official criterion of disease, through observation of motor disorders. The symptoms of this disease have many implications in the daily life activities of the patient and his caregivers, with a gradual loss of autonomy (Marder et al., 2000), and a gradual degradation of social interactions. To date, there is no cure for Huntington's disease. However, current innovative therapies such as neural grafts (Bachoud-Lévi et al., 2000) and therapeutic methods for lowering huntingtin expression are being tested on patients (Tabrizi et al., 2019).

Since the discovery of the gene causing the disease (Gusella et al., 1983), various biomarkers for the monitoring of HD have been proposed: biological markers such as the cerebrospinal fluid (Scahill et al., 2020), imaging to measure brain volumes (Tabrizi et al., 2013), or with digitized tests (Stout et al., 2014; Lunven et al., 2021). However, these methods are still expensive, complex to implement and require the presence of experts. In addition, due to the multiplicity of symptoms, one single marker can not reflect all dimensions in real life. On the first hand, the ideal biomarker in HD (and in NDDs in general) follow-up should have predictive capabilities at the individual level to make diagnostic of the disease, to reflect the different symptoms at a given time, and potentially have prognostic capabilities, i.e. predict the likely future course of these symptoms. On the other hand, the ideal biomarker should remain inexpensive, non-invasive, safe, non-exhausting, insensitive to demographics unrelated to the pathology, and potentially invisible to the participant to not influence behavior.

One of the many symptoms of the disease is the perturbation of production of *spoken language*. The gradual degradation of spoken language, until the end of the patient's life, is an important factor in the social disintegration and isolation of patients. The

spoken language production invokes various cognitive, social, and motor skills, and its realisation is greatly influenced by the mental state and neurological condition of the individual.

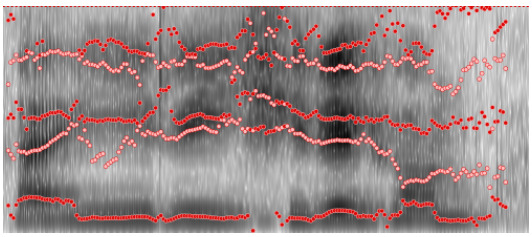
The main goal of this PhD thesis is to investigate *the automatic assessment of motor, cognitive and functional statuses in HD through the analyses of spoken language production*. In this thesis, we hypothesize that through the inspection of the speech and its content we can access the inventory of these different skills. Indeed, several studies have observed the alteration of the different processes involved in spoken language production in HD: phonatory dysfunctions (Rusz, Saft, et al., 2014; Rusz, Klempir, Baborova, et al., 2013; Romana et al., 2020), timing (Perez et al., 2018; Vogel et al., 2012), morphology and rule applications (Ullman et al., 1997; Teichmann et al., 2005), semantic diversity and syntactic complexity (Hinzen et al., 2017). This hypothesis is reinforced by the recruitment of the striatum at different step of spoken language production (Friederici, 2017; Guenther, 2016; Jacquemot & Bachoud-Lévi, 2021). Past studies have evaluated extensively the differences between the speech of HD and healthy controls. To obtain usable biomarkers from speech and language, it is necessary to go one step beyond and assess the clinical validity and reliability of these markers as predictors of the different symptoms in HD. In this thesis, we extended the validity of spoken language markers as biomarkers by trying to predict clinical scores in HD.

Spoken language markers also potentially meet the requirements to be a non-invasive, economical, non-exhausting marker, that remain short to collect. Thanks to recent technological developments, the acquisition of acoustic data is much easier and more affordable than the other types of biological, imaging or cognitive, markers mentioned above. In addition, the acquisition of speech data requires less effort for patients and clinical practitioners, and even could be collected remotely.

One may ask why have these analyses not already been deployed in clinical practice and why did we introduce the term 'automatic' in our thesis goal statement? Speech analyses require long and tedious listening by experts in linguistics and the use of complex computer tools for the annotation of recorded speech. These expert-based solutions can not be scaled to real-life conditions because the amount of data is too large. Moreover, the current understanding of the difference between typical speech and pathological speech is limited by the knowledge gained from small-group clinical studies. Going from the lab to the clinical application would therefore require to build robust speech processing pipelines that automatically extract meaningful insights from large naturalistic datasets (Lahiri et al., 2020) that are easy to obtain from individuals with speech and language impairments.

(a) Waveform of the speech utterance.

(b) Excerpt waveform of the vowel /i/ in the utterance.



(c) Time-frequency representation (Spectrogram) of the speech utterance along the fundamental frequency and the formants.

Fig. 1.1.: The waveform of a speech utterance pronounced by a participant in our thesis saying "(il) y a les gilets jaunes quoi" (en: (there) are the yellow jackets duh).

Yet, the automatic analyses of speech and language from audio recording to predict HD symptoms still pose significant scientific and engineering challenges. Even though, audio is easy to collect, its automatic processing is not. 15 minutes of audio sampled at 44.1kHZ represents more than 39 millions of audio data points. Figure 1.1 shows the correspondence between a speech waveform of an utterance, an excerpt of the speech waveform representing a single vowel (/i/) from this utterance, and the time domain representation (its magnitude) of the utterance. How can we detect automatically the clinical information hidden in this type of data?

This thesis addresses a number of related research questions: How do we automatically map raw audio data to interpretable linguistic markers? What kind of interview protocol can elicit maximally informative speech from participants? What are the speech and linguistic markers that are clinically relevant? How do we make sure that the clinical predictions are reliable and valid enough to inform decision making?

These different sub-questions are reflected in the construction and the chapters of this thesis. In the following sections of the introductory chapter 1, we first review the current state of the scientific knowledge on the speaking ability of humans. Then, we present recent work in computational models and technologies applied to pathological speech, how spoken language collection represents an opportunity to access the mental and medical condition of individuals. Finally, we briefly review some key results in the study of HD. We focused our reviews on the main points and concepts to validate speech biomarkers in HD relevant for this thesis.

In Chapter 2, we describe the main database used in this thesis, including the development rationale, BasalVoice, its collection, the annotation protocol and its parsing. In Chapter 3, we describe the development and evaluation of speech algorithms to detect turn-takings and tracks of communication in disordered speech. Then, we summarize the different contributions on the evaluation of speech and linguistic markers as biomarkers of motor, cognitive, psychiatric and functional symptoms in HD in Chapter 4. Finally, we conclude this thesis with a general discussion, limitations, and potential future avenues of work in Chapter 5.

This thesis is composed of several research publications that we included in the format of each venue. We introduce these papers in the adequate chapter and we will introduce them how they fit in the whole picture of this thesis. You can see below the publications that are included in this thesis and the other contributions included in the appendix.

Publications

The work carried during this thesis lead to the following publications:

- Titeux*, Riad*, Cao, Hamilakis, Madden, Cristia, Bachoud-Lévi, and Dupoux, "Seshat: A tool for managing and verifying annotation campaigns of audio data", *: equal contribution, LREC 2020
- Riad, Bachoud-Lévi, Rudzicz, and Dupoux, "Identification of primary and collateral tracks in stuttered speech", LREC 2020
- Riad, Titeux, Lemoine, Montillot, Bagnou, Cao, Dupoux, and Bachoud-Lévi, "Vocal markers from sustained phonation in Huntington's Disease", INTERSPEECH 2020
- Titeux and Riad, "pygamma-agreement: Gamma measure for inter/intra-annotator agreement in Python" Journal of Open-Source Software (JOSS) 2021
- Riad*, Titeux*, Lemoine, Montillot, Sliwinski, Bagnou, Cao, Bachoud-Lévi, and Dupoux, "A comparison study on patient-psychologist voice diarization", *: equal contribution, SLPAT 2022
- Riad, Lunven, Titeux, Cao, Bagnou, Lemoine, Montillot, Sliwinski, Youssov, de Langavant, Dupoux, and Bachoud-Lévi, "Predicting clinical scores in Huntington's Disease: a lightweight speech test", Journal of Neurology (JOON) 2022

- Gallezot*, Riad*, Titeux, Cao, Bagnou, Lemoine, Montillot, Sliwinski, Youssov, de Langavant, Dupoux, and Bachoud-Lévi, “Vocal and linguistic expression impairments of emotions in Huntington’s disease”, *: equal contribution, Cortex 2022

Other contributions, which we are in the appendix, were done during the same period.

- Riad, Karadayi, Bachoud-Lévi, and Dupoux, “Learning spectro-temporal representations of complex sounds with parameterized neural networks”, The Journal of the Acoustical Society of America (JASA) 2021

- Riad, Teboul, Grangier, and Zeghidour, “Learning strides in convolutional neural networks”, Oral and outstanding paper award at ICLR 2022

1.1 Speaking

Speaking is one of our most complex skill (Levelt, 1993). Speaking is unique to our species, especially our capability to generate an infinite amount of sentences that remain meaningful based on a limited number of elements (Friederici, 2017; Hauser et al., 2002). We use this skill to do a number of things, such as chitchatting, coordinate actions with someone else, teach, debate over politics or pandemic, gossip, share our emotions (Tomasello, 2010; Clark, 1996; Stokoe, 2018; Scherer, 1995). Neurotypical children manage to pick up the language that is in use in the social environment they are immersed into (Mehler & Dupoux, 2002). We could go as far as to that speaking ability is in constant evolution through lifespan. Adults that already master their language can easily pick up new words that emerge and expand their lexicon.

The scientific inquiry of the faculty of speaking require a large interdisciplinary research task force, with methods from social science (Clark, 1996), acoustics (Johnson, 2004), cognitive science (Levelt, 1993), neuroscience (Friederici, 2017; Guenther, 2016), neuropsychology (Shallice, 1988) and of course, computer science (Jurafsky & Martin, 2000). In Section 1.1.1, we briefly overview the main processing components that transform our intentions into fluent articulated speech through sound. Then we introduce in Section 1.1.2 the specifics of naturalistic speech.

1.1.1 Model of spoken language production

In this section, we introduce the main concepts and their organization that transform our intentions into speech. We focus on three models that are not exclusive, the global model of spoken language production of Levelt (Levelt, 1993), and two models of Guenther for articulation and coordination, DIVA and GODIVA, to articulate speech (Guenther, 2016). In Figure 1.2, we reproduced the bluebrint that illustrates the main processing components to generate fluent speech introduced by Levelt (1993). Each component has specific type of inputs and specific type of outputs that are fed to other components.

First, the production of speech involves the conception of an intention, a proper selection of information to transmit to the interlocutor(s). This conception module is referred as the *Conceptualizer*. Speaker needs to order the information, keeping track into memory what has been said and by whom. This demands an important attention to personal productions, monitor what has been actually pronounced. The output of the Conceptualizer is the *preverbal message*. To generate such message, speaker need *procedural knowledge* to properly coordinate actions/intentions to a specific message and check if a message is correct. In addition, speaker needs *declarative knowledge* to properly generate messages, in the form of boolean values attributed to assertions such as 'Paris is a great place'. Besides, there is a requirement to track the situation and the current state of the conversation to avoid redundancy and to be efficient.

This pre-verbal message is also the input of the *Formulator*, the information processing system that transform this pre-verbal message into a phonetic plan. The Formulator proceeds into two main steps. The first step is the *grammatical encoding* of the message. There is need to access the relevant lemma information and the adequate grammar construction for the message. A given word has specific lemma information containing both its *meaning* and its individuals rules for combinations. Based on the relevant retrieval of lemma information, the grammatical encoder produced a *surface structure* , an ordered sequence of strings of words that are grouped in sentences and sub-sentences. The grammar encoding is subdivided into *syntax*, *morphology* and *phonology*. Syntax describes the process that combines words into sentences/utterances. Morphology describes rules for the combination of words and meaningful word parts (morphemes) to get larger words ('walk'+'ed'='walked'). Finally, phonology depicts the sound organization of a language and the rules to combine sounds to form words. This results of these processes are now stored in what Levelt (1993) called a *syntactic buffer*. This syntactic buffer is read out by the *Phonological encoder* to get a phonetic/articulatory plan for each word. This plan

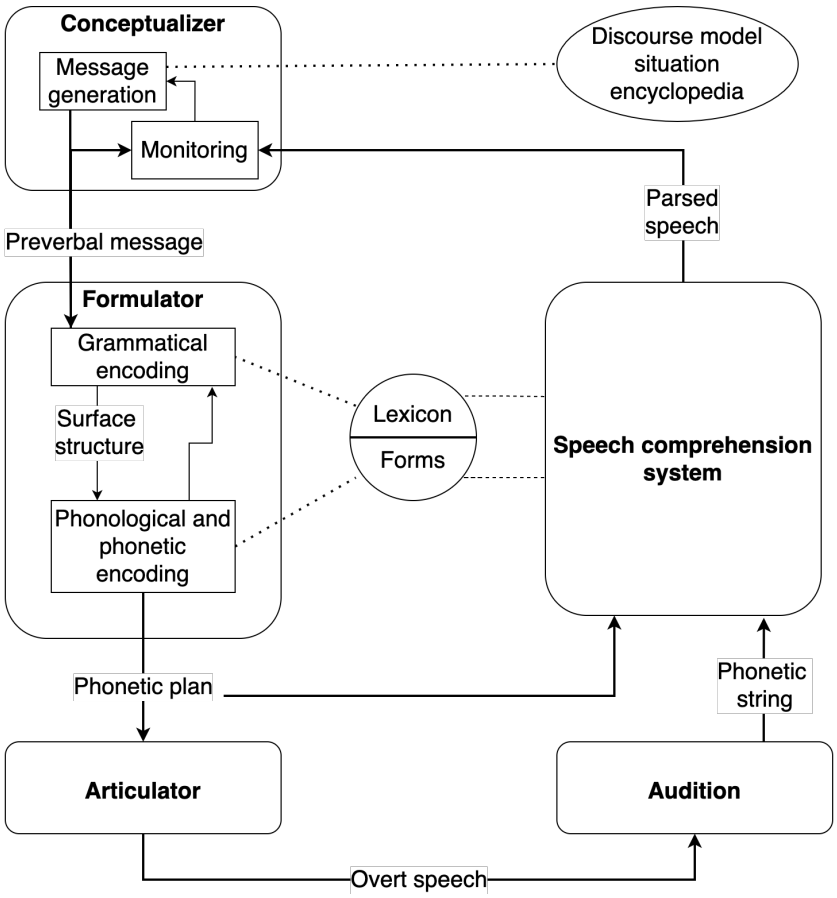


Fig. 1.2.: Schematic Levelt's model of spoken language production (Levelt, 1993).

becomes the next input to the *Articulator*, the next processing unit to finally obtain overt speech.

Articulation is the mechanism that transforms this phonetic plan into a sequence of coordinated movements of the *respiratory*, *phonatory* and *supralaryngeal* sub-systems. The articulation mechanisms have been described and summarized in great details by Guenther (2016). Guenther (2016) tried to specify the computations but also locate which part of the brain is involved for specific component. Based, on the phonetic plan, the Articulator need to command muscles movements to be perfectly timed, and to aim for the good acoustic adequate targets that represent each phones. This planning, chunking and timing of phonetic plan had been specified in the GODIVA model from (Guenther, 2016) (See Figure 1.3). Now, based on this information, commands needs to be executed to obtain the vibrations of the air to transmit speech. Overt speech production is surely the most complicated motor skill performed by human. There are approximately 100 muscles that are involved in the *respiratory*, *phonatory* and *supralaryngeal* sub-systems, each one contain approximately 100

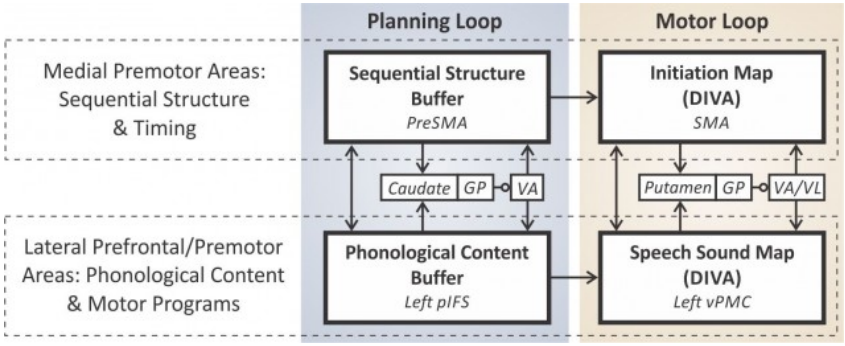


Fig. 1.3.: Schematic neurobiological GODIVA's model of multisyllabic planning, timing, and coordination speech production (Guenther, 2016). Image is extracted from (Guenther, 2016). Putamen and Caudate are two of the main structures of the striatum.

motor commands to pilot (Guenther, 2016). This means during the pronunciation of a word lasting 1 second, 10000 commands need to be generated and transmitted in less than 1 second to motor neurons to obtain the good acoustic target. To do so, it is required to inverse the required spatial movements from these planned acoustic targets (*inverse kinematics*), then derive the muscle commands to have these spatial movements (*inverse dynamics*). A model has been proposed into computational terms and neural localization in the DIVA model from (Guenther, 2016) (See Figure 1.4).

So far, we supposed that everything goes well when we transform our intentions into overt speech. Yet, troubles can occur at every step we mentioned. Speakers needs to attend to their speech thank to their auditory and speech comprehension systems to check for correctness with initial intentions. In addition, to the auditory targets, speakers monitor the somatosensory targets for a given articulation (See blue panel in Figure 1.4). Speakers also have access and correct the internal intermediate outputs of the processing components: the pre-verbal message, the surface structure and the phonetic plan. When the speaker spots serious errors, he can decide to halt formulation, pause or add an additional message.

1.1.2 Naturalistic conversational speech

In previous section, speaker's information processing was described in a standalone fashion. But it is important to consider spoken language production in the more classic canonical ecological setting, with the speaker's involvement in conversations (Clark, 1996). Levelt's and Guenther's processing components and their links only depict an isolated speaker. Yet, conversing, as argumented by Clark (1996),

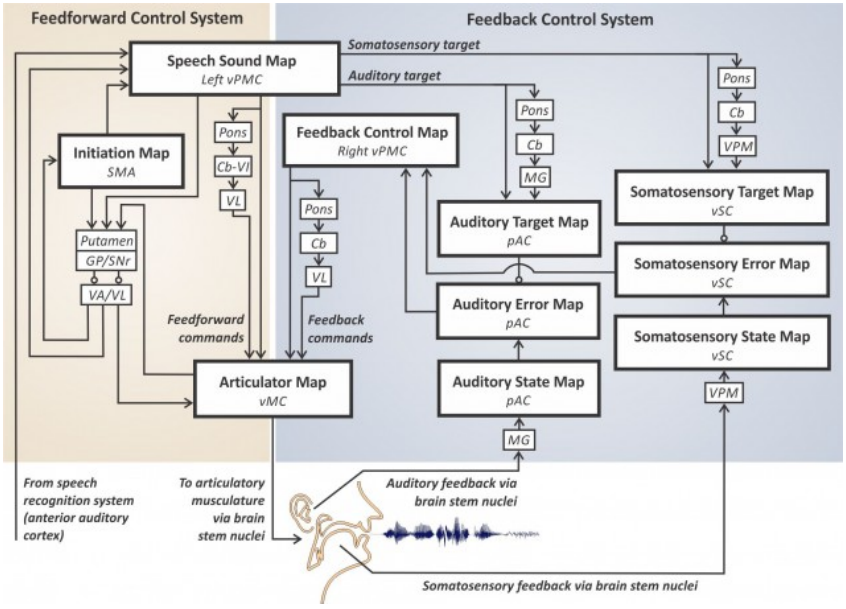


Fig. 1.4.: Schematic neurobiological DIVA's model of speech sound production (Guenther, 2016). Image is extracted from (Guenther, 2016). Putamen and Caudate are two of the main structures of the striatum.

is a joint action. We mentioned the fact that speaker self-monitor their speech productions, but speaker also monitor the comprehension of their interlocutors and adapts their productions so that conversations unroll smoothly. However, we did not mention *how* these interruptions occur and what are the strategies *within* an utterance to correct spoken language productions (We do not review cross-turn strategies and elements in this section and do not analyze them in this thesis). Two main descriptions/terminologies have been proposed concerning these aspects of spontaneous speech. On the first hand, Shriberg (1994) proposed to describe *how* an interruption of flow occurs. Especially, the description focused on how different vocalization are used for suspension (Interregnum), pre-empted (Reparandum) for comprehension, the specific moment of suspension (Interruption point) and the corrections of the errors (Repair). On the other hand, Clark (1996) proposed a segmentation of vocalizations into *tracks of communication* based on their *roles for transmitting signal*. We will come back in greater details to the Clark's theory of tracks of communication in Section 2.3. We illustrate these theories of interruption of speech for the same utterance in Figure 1.5 and Figure 1.6.

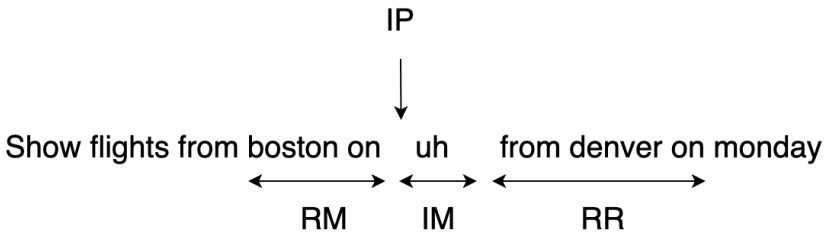


Fig. 1.5.: Terminology for regions involved in the disruption of flow, as introduced by Shriberg (1994). IP: Interruption point, RM: Reparandum, IM: Interregnum, RR: Repair

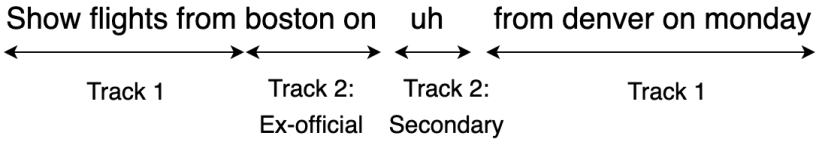


Fig. 1.6.: Terminology for the different regions in an utterance, as introduced by Clark (1996). Track 1 refers to the primary track of communication. Track 2 encompasses several elements: incidental elements (ex: audible respiration, vocal noises), secondary presentation

1.2 Dealing automatically with pathological spoken language

In Section 1.2.1, we review the basics of statistics and machine learning to validate biomarkers in HD. Then, in Section 1.2.2, we examine the algorithms and methods to extract speech information from sounds. Finally, we briefly describe how speech processing methods are used to obtain a window into the mind (Picard, 2000; Pinker, 2007), i.e. getting traits and states of individuals from speech.

1.2.1 Statistics or Machine learning?

Statistics and Machine Learning (ML) are methodologies to ask different types of questions. On the first hand, statistical methods probes pre-defined single markers and analyses are usually performed at the group level, comparing the average person of a given group to the average person of another group. Inferential or Bayesian statistical methodologies usually answer questions at the group level, through the probability that an observed difference could have occurred just by random chance (the p-value). Currently, null hypothesis significance testing allows to validate rigorously methods that are transferred into the neurological practice (Lunven et al., 2021; Cavanna et al., 2008; Gramunt et al., 2016). From a clinical perspective,

these methods allow to augment our current knowledge on specific diseases. To compare groups, researchers can make assumption on the generative process of the data. For instance, group difference between unpaired data can be tested thanks to the Student's t-test if the variance of the two groups are equal and each group data follows a normal distribution. These methodologies are referred as *parametric* tests as they make assumptions on the parameters that generate the data. On the other hand, researchers can choose to use *non-parametric* tests, that make no assumptions on the data and the underlying distribution. In this thesis, we used parametric tests when conditions were met, otherwise we focused on non-parametric tests to obtain more robustness in our conclusions.

In addition, statistical methodologies allow researchers to accumulate evidence in a distributed fashion, thanks to meta-analyses (Brockwell & Gordon, 2001). This means that, even though a study is under-powered, this allows to contribute to the falsification of hypotheses and/or reinforce the value of biomarkers in neurology and in medicine in general. Yet, such statistical approach requires to predict and guess which variable is of interest beforehand. One way to circumvent this issue, is to go through an *exploration* phase on a separate cohort, and identify the best variable candidate. In addition, to obtain a perfect single clinical biomarker in HD, even in NDDs in general seems utopian. Yet, it is more likely that different biomarkers measured at the same point in time with different information can potentially be more useful.

Thus if researchers want to test a large number of these biomarkers, or novel ways to combine them, this comes at the cost to increase the dimensionality of the data and the number of hypotheses to be tested. In limited data setting such as clinical trials in neurology, the *multi-comparison* issue increases the uncertainty of the findings (Reshef et al., 2016). Classical statistical methodologies were designed for data with a limited number of variables and small sample sizes in comparison to ML nowadays standards (Bzdok & Ioannidis, 2019). NDDs present a large variability at the individual level for the expression of the symptoms.

On the other hand, current research in Neurology does not escape the deluge of ML methodologies (Bzdok & Ioannidis, 2019; Myszczynska et al., 2020). Their uses start to be promising for diagnoses, prognoses and optimal treatment regime selection (Perez et al., 2018; Dwyer et al., 2018; Wilkinson et al., 2020; Varoquaux, 2018; Koval et al., 2021; Gajos et al., 2020). The objective of ML is to *generalize* from experience to a new unseen set of data, the test data. ML offers a new way to make sense of high-dimensional data, and select data for the objective at hand. From a clinical perspective, this allows to obtain a new tool that could be useful

for clinicians in the daily practice. ML has different potential settings, but in the scope of this thesis, we focused only on the supervised learning setting to predict clinical status and clinical scores. The supervised learning approach in ML offers to model the transformation of input data x into given and pre-defined label y. This label y can be categorical (classification) or numerical (regression). In the setting of biomarkers for NDDs, estimating the link between proposed quantitative biomarkers (See section 1.15) and the clinical status of individuals, is the way to answer the following question:

> Can I predict the clinical status of an individual based on objective measured biomarkers?

For instance, in one of our study, the goal will be to distinguish healthy controls (HC), preHD and HD based on the vowel /a/ features (See Section 4.1.1). Solving a task in machine learning is obtained through the solving of an optimization problem (Boyd et al., 2004) based on labelled training data $\left\{(x_i, y_i) \in \mathbb{R}^K \times \mathbb{R}, i = 1 \cdots N\right\}$. Thus, with pre-defined set of functions $\mathcal{H}$ parametrized by θ, the goal is to find the best f_θ that minimizes an objective function J based on input training data. To avoid overfitting and memorize the training data, the objective function is usually combined with a regularization function R. This can be formalized as:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} J\left(f_\theta, x_i, y_i\right) + \lambda \mathrm{R}(\theta) \tag{1.1}$$

The choice of the features x to use and the modelling between x and y, through the choice space of functions $\mathcal{H}$, are the core and biggest challenges. Machine Learning does not escape an exploration phase. During this phase, practitioners iterate over the feature and model choices. These findings are validated through replication of other groups, and accelerated thanks open benchmarks in Machine Learning.

In the case of validation of biomarkers in NDDs, this is currently unfeasible due to the sensitive nature of clinical data. In addition, reviewers sometimes prioritise improvements over benchmarks, which potentially affects the published type of research and ideas being shared. Researchers potentially overfit the training data to boost the test performance, especially in setting where the test data is small.

This specific application of machine learning in NDDs calls for new approach to safeguard ourselves against these problems. One approach would be that the validation of biomarker is restrained to specific space of modelling. Another way to

protect against inflated results is the *pre-registration* of methods before analyses. We argue for that machine learning methods and the features used for prediction should be pre-defined. At time of writing, this approach to the scientific process is gaining momentum across different sciences (Nosek et al., 2018). Despite the complexity during the conception of pre-registration without analyzing any validation data, pre-registration studies are more reproducible and worthwhile for progresses (Nosek et al., 2019). Machine learning is not escaping this new avenue with the organization of workshops in this theme https://preregister.science/.

We used this approach of pre-registration in two of our studies in this thesis to validate spoken language as markers or investigate emotional speech impairments in HD (Riad, Lunven, et al., 2022; Gallezot* et al., 2022).

1.2.2 Speech processing

In the section 1.1, we looked into how human generate overt speech through sounds. In practice, sounds are vibrations that propagate. These vibrations are captured through microphones that sample pressure, to yield a *discrete* representation of the sounds, the audio waveform. We illustrated in Figure 1.1a the waveform of an utterance. Audio scenes can be captured through a single microphone, *single-channel signal*, or multiple ones, *multi-channel signal*. In this thesis, we focused only on the single-channel treatment of audio.

Usually, audio signals are sampled at 16kHz or 44.1kHz with today's microphones, and each data points encodes the intensity of the waveform, with a resolution of 16 bits. Speech events vary in duration; phones ($\sim 80ms$), words ($\sim 300ms$), sentences ($\sim 2-3s$); thus there is need of a compression and selection of the waveform information to capture these events.

The frequency content of audio signal is the main avenue to do so. Harmonic analysis is used to transform a given signal in its frequency equivalent, notably thanks to the Discrete Fourier Transform (DFT). The DFT is the projection of the signal onto the basis of complex exponential that are linearly spaced in frequency. Thus, the DFT $\hat{x}[k]$ of an audio signal $x[n]$ of size N is given by:

$$\hat{x}[k] = \sum_{n=0}^{N-1} x[n] \exp\left(\frac{-i2\pi kn}{N}\right) \tag{1.2}$$

However, most audio events are short in time and their frequency content change over time. That is why the most accepted way to represent the transformation of the audio signal from the time domain into a *time-frequency representation* $X(m,k)$. The classic time-frequency representation is the discretized version of the short-time Fourier transform (STFT):

$$X(m,k) = \sum_{n=0}^{N} x[n]w[n-m] \exp\left(\frac{-i2\pi kn}{N}\right) \tag{1.3}$$

w represents the analysis window, usually $25ms$ in speech processing, and this window is usually computed every $10ms$ to form the time-frequency representation, the STFT. The STFT can be decomposed into magnitude and phase components, they can be intuitively seen as *how much* a frequency component is present in the audio signal at a given point in time, and *where* these components are placed relatively.

The distribution of energy in these frequency bands is not random, and our ear are adapted to this non-uniform distribution, with a limited range of hearing from 20Hz to 20kHz. Usually, this is the Mel Scale that is used to mimic the non-linear human hearing perception, with more emphasis on lower frequencies (Stevens et al., 1937).

Therefore, to obtain a closer representation to human hearing that is more compact, the magnitude is convoluted along the frequency axis with triangle filters (ψ_λ) that are equally spaced on the Mel Scale, to obtain the mel-filterbanks. In order to mimic the compression in terms of volume, the logarithm is also computed afterwards, to obtain the *log-mel Filterbanks*

$$Y(\lambda,k) = \log_{10}\left(|X(.,k)|^2 * \psi_\lambda\right) \tag{1.4}$$

Slow spectral and temporal modulation built on top of the STFT or the log-mel Filterbanks have been shown in psychophysical tests to be relevant for different human behaviors, emotions and speech intelligibility (Elliott & Theunissen, 2009; Elhilali et al., 2003; Edraki et al., 2019; Arnal et al., 2015), and boost performance for speech processing algorithms in noisy environments (Mesgarani et al., 2006; Chang & Morgan, 2014; Vuong et al., 2020). We illustrate spectro-temporal information that is computed through the Modulation Power Spectrum (MPS) and its relevance for speech content and the information about the speaker based on (Arnal et al., 2015). To put it in other terms, the MPS is basically the 2D Fourier analysis applied directly to the STFT or the log-mel Filterbanks, and only the magnitude is used.

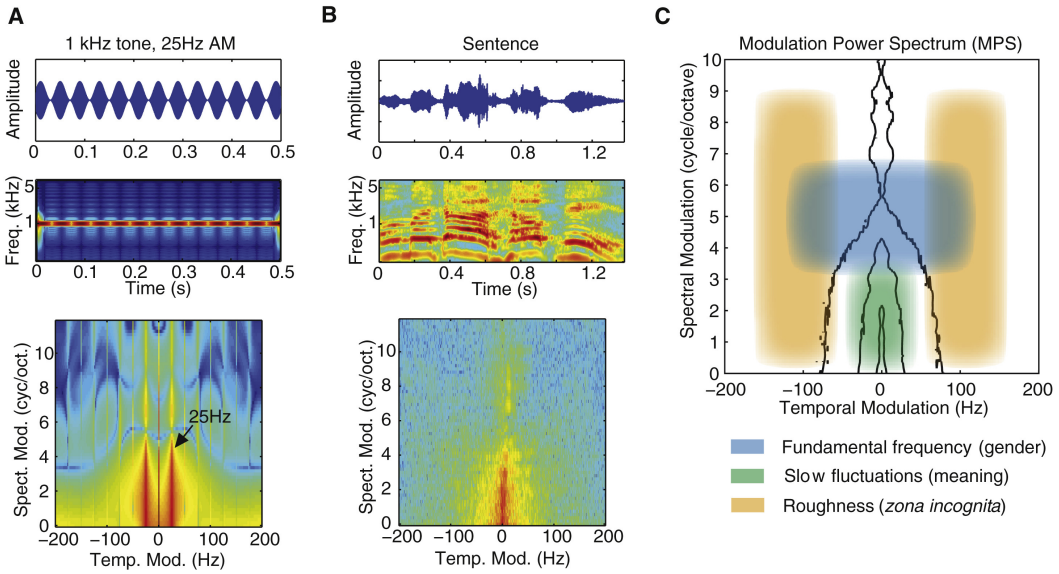


Fig. 1.7.: Examples of spectro-temporal information through the Modulation Power Spectrum (MPS) and its ecological relevance. Image is extracted from (Arnal et al., 2015). Examples are from a modulated tone at 1kHz with 25Hz and a human sentence.

The spectral modulations version of the log-mel Filterbanks is also widely used in the speech processing community, to obtain an even more compressed version of the audio signal. The coefficients that made this representation are referred as the *Mel-Frequency Cepstral Coefficients* (MFCC). The number of coefficients to be used depend on the application at stake, when the slower modulation are useful for speech recognition and higher components for speaker information (See (Elliott & Theunissen, 2009)).

In speech and phonetic sciences, there are other representations to represent what is said and who is speaking based on assumption about the speech production model. A widely accepted model is the source-filter model. The source–filter model represents speech as the superposition of a sound source, based on the vibration of the vocal cords, and a linear acoustic filter due to articulatory placements, the vocal tract. The lowest frequency of the speech waveform is referred as the fundamental frequency (F0) and is computed for instance with the RAPT algorithm (Talkin & Kleijn, 1995). In addition, to obtain representation that distinguish quantitatively the speech sounds, it has been also proposed to extract the distinctive distinctive frequency components of the acoustic signal, the *formants* (See (Ladefoged & Maddieson, 1996) for a complete phonetics review on what distinguish speech sounds). The formants are numbered from the lowest frequency to higher ones, and the three first formants are usually sufficient for speech description F1, F2, F3.

1.2.3 Spoken language as a window into the mind: objective quantification of speaker characteristics from speech

During conversation, spoken language production also engages a number of cognitive skills such as attention, memory and planning, but also social skills such as Theory of Mind and emotional processing. In addition, the different characteristics of the individuals also influence the way we talk, and represents a whole field of study in speech processing, the study of paralinguistics. All these factors are intertwined with the identities of speech sounds (phones, words, sentences) inside the produced waveform. These paralinguistics traits and states can be broadly decomposed into three main categories, and it has been shown to be potentially be captured from speech signals:

Long term traits Biological traits such as height, weight, age (Mporas & Ganchev, 2009; B. Schuller et al., 2010), neurological disorders. Personality traits such as likeability (Weiss & Burkhardt, 2010).

Medium term traits and states Induced states such as sleepiness (Krajewski et al., 2009), mood (Gideon et al., 2016), intoxication with alcohol or drugs (B. Schuller et al., 2011). Structural roles in social groups (Laskowski et al., 2008).

Short term states Emotions (B. W. Schuller, 2018), stress (Hansen et al., 1997), or pain (Simon et al., 2008).

Previously, we introduced the processing methods to obtain more compact representation of the waveform. These methods have been mostly developed with the idea to capture *what is said* by the individuals, to solve Automatic Speech Recognition (ASR). However, the statistical moments of all these components have been proven to be useful for the paralinguistics aspects of speech (B. Schuller & Batliner, 2013; B. Schuller et al., 2013). These moments refer to the statistics (ex: mean, standard-deviation, quantiles) obtained from the corresponding waveform of utterances, words, phones pronounced by a given speaker. One of the most widely used toolkit to obtain these speech representations and statistics is the open-source software OpenSmile (Eyben et al., 2010). A plethora of novel features for the estimation of speaker characteristics have been proposed in the past few years. We review some approaches and features related to the current thesis, and this does not represent a complete review of speech features for paralinguistics.

For instance, it has been proposed by the Thomas Quatieri's group that the vocal tract coordination information can be very informative for different speaker traits

such as depression (Williamson et al., 2014), Parkinson's Disease (Smith et al., 2017) or cognitive load (Quatieri et al., 2015). This method consists in the computation of auto-correlations and cross-correlations of MFCCs of a given speech sequence. This method has been recently extended (and open-sourced) in (Perez et al., 2021), and showed promising results to estimate the motor symptoms in HD.

These statistics of time-frequency representations, such as the spectrogram, log-mel filterbanks or MFCCs, present very informative information concerning the speaker. Yet, the different influences of speaker traits are not directly separable in these speech representations. The spectro-temporal representations, through the computation of the MPS are now being investigated as potential markers of the speaker mental state and traits. It has been shown that these components better assess the intelligibility of individuals with dysarthria (Moro-Velazquez et al., 2015; Kodrasi & Bourlard, 2020; Janbakhshi et al., 2019), but also capture depressive symptoms (Bozkurt et al., 2014) or sleep deprivation (Thoret et al., 2020).

Another line of approach, that combines the knowledge of the speech task and speech technologies, represents a promising path to quickly estimate speaker characteristics. When the speech to be pronounced is known in advance, such as reading or automatic series (counting, months of the year, days of the week), the word and phonetic targets are known by speech researchers. It is potentially possible to transcribe with speech features and ASR systems what is *actually said* by individuals and being compared to the theoretical phonetic and word targets. The comparison is usually carried with edit-based distances such as the Levenshtein distance, that accounts for insertion, deletion, substitution, and transposition of units. Such approach proved to be useful to measure intelligibility in patients with dysarthria (Ghio et al., 2018; Fredouille et al., 2019) or to extract cognitive measure in Parkinson's Disease (Romana et al., 2021).

1.3 Huntington's Disease

1.3.1 History: from George Huntington to 2022

George Huntington was an American medical physician, from Long Island, New-York. He contributed to the first clinical description of the disease that carries his name. He was struck during his childhood by the vision of the abnormal movements, the chorea, and returned after his medical degree in his hometown to study in-depth the disease, where his father and great-father also had notes on the disease.

He introduced the first description of the disease in his 1872 paper "On Chorea" (Huntington et al., 1872), when he was only 22 years old.

George Huntington identified key characteristics of the disease, that are still up-to-date today. At the time, he described three main clinical characteristics for the disease:

1. He described the hereditary nature of the disease, especially the fact that if one of the parent is affected by the disease, a number of the children will develop the disease. He also noted the fact that HD can not skip generations.

2. He also characterized some psychiatric and cognitive aspects of the disease, a "tendency to insanity" that can leads to suicide. He mentioned the progressive aspect of the disease both for the "mind and body" until the death of the patient.

3. Finally, George Huntington also described the time course of the disease, where choreic symptoms start to appear around the thirties and fourties of individuals, without any juvenile case earlier. He referred to the chorea as unpredictable and spasmodic movements in the faces, legs or arms. He always observed a gradual rise of these symptoms and no recovery.

It took about a century, in 1983, to locate the gene 4 as responsible for HD (Gusella et al., 1983). This finding relied on high prevalence of HD in a given huge family in Venezuela, due to high consanguinity (Gusella et al., 1983).

In 1993, a consortium of researchers (Huntington Disease Collaborative Research Group) gathered to share their knowledge concerning the disease, and made significant progresses. By sharing resources and through extensive collaborations, the consortium managed to locate the exact portion of the gene responsible for HD. The gene provoking HD is on the chromosome 4 and is associated with an expanded trinucleotide repetition. The healthy version of the gene contains CAG repeats, yet, when the number of these repeats exceeds 40, the disease is fully penetrant (Rubinsztein, 2003). Full penetrance means that all individuals carrying the mutation will develop the disease.

Only 3 years after the localisation of the portion of the gene, in 1996, the first mouse models of HD were developed by researchers from the United-Kingdom (Mangiarini et al., 1996), from Gillian Bates' group. Summary of the history of research in HD up until the 2000s can be found in Figure 1.8 (extracted from (Walker, 2007)).

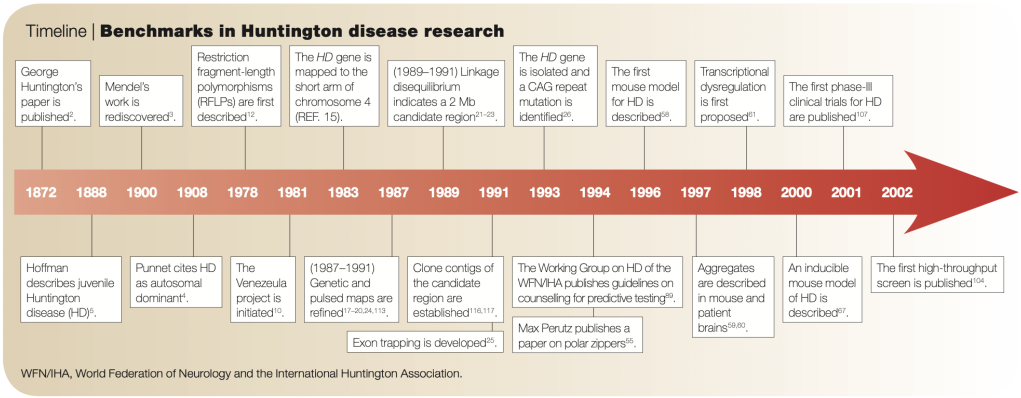


Fig. 1.8.: Timeline of Huntington's Disease research until the 2000s, extracted from (G. P. Bates, 2005). WFN/IHA, World Federation of Neurology and the International Huntington Association.

1.3.2 Epidemiology

HD is an autosomal dominant NDD with complete penetrance when the number of CAG is over 40. In other words, every individual with HD has one of his/her parent with HD, and has a 50% chance to transmit the disease to a child (Shaw & Caro, 1982). However, there are also cases where new mutants of HD appear, where parents did not exhibit the disease, when there is not complete penetrance. New mutants HD where the threshold of CAG exceeded the normal threshold probably account for around 0.1% of all HD individuals (Shaw & Caro, 1982).

HD is a rare NDD, especially in comparison to Alzheimer disease and Parkinson disease, with a stable prevalence in Caucasian individuals of about 5-7 affected individuals per 100000 (Walker, 2007; G. Bates et al., 2002). In Japan, the prevalence of HD is around 0.5 per 100000, and the rate is much lower in most of Asia (Takano et al., 1998). The prevalence in African populations is also very low, even though in places where there are marriages with white people the frequency is higher (G. Bates et al., 2002; Wright et al., 1981).

The observed differences in prevalence can be explained by the difference in the number of CAG repetitions. Indeed, it has been estimated that European ethnic populations have around of 18.4-18.7 CAG repeats while Asian ethnic populations have 16.9-17.4 CAG repeats (G. Bates et al., 2002; McColgan & Tabrizi, 2018).

In the US, there are approximately 30000 individuals with HD with a clinical diagnosis and symptoms, and around 150000 carrying the HD mutation, thus potentially being affected in the future by clinical symptoms (Walker, 2007; Crowell

Fig. 1.9.: Illustration of the Chorea, sometimes described as "Danse de la Saint-Guy" or "Saint Vitus's dance". Around 1880.

et al., 2021). In France, there is an estimate of 6000-7000 individuals with HD with a clinical diagnosis and symptoms, and around 12000 asymptomatic individuals.

HD affects both men and women in the same proportion, and the onset of symptoms is between 30 and 50 years old (Walker, 2007; Tabrizi et al., 2020). Even though, there had been case reports of individuals ranging from 2 to 80 years old. The juvenile form of HD (onset of symptoms before 20 years old) accounts for 5% of the HD population (Quarrell et al., 2012). Late form of HD (more than 60 years old) of the disease accounts for around 15% of the HD population (Chaganti et al., 2017). As the copy of the gene 4 becomes unstable when the number of CAG repeats exceeds 28, successive generations can exhibit higher number of CAG and higher severity. This phenomenon in genetics is called *anticipation*. Anticipation is aggravated when the father is transmitting HD to the children in comparison to the mother, due to higher instability to produce male gametes than female ones.

1.3.3 Patient care

Genetic predictive testing

In the healthy population, the mean number of CAG in the portion of the gene 4 is between 16 and 20 repetitions, while 36 is the minimum number for HD (Walker, 2007; G. Bates et al., 2002; Tabrizi et al., 2020). Since the discovery of the gene

responsible for HD, and the localization coding for the huntingtin protein, individuals can go through a predictive test for HD (Cox & McKellin, 1999). Even though there is this possibility, only a few of the individuals at risk for HD (around 15%) decide to go through it (Quaid & Morris, 1993), even though they mostly declare to be willing to do it (Kessler et al., 1987). The predictive test is given in HD certified centers around a multidisciplinary teams of neurologists, neuropsychologists, and social workers. There are complicated reasons to go or not undergo through the test. Indeed, a positive test can have profound impact for the individual and its relatives. Pretest preparation is of the utmost importance, as it has been observed that between 2 and 6% may have severe psychiatric or suicidal ideas due to a positive outcome (Kessler et al., 1987). Based on the results of the genetic test, an estimation of the onset of the symptoms can be computed (Langbehn et al., 2004).

There is incomplete penetrance of HD, when the number of CAG is between 36 and 40, not necessarily associated with the disease. Yet, there is a higher risk to transmit the disease to the children. Above 40 repeats there is complete penetrance of HD, provoking the formation of the mutant Huntingtin protein (mHTT). The roles of this protein is not totally understood, but we know it plays a critical part during development and for normal functioning of neurons (Schulte & Littleton, 2011).

There is now the possibility for predictive testing concerning the future children for couple at risk to transmit the disease. Due to advances in genetic testing, it is possible to know if a fetus is carrying the mutation of HD. Based on the results, the couple can be faced to the choice of abortion. These questions around the selection of fetuses with and without the HD mutation remains a controversial and ethical question (Braude et al., 1998). On the one hand, there are arguments concerning that selective abortion is unacceptable, since the child carrying the HD mutation can still have decades of a life without the disease. On the other hand, to not do the test has implications for parents that desire not want to know, potentially to respect the right of children to not know their statuses without their consent. There is also the possibility now to run genetic analysis for embryo produced by in vitro fertilization (Van Rij et al., 2012). This allows the parents to select only for the healthy embryo and avoid termination of fetuses.

Evolution of the disease until death

The evolution of HD is divided into two parts: an asymptomatic pre-manifest (preHD) phase and a symptomatic manifest (HD) phase (Walker, 2007; Ross, Aylward, et al.,

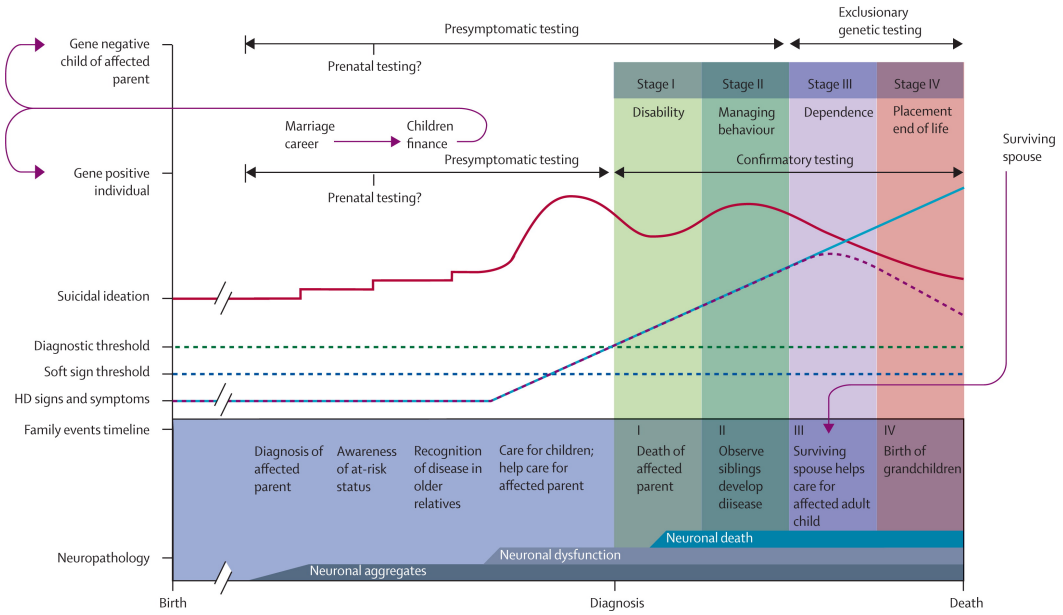


Fig. 1.10.: The course of Huntington's Disease. The events can occur in different orders, as the disease is highly heterogeneous. It illustrated the recurrent and family nature of the disease. This images is extracted from (Walker, 2007).

2014). In this thesis, we followed this system, and did not consider sub-staging of HD in the preHD phase or HD phase.

The official clinical onset is declared by a certified neurologist, that observed symptoms related to HD with a great confidence. HD is a progressive disease, and the burden of the disease keeps increasing as years passes on. The quality of life of HD, their spouses and children is also impacted as disease progresses (Ready et al., 2008). It has been observed that the cognitive and independence statuses of HD carriers were the most impactful of the quality of life of carriers *and* their spouses (Ready et al., 2008). The psychiatric aspect of the disease impacts differently the HD carrier and the spouse, where caregivers suffer the most.

Based on the data from the European HD Network prospective study (REGISTRY), it was also possible to assess the survival, mortality and causes of death in HD in Europe (Rodrigues et al., 2017). In this cohort, the mean age at diagnosis was 49 years old and the mean age of death was 59 years old. The median survival is 24 years from diagnosis, 35 years from the symptoms onset, and the most observed causes of deaths were pneumonia (19.5%), other infections (6.9%) and suicide (6.6%) (Rodrigues et al., 2017). All the different events in the life of a carrier of HD mutation are summarized in Figure 1.10.

1.3.4 Neuropathophysiology

The Huntingtin protein is expressed in all human and mammalian cells, with high production in the brain (Tabrizi et al., 2020). The Huntingtin is believed to play a part in the different steps of the cycle of life of cells, especially in the programmed cell death, the aptosis (Rangone et al., 2004). See Tabrizi et al. (2020) for a recent review of the molecular pathogenesis of HD.

The associated mutant HTT protein (mHTT) that results from the CAG repeat on the gene HTT leads to selective neural dysfunction in HD and neurodegeneration. The mechanisms that provokes these disturbances are not completely understood (and beyond the scope of this thesis). Yet, the mHTT is believed to potentially harm the ability of cells to degrade them, leading to an increase of unmanageable aggregates of proteins (Walker, 2007; Tabrizi et al., 2020; Rangone et al., 2004). These aggregates of proteins in return interfere with other normal proteins or retain active binding sites for the normal huntingtin protein (Walker, 2007; Rangone et al., 2004).

In particular, the medium spiny neurons in the striatum are particularly vulnerable and the most affected in the brain to the presence of the mHTT (Tabrizi et al., 2020; De Diego Balaguer & Bachoud, 2006). HD is sometimes charaterized as a basal ganglia disease, as the Parkinson's Disease (PD), since it affects selectively the striatum (Scahill et al., 2020; Douaud et al., 2006) and its white matter connections (Stoffers et al., 2010; Phillips et al., 2014; Matsui et al., 2015) with the rest of the brain.

The basal ganglia is composed of several nuclei, the striatum, the internal and external globus pallidus, the subthalamic nucleus and the substantia nigra. The striatum is itself composed of different nuclei, the caudate, the putamen and the ventral striatum. These nuclei are involved in a number of loops to control motor, cognitive, and limbic behaviors (Graff-Radford et al., 2017; G. E. Alexander et al., 1986).

The 5 "parallel" loops that recruit the basal ganglia and the striatum that support the different behaviors in concert with the different parts of the cortex and the thalamus, are illustrated in the Figure 1.12. Even though these loops are illustrated as completely separated, there is a large overlap between them and this represents only an approximate schematic involvement of the striatum.

Thanks to autopsies (J.-P. Vonsattel et al., 1985; Roos et al., 1985) and the progress of neuroimaging techniques (Douaud et al., 2006; J. P. G. Vonsattel, 2008) allowed

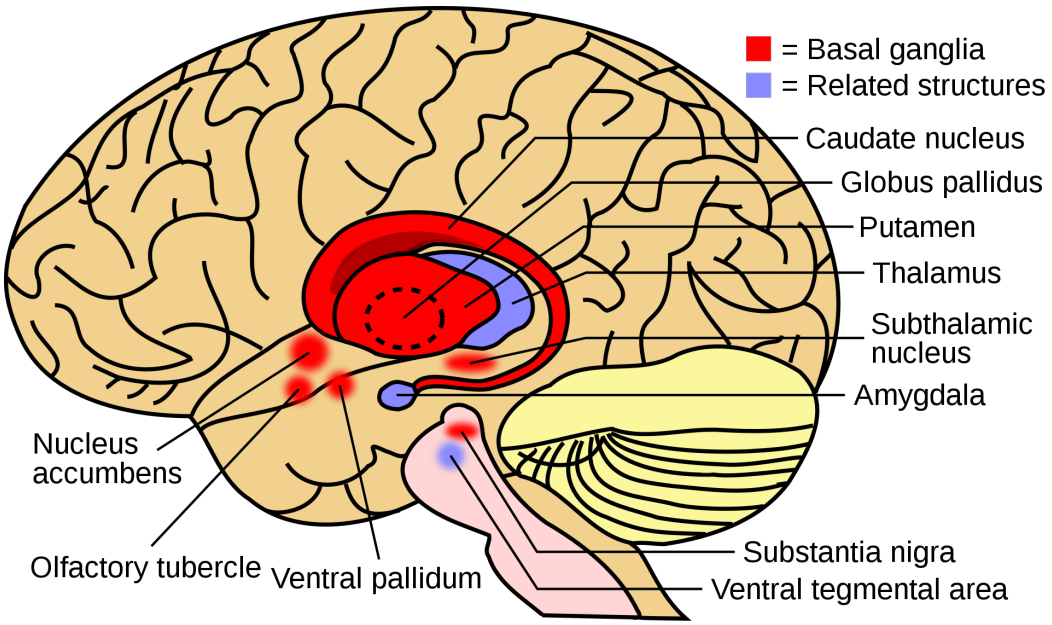


Fig. 1.11.: Structures composing the Basal Ganglia. This image is extracted from Wikipedia.

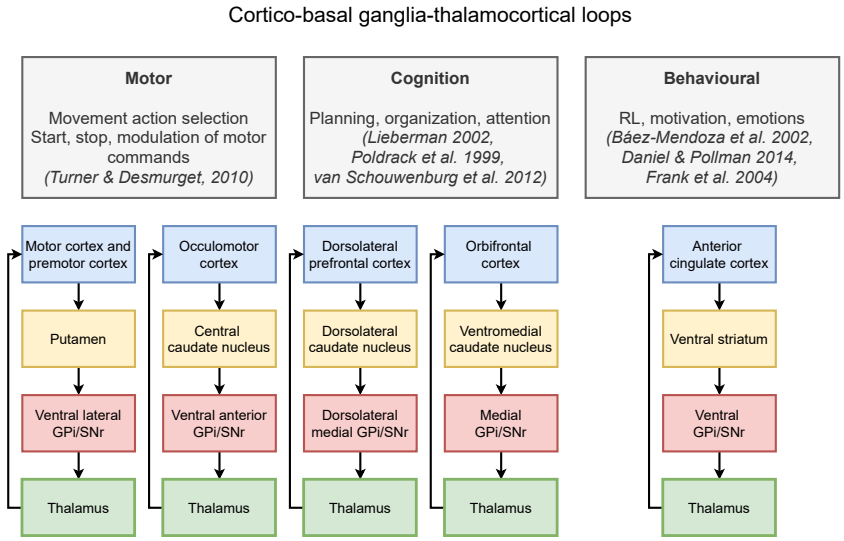


Fig. 1.12.: Schematic "parallel" anatomical organization of the cortical-basal ganglia-thalamocortical loops. Schema is adapted and extended from (Graff-Radford et al., 2017; G. E. Alexander et al., 1986) RL: Reinforcement Learning, GPi: Internal Globus Pallidus, SNr: Substantia Nigra Pars reticulata. References in the schema: Motor (Turner & Desmurget, 2010), Cognition (Lieberman, 2002; Poldrack et al., 1999; van Schouwenburg et al., 2012), Behavioural (Báez-Mendoza & Schultz, 2013; Daniel & Pollmann, 2014; Frank et al., 2004)

to confirm selective striatal atrophy, and its progression through lifespan. The loss of neural cells in the striatum and the disorganization of white matter tracts is present even before the official clinical onset of HD (Scahill et al., 2020). As shown in Figure

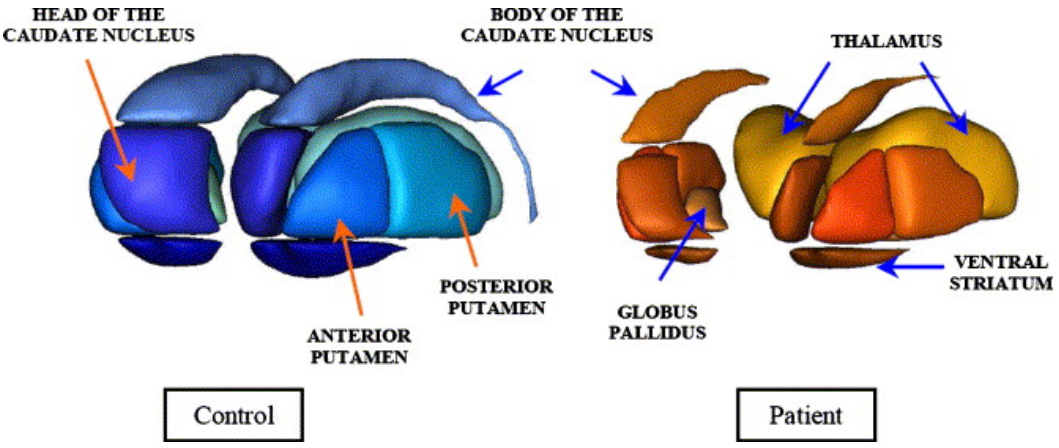


Fig. 1.13.: 3D model reconstruction of the different sub-cortical nuclei of a control (blue left) and of a patient HD (orange right) matched in age and sex. This image is extracted from (Douaud et al., 2006).

1.13 extracted from (Douaud et al., 2006), all the different nuclei in the striatum are affected by the disease. Even though, HD selectively damages the striatum, with around 50% volume reduction, the disease affects also almost globally affects the brain, with a 19% loss of the total volume of the brain (Halliday et al., 1998).

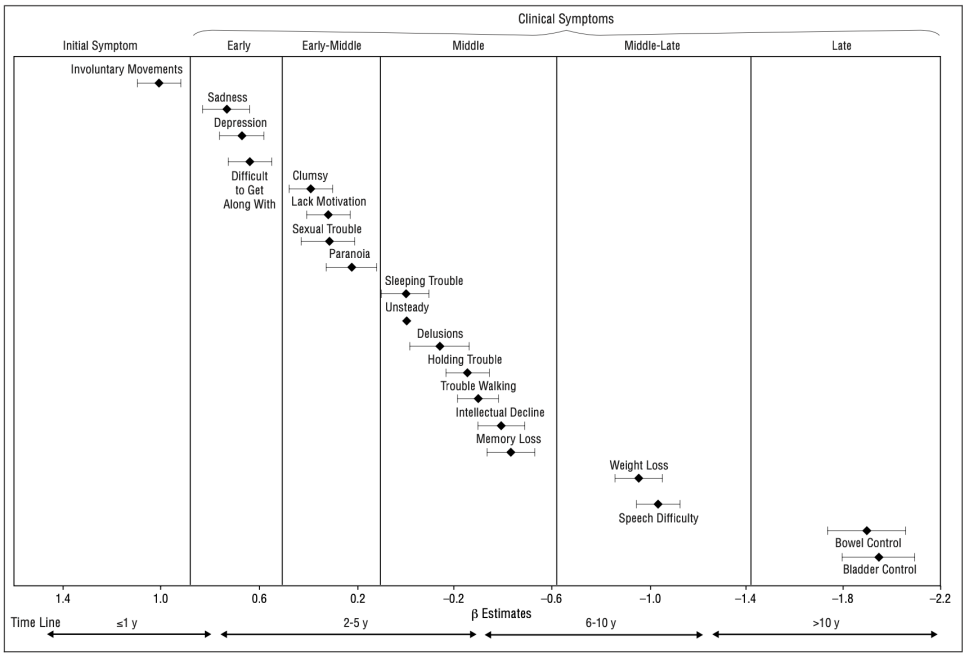
1.3.5 Clinical symptoms

Due to the different involvements of the striatum in the different loops described in section 1.3.4, HD provokes a triad of symptoms: motor, cognitive and psychiatric/behavioral (Walker, 2007; Novak & Tabrizi, 2010). These symptoms evolves as the neural cell deaths progresses, up to the point where an official diagnosis of the disease can be pronounced by a certified, renewed every year, neurologist. See Figure 1.14 for the average time course of symptoms in HD. The neurologists declares the beginning of the disease when there are signs extrapyramidal chorea, poor coordination, dystonia or bradykinesia (Kremer, Group, et al., 1996). This focus on motor symptoms for HD is quite arbitrary as cognitive and/or behavioral symptoms can start even before motor ones (Duff, Paulsen, et al., 2010; J. S. Paulsen et al., 2005). All these symptoms contribute to the deterioration of the functional capabilities of individuals with HD (Schobel et al., 2017). The functional capacity refer to the capacity of individuals to carry out their activities of daily living (eating, bathing) and instrumental activities of daily living (managing medication or personal finances). The functional capacity assessment represents the closest evidence for "real-world" performance of HD.

In this section, we will described on the clinical HD symptoms, and we will come back later in the thesis, how some of these clinical symptoms are measured through clinical valid tests. In addition, this small review of clinical symptoms in HD is not exhaustive and readers can refer to Walker (2007), G. Bates et al. (2002), G. P. Bates et al. (2015), and Youssov and Bachoud-Levi (2008) for more complete reviews.

Motor symptoms

The motor symptoms in HD can be divided in two main categories (Novak & Tabrizi, 2010): (1) the addition of involuntary movements such as chorea and (2) impairments of voluntary movements that cause incoordination of limbs and hands.



Mean and 1 SD of β estimates for 19 symptoms of Huntington disease. Vertical bars represent proposed divisions based on clustering of β estimates into 6 onset periods.

Fig. 1.14.: Overview of the average time course of the symptoms in HD after clinical diagnosis. This is based on given responses by a close relative of the affected participant such as spouse or child. This image is extracted from (Kirkwood et al., 2001).

Chorea is the main motor symptom in HD, thus the former name of HD "Huntington's chorea". Chorea refers to movements that are spontaneous, excessive, irregular, unpredictable in time, randomly distributed on the body (G. Bates et al., 2002). The chorea is always present when not asleep, and can not be voluntary suppressed. Chorea is present in 90% of HD cases during first phases of the disease. As the

disease progresses, chorea tends to slowly disappear and being replace by other more debilitating motor symptoms.

Bradykinesia (slowness of movements), and *rigidity* are usually the symptoms that mostly associated wtih Parkinson's Disease (PD), and infrequent in the early stages of HD. Yet, these symptoms tend to replace the chorea in the late stages of the disease (Kirkwood et al., 2001). *Dystonia* is also infrequent in the early stages of the disease but worsens over time, and is prominent in the last stages (Louis et al., 1999). Dystonia refers to slow abnormal movements along abnormal posturing. In HD, dystonia is present in several body regions, shoulders, fists, knee and feet. In addition, it was noted it was not bothersome for the patients for their daily functioning.

These types of symptoms (Bradykinesia, rigidity and dystonia) are also function of the medications to the patients, such as anti choreic drugs or anti psychotic drugs (Unti et al., 2017) and can increase their severity as side-effect. As the disease progresses, there is also gait disturbances and patients can face frequent falls associated with morbidity and confinement to wheelchair.

Swallowing impairments, or Dysphagia in HD start to occur in the late stages of the disease, primarily affecting fluid intakes and also solid intakes later. Choking due to these impairments is a common cause of morbidity. Finally, there are also motor speech abnormalities (which we will review into more depth in Section 1.3.7) that provokes social isolation of individuals with HD.

Cognitive symptoms

Even though the motor symptoms are the most evident at first examination, there is no doubt that non-motor symptoms have the biggest impact on HD daily routines, and contributes the most to the loss of autonomy (Ross, Pantelyat, et al., 2014; Hamilton et al., 2003). These cognitive symptoms are also the object of a number of plaints from patients concerning the impact of the disease. When the severity of the cognitive symptoms in HD is aggravated up to a loss of autonomy, they are usually referred as *dementia*. The cognitive symptoms may also have a huge impact on the social environment and the relatives of HD individuals.

Individuals with HD, can experience *attention* deficits and perturbation of *executive functions*. Executive function refers to a number of capacities that are related to the regulation, control of cognitive performance. As individuals carry planned and goald directed behavior, actions need to be organized, attend to the relevant stimuli,

ignore irrelevant ones, shift attention when necessary as task unfolds. Individuals need to build abstract representation, and integrate it with prior knowledge, and be able to self-correct errors in line with intended goals. *Processing speed* of information is also an important cognitive feature in HD and represent one of the most frequent cognitive impairment (Duff, Paulsen, et al., 2010). It affects almost all the goal-directed behaviors and decrease the performance in time-dependent tasks.

These functions rely heavily on the fronto-striatal loops as seen before, and individuals of HD face difficulties in a number of these functions (G. P. Bates et al., 2015; Youssov & Bachoud-Levi, 2008; Duff, Paulsen, et al., 2010).

Learning new information/competence are also impaired in HD (De Diego-Balaguer et al., 2008), especially due to disturbances of reinforcement learning loops relying on the ventral striatum. Even though, memory is less affected as much as Alzheimer disease (AD) (Aretouli & Brandt, 2010). Even though, HD are not considered aphasic, it has been demonstrated that HD exhibit linguistic impairments We dive into more depth into these troubles in the section 1.3.7.

HD also provokes perceptual and spatial impairments (Glikmann-Johnston et al., 2019). They do not exhibit recognition and identification deficits for objects, but they have more difficulties concerning *spatial memory* tasks (Glikmann-Johnston et al., 2021), (ex: how objects displayed previously on a screen, were placed relatively to each other?).

Psychiatric symptoms

Even though the psychiatric features of HD are not as consistent as motor and cognition ones, they can be very detrimental to the individual and its relatives. *Depression* is very common and up to 50% of these depressive symptoms have been reported at least once during the course of the disease (Thompson et al., 2012). Irritability and apathy also common features in the psychiatric portrait of HD (J. S. Paulsen et al., 2001). *Suicide* and its ideation are also more prevalent in HD populations than in the general population (Solberg et al., 2018).

1.3.6 Biomarkers in Huntington's Disease

In this thesis, our goal is to evaluate the potential of spoken language as potential biomarkers in HD. Biomarkers are those measures that reliably track the progression of the disease, potentially reflect the disease severity (Killoran & Biglan, 2016).

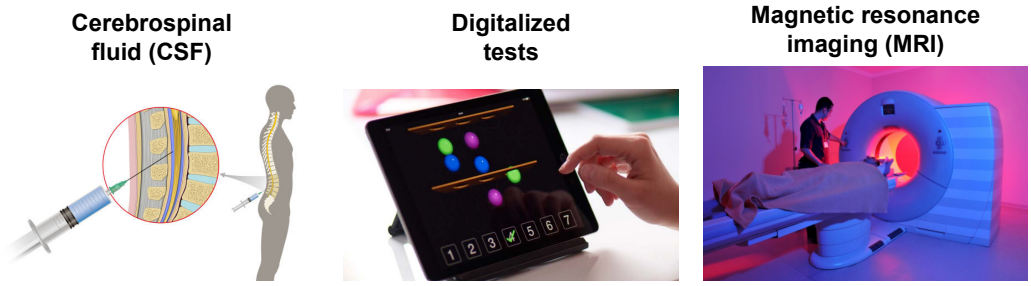


Fig. 1.15.: Illustration of the current candidate biomarkers in HD.

In this thesis, biomarker refer to objective phenomena of the neurological state observed without the introspection of the patient, which can be measured accurately and reproducibly. The ideal biomarker in NDDs should remain affordable, within reach for all individuals, unaffected by a comorbidity, and should not have huge variability among healthy individuals (Henley et al., 2005). Furthermore, the collection and validity of biomarkers should not vary across facilities.

We illustrate in Figure 1.15 the main candidate biomarkers in HD that have been proposed in the past few years. Especially, cognitive batteries with digitalized tests (Stout et al., 2014; Lunven et al., 2021), brain imaging, focusing on striatal volumes (Mason et al., 2018; Zeun et al., 2019), or biofluids with Human Cerebrospinal Fluid (CSF) (Scahill et al., 2020; Rodrigues et al., 2020) showed promising results. However, with the current state of technologies, all these biomarkers require the presence of the patient at the hospital and a high level of expertise or equipment.

1.3.7 Spoken language production in Huntington's Disease

As mentioned in the previous section 1.1, spoken language production encompasses the (1) production of well-formed sentences based on intentions and (2) the fluent production of speech through the coordination of respiratory, phonatory and articulatory sub-systems. Here, we overview the current knowledge of spoken language production in HD, and its potential to be a biomarker in HD. The speech disorders in HD are well known and had been the subject of more studies than the linguistic aspects.

The speech disorders in HD are usually referred to as *hyperkinetic* dysarthria, and is present in 93% of individuals (Rusz, Klempir, Tykalova, et al., 2014). The first description of hyperkinetic dysarthria comes from Darley et al. (1969). This categorization of dysarthria is usually depicted speech as having variable rate, abnormal prosody, imprecise consonants and distorted vowels, phonation deviations,

and sudden forced breath. Hyperkinetic dysarthria is broad description of HD's speech production, but in practice speech impairments in HD are diverse. Indeed, there is a recent perceptual study showing that the speech motor impairments in HD are highly heterogeneous (Diehl et al., 2019). For a complete review of speech impairments in HD, readers can refer to (J. C. Chan et al., 2019).

Concerning speech perception, HD individuals can exhibit prosody perception deficits (Speedie et al., 1990), and have an overall decrease of speech understanding, especially under demanding conditions (Profant et al., 2017). These perceptual deficits are observed despite the fact that individuals with HD seem to have normal hearing.

Concerning the content of discourse, it has been reported language disorders both in *perception* and *production* at several linguistic levels:

1. *Perception*: syntax (Sambin et al., 2012; Giavazzi et al., 2018), rule application (Teichmann et al., 2008; Teichmann et al., 2006), word comprehension (PODOLL et al., 1988), sentence comprehension (Wallesch & Fehrenbach, 1988; Saldert et al., 2010), pragmatics (Chenery et al., 2002).

2. *Production*: morphology and rule applications (Ullman et al., 1997; Teichmann et al., 2005), semantic diversity and syntactic complexity (Hinzen et al., 2017; Gordon & Illes, 1987), distribution of words (Wallesch & Fehrenbach, 1988).

For a complete review of linguistic impairments in HD, recent at time of writing, readers can refer to Gagnon et al. (2018). The different studies around the speech and linguistic impairments offer a broad view around the spoken language production, but consensus is not reached as underlined by motor speech patterns controversies (Diehl et al., 2019; Rusz & Tykalova, 2021) or morphology rule applications controversies (Ullman et al., 1997; Teichmann et al., 2005; Longworth et al., 2005).

For instance, Diehl et al. (2019) collected speech from 48 manifest HD at several stages of the disease, and found out 4 distinct sub-groups based on perceptual listening. These groups were most distinctive according to severity of dysarthria and also the speed of spoken language production. These findings triggered an official answer in the Neurology journal from Drs. Rusz and Tykalova, as the Diehl's findings were at odds with their own (Rusz & Tykalova, 2021).

In addition, the studies around HD's speech and language production do not use the same methodologies and cohorts remain small and heterogeneous across studies. This heterogeneity also extend in the choice of speech tasks and the markers that

are used and being evaluated. In addition, there is a variability in the research question and thus analyses. Researchers can use statistical methodologies and draw conclusions at the group level. On the other hand, there are more and more studies using machine learning methodologies, to combine a number of speech and linguistic markers to evaluate the predictive capabilities. In these machine learning studies, analyses, splitting of data for evaluation, models remain also heterogeneous.

That is why, we made reviews of the main studies around spoken language production in HD, with a special focus on studies with reproducible markers that can be extracted in a programmatic way, excluding pure subjective perceptual studies. We made the distinction between the study using statistical methods (see Table 1.1) and the ones using machine learning methods (see Table 1.2).

As we can see in both tables, the sample size of each study remain limited and never exceeded 100 participants. The participation of pre-symptomatic individuals (preHD) remain also limited. Based on these studies it is complicated to conclude if preHD exhibit clear speech patterns that could be spotted to differentiate them from healthy controls, both at the group and individual levels.

Finally, we relocate these different studies in the context of the Levelt's model of spoken language production introduced in the section 1.1.1, in Figure 1.16. We emphasized in this figure which potential processing component have been reported impacted in HD. As underlined by Jacquemot and Bachoud-Lévi (2021) and the DIVA and GODIVA models of Guenther (2016), the striatum is involved in the different steps of spoken language production.

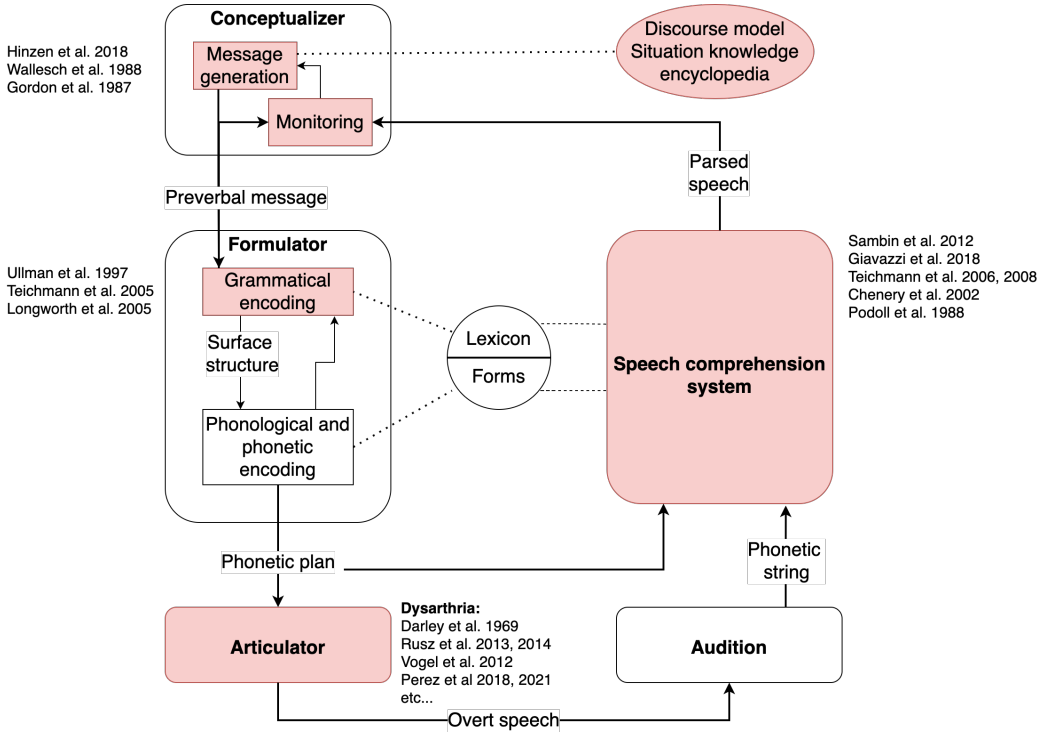


Fig. 1.16.: Approximate localization of speech and linguistic impairments in Levelt's psycholinguistic model of spoken language production (Levelt, 1993). In red, we underline the components that are presumed to be impaired in HD.

Tab. 1.1.: Review of speech and linguistic studies in Huntington's Disease using *only* statistical methods at the *group level*. Analysis reflects the research question in the given study. We focused on HD populations. HC stands for Healthy Controls. We include studies with at least 20 participants, using features from audio or linguistic annotation (no perceptual studies) and with clear explanations of speech and language features.

Study	Levels	Speech tasks	Cohort	Features	Analyses	Results
(Gordon & Illes, 1987)	Lexicon, Hesitations, Syntax, Intelligibility	Cookie-theft + travel story	12HD 24preHD	(un)filled pauses, prolongation, repetitions, interjections, syntactic complexity, paraphasia	HD vs preHD	Princeps study, less words, more silences, less filled pauses, more prolongations, more paraphasia
(Ullman et al., 1997)	Inflectional morphology	Sentence completion	17HD 8HC	(In)correct answer	HD vs HC	Overregularized irregulars verbs, small N
(Hertrich & Ackermann, 1994)	Unit timings voicing	Sentence completion with nonsense target word	17HD 8HC	Utterance duration, syllable lengths, vowel duration,	HD vs HC	Variability of utterance duration and VOT were the most significant.
(Hinzen et al., 2017)	Lexicon, Hesitations, Syntax, Intelligibility, Rhythm, Grounding	Cinderella	9PreHD 19HD 28HC	Quantitative, Fluency, Reference, Connectivity, and Concordance	HC vs preHD vs HD, brain and clinical scores correlations	HD differ from HC in all dimensions, preHD differ from HC for Reference and Connectivity, Quantitative correlated with motor, cognitive scores and striatum volumes
(Vogel et al., 2012)	Timing, pitch and intensity	Reading, monologue, days of the week.	13PreHD 17HD 15HC	Speech rate, Silence and speech times, intensity α ration, F0 cov (mean/SD)	HC vs preHD vs HD, clinical scores correlations	HD differed in all dimensions, preHD in between HC and HD. Speech rate and silence correlated with burden score and TMS. Effect of cognitive load.
(Skodda et al., 2016)	Phonation, Speech rate	Reading task, diadochokinesis (DDK) task	28PreHD 28HC	Speech rate (reading), maximum phonation time, maximum and steadiness of syllable repetitions	preHD vs HD, brain and clinical scores correlations	Speech rate correlated with TMS, DBS, cognitive score, *motivation* of preHD>*motivation* of HC
(Chenery et al., 2002)	"High linguistic" levels	Western Aphasia Battery, Boston Naming test, Test of Language Competence, Test of Word Knowledge	13HD 13HC	In(correct) answer	HD vs HC	spared primary language functions in Western Aphasia Battery, impairments in more complex abilities (cognitive–linguistic flexibility and planning for production)
(Rusz, Tykalova, et al., 2021)	Fundamental Frequency, DDK task, Vowel articulation, Voicing, timing	Sustained vowel, sequential and alternating motion rates, reading, monologue	30HD 50HC	Loudness, F0 range and SD, Pitch breaks, Shimmer, Jitter, Harmonics-to-Noise Ratio (HNR),DDK rate and regularity, Vowel Space Area(VSA)	HC vs HD	HD differed in all dimensions. Speech rate had biggest effect size.
(Novotny et al., 2016)	Hypernasality	Sustained vowel /i/	37HD 37HC	Mean, SD and trend of the energy in the filterbank centered around 1 kHz (passband from 890.9 Hz to 1122.5 Hz).	HC vs HD, correlation with UHDRS chorea composite subscore	HD differed for mean and SD. 54% of HD affected. Significant correlation between the UHDRS chorea subscore and the SD of the energy.
(J. Chan et al., 2021)	Speech and silence timings, Speech rate	Reading of Grandfather passage, Monologue	16preHD 16HD 30HC	Speech rate, and silence rate, Mean and SD duration of pauses	HC vs preHD vs HD, correlation with SDMT	Speech rate in reading separated all three classes, and correlated with SDMT.

Tab. 1.2.: Review of speech and linguistic studies in Huntington's Disease using machine learning methods to make predictions at *individual level*. We focused on HD populations. HC stands for Healthy Controls. Split refers to train/test splitting method for evaluation.

Study	Levels	Speech tasks	Cohort/Split	Features	Analysis and models	Results
(Rusz, Klempir, Baborova, et al., 2013)	Airflow insufficiency, Aperiodicity, Irregular vibration of vocal folds, Signal perturbations, Increased noise, Vocal tremor, Articulatory deficiency	Sustained vowel /a/	136 phonations from 34HD and 136 from 34HC, 20 repeats of 80% train/20% test, split at phonation level	Maximum phonation time (MPT), Maximum phonation time until voice break (MPT_{VB}), number of voice breaks, degree of voicelessness, F0 SD, Recurrence period density entropy (RPDE), Pitch period entropy, Jitter, Shimmer, HNR, Detrended fluctuation analysis (DFA), 12st MFCCs.	HC vs HD, Support-Vector-Machine (SVM) classifier with Gaussian radial basis kernel	sensitivity ($95.1 \pm 4.0\%$) and specificity ($93.2 \pm 4.3\%$), very high reported performance, SVM parameters and features selection cross-validated on the same data.
(Rusz, Saft, et al., 2014)	Airflow insufficiency, Aperiodicity, Irregular vibration of vocal folds, Signal perturbations (Jitter/Shimmer), Increased noise, Articulatory deficiency	Sustained vowel /a/.	28PreHD 28HC, 20 repeats of 4-fold 75% train/25% test	MPT, MPT_{VB}, number of voice breaks, degree of voicelessness, F0 SD, RPDE, Pitch period entropy, Jitter, Shimmer, HNR, DFA, frequency tremor intensity index (FTRI) and amplitude tremor intensity index (ATRI), MFCCs Mean of SDs.	HC vs preHD, Support-Vector-Machine (SVM) classifier with Gaussian radial basis kernel	sensitivity ($91.5 \pm 4.0\%$) and specificity ($79.2 \pm 4.3\%$), high reported performance,
(Rusz, Klempir, Tykalova, et al., 2014)	Vowel articulation, pitch, loudness and timing	Reading passage and monologue	40HD (22 with antipsychotic medication) 40HC, 20 repeats of 60% train/40% test	VSA, Vowel articulation index (VAI), Mean and SD of F0, pause number and ratio, speech rate.	HD vs HC, sub-analyses per gender, SVM with classifier with Gaussian radial basis kernel.	Males' best features are from reading passage with sensitivity ($99.8 \pm 1.6\%$) and specificity ($98.5 \pm 4.1\%$). For females, intensity variation and VAI from reading, number of pauses from monologue predicted HD with sensitivity ($99.8 \pm 1.6\%$) and specificity ($98.5 \pm 4.1\%$). Medication influenced pauses in monologue.
(Perez et al., 2018)	(Un)filled pauses, speech rate, Goodness of Pronunciation	Reading of Grandfather Passage	12preHD 19HD 31HC, Leave-One-subject-out evaluation	Utterance-level (dynamic) and Speaker-level (static) pooling of features.	HD vs (preHD/HC), comparison of k-NN, DTW, DNN, LSTM-RNN modelling of utterance	Automatic Speech Recognition (ASR) methods to retrieve features. First fully automatic HD speech study. Accuracy 87% with oracle features, 87% with ASR.
(Romana et al., 2020)	Vowel lengths, and trends and fluctuations of vowels	Sustained vowel /a/ and reading of Grandfather Passage	12preHD 18HD, Leave-One-subject-out evaluation	MPT, DFA, trend and fluctuation measures	HD vs preHD, Logistic regression	Accuracy 88% with all vowel features. Fluctuations and duration of vowels of Grandfather passage outpeformed features from (Perez et al., 2018). Vowel duration stats informative about HD (sustained vowel, reading).
(Perez et al., 2021)	Vocal Tract Coordination	Reading of Grandfather Passage	12preHD 19HD, 100 repeats of 80% train/20% test of speakers	Auto and cross-correlations of 16 MFCC and delta of MFCC.	prediction of TMS, Elasticnet regression model	RMSE of 17.9 (3.5) using the delta of MFCC, no need of ASR to extract features.

2

BasalVoice: speech database of individuals carrying the mutation in the huntingtin gene

> *80 percent of our work is data preparation, then ensuring data quality is the important work of a machine learning team.*
>
> — **Andrew Ng**
> Former Chief Scientist of Baidu
> Co-Founder of Coursera and Google Brain
> Adjunct Professor at Stanford University

This chapter describes the acquisition, the annotation, the quality assessment, and linguistic formatting of a new database of French speech produced by individuals with HD, along their clinical assessments. This database, called BasalVoice, is the result of a collaboration between the Cognitive Machine learning team at the National Institute for Research in Digital Science and Technology, and the Interventional Neuropsychology team, at the Neurology department in Hospital Henri-Mondor in Créteil. This database is the first necessary and unavoidable step to test the potential of speech and language as potential biomarkers in HD and to build the algorithms to democratize their uses. Only a few reports and research studies exist concerning the creation of database with individual with neurological troubles and spoken language impairments, such as (Ghio et al., 2012).

The collection for the BasalVoice database is still ongoing and at time of writing currently includes 300 interviews of individuals with speech impediments caused by HD, premanifest individuals carrying the genetic mutation, the preHD individuals, and healthy control subjects. All the participants went through standardized assessments of motor, cognitive, psychiatric and functional capabilities by professional neurologists, neuropsychologists and speech pathologists. Among these interviews, 145 were annotated at different linguistic levels by speech pathology students. The

goal of this resource is three-fold, with different scientific perspectives and ultimate goals:

Clinical Neurology Derive new ecological biomarkers that reflects the different symptoms of HD.

Neurolinguistics Quantify the HD spoken language deficits at different levels, to inform on the role of the striatum in the production of spoken language.

Speech technology Design and build tailored engineering pipelines to automatically quantify the spoken language deficits provoked by HD.

In Section 2.1, we describe the interview protocol, instrumentation and circuit of the audio data, and the demographics of the interviewee. Then, in Section 2.2, we describe the annotation protocol of different linguistic information, the management of the fleet of annotators and the quality assessments of the linguistic annotations. Finally, in Section 2.3, we describe the post-processing and parsing methods of the audio and linguistic data that we used to be able to analyze in a automatically the BasalVoice database.

2.1 Data collection

2.1.1 Demographics of interviewers and interviewees

Data collection began in June 2018 at the hospital Henri-Mondor in Créteil, France and is still ongoing at time of writing. The following sections describe the different aspects of the interviews and their audio collection. So far, six interviewers (1 male, 5 female) have carried out the speech interviews in the collection of the BasalVoice database. They were between 23 and 36 years old at the start of the study and they were all certified neuropsychologists without any speech deficits.

Among all the interviewees, we find: Healthy controls and individuals tested with a number of CAG repeats on the Huntington gene above 40. These latter feature different stages of the disease ranging from asymptomatic preHD to the stade 4 of the disease HD. Their statuses change as a function of the visit. All the participants had between 1 and 4 interviews spaced in time between June 2018 and October 2021. The Healthy controls, the preHD individuals and the HD patients were enrolled through the RepairdHD clinical study (NCT03119246) and the BioHD clinical study (NCT01412125). All participants signed an informed consent. Ethical approval was given by the institutional review board from Henri Mondor Hospital (Créteil, France)

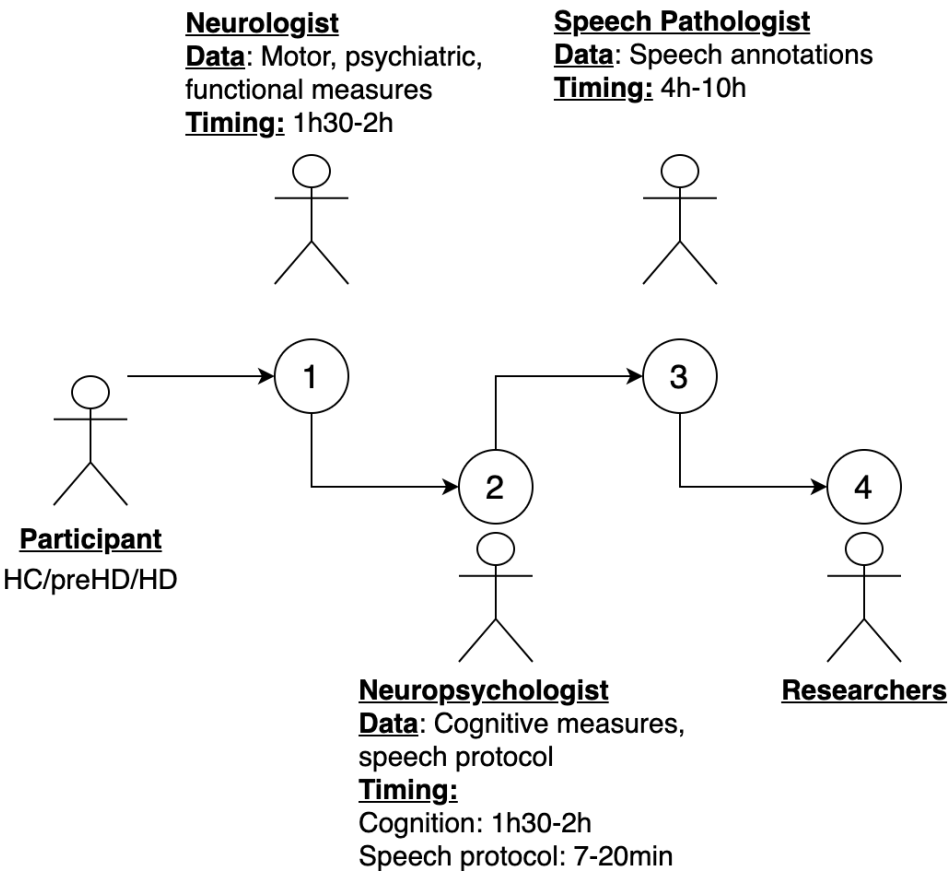


Fig. 2.1.: Collection of data along the healthcare circuit of participants. 1) Participants meet neurologists. Neurologists assess motor, psychiatric symptoms and functional capabilities. 2) Participants meet neuropsychologists. Neuropsychologists assess cognitive functions and record the speech tasks of our protocol. 3) Audio data is annotated by Speech pathologists. 4) Audio data and its annotations are handed for analyses to researchers.

for all the protocol. It complied with the Helsinki Declaration, current Good Clinical Practice guidelines, and local laws and regulations.

The participants were between 25 and 80 year-old. All participants were required to have no history of severe hearing or speech problems prior to the disease onset. They were all french-native speakers. During each visit of the day of the interview, the participants were assessed by certified examiners through a large panel of measures classically used for both clinical practice and clinical trials, described in the following section. The full circuit of individuals of the collection of clinical and speech data is represented in Figure 2.1.

Tab. 2.1.: Messick's Model of Validity of psychometric test, table adapted from (Messick, 1995).

Type of Evidence	Definition
Content-related	Relevance, representativeness, and technical quality of test content
Substantive	Theoretical rationales for the test and test responses
Structural	Fidelity of scoring structure to the structure of the construct measured by the test
Generalizability	Scores and interpretations generalize across groups, settings, and tasks
External	Convergent and divergent validity, criterion relevance, and applied utility
Consequential	Actual and potential consequences of test use, relating to sources of invalidity related to bias, fairness, and distributive justice

2.1.2 Clinical evaluation of the symptoms in HD

As underlined in the introduction, HD is characterized by a triad of progressive deficits: *cognitive*, *emotional–behavioral* (psychiatric), and *motor* symptoms. All these symptoms contribute to the deterioration of the *functional* capabilities to live a 'normal' life.

The measures of clinical symptoms is obtained through tests, that have to be scientifically *reliable* and *valid*. In this psychometric context, reliability indicates the degree to which a test is free from the measurement error (Strauss et al., 2006). The reliability can be decomposed into: (1) consistency within itself (internal reliability) (2) consistency over time assuming no damage in clinical condition (test-retest reliability) (3) consistency across different forms, such as pen/paper vs digital device (alternate form reliability), (4) consistency across raters (inter-rater reliability). Second, the validity of a test can be defined as to what extent a test actually measures what it is intended to measure (Strauss et al., 2006). Especially, the validity of a test needs to be understood in a specific context. Decisions taken from these tests need to take into consideration individuals characteristics, testing conditions or examinee willingness to cooperate. Messick (1995) proposed a set of six different types of evidence to construct the validity of a test. We summarize in Table 2.1 these six types of evidence, adapted from (Messick, 1995).

All the tests that are performed in the follow-up of HD need to be assessed for their reliability and validity through careful research studies. For instance, the cognitive battery, the HD-CAB (Stout et al., 2014), was validated for its reliability by repeating

tests across an interval so brief that it can not be associated with detectable disease progression. The tapping test validated its quality by its reflection of motor functions in (Bechtel et al., 2010) and external validity with the correlation with striatal volumes (external validity in HD).

Now, we describe in more details, the different tests that we used to follow up participants in our database. First, HD and preHD individuals were assessed through the Unified Huntington's Disease Rating Scale (UHDRS) (Kremer, Group, et al., 1996). The UHDRS is a complete battery assessing the motor function, cognitive function, behavioral abnormalities, and functional capacity, specifically designed to the HD population. The different assessed *motor* functions in HD are:

- Ocular pursuit
- Saccade initiation
- Saccade velocity
- Dysarthria
- Tongue protrusion
- Maximal dystonia
- Maximal chorea
- Retropulsion pull test
- Finger taps
- Pronate/Supinate-Hands
- Luria (fist-hand-palm test)
- Rigidity of the arms
- Bradykinesia of the body
- Gait
- Tandem walking

These motor functions are scored by neurologists that were additionally certified experts on Huntington's Disease. These certifications are renewed every year with virtual training and validation. All these dimensions are then assessed with integer scores on a scale from 0 to 4, with a score of 0 indicating no impairment up to 4 indicating the most severe impairment. All these scores are summed to form a global score that reflects the progression of the overall motor symptoms: the Total Motor Score (TMS). The maximum possible score for the TMS is 124. Of the various motor symptoms assessed, it is worth underlining that there is a dysarthria item. It is based *only* on the perception of the neurologist after discussions with the patient providing a scale from 0 to 4. This item lacks proper quantitative metric to track the different speech production aspects.

The *cognitive* part of the UHDRS includes verbal fluency tasks, the symbol digit modalities test (SDMT), and the 3 tests of the Stroop: word (SW), colour(SC) and interference (SI). The Stroop interference task is classic in Neuropsychology to test for executive functions, especially the inhibition capabilities of automatic behaviors.

We illustrate this Stroop effect in Figure 2.2. The measures of cognitive symptoms contrasts with the ones of motor symptoms. Indeed, these are not "pure" tests that measure a single aspect of the cognitive symptoms. The performance and success during these tasks engage different cognitive components. We review briefly the different components invoked in each cognitive test in Table 2.2.

Tab. 2.2.: Review of the cognitive components that are invoked during the cognitive tasks in Huntington's Disease follow-up. + represents the presence of a specific dimension in a test and − represents the relative absence.

Test	Verbal Fluency	SDMT	Stroop Word	Stroop Color	Stroop Interference
Cognitive component					
Processing speed	+	+	+	+	+
Executive function	+	+	−	−	+
Language	+	−	+	+	+

These cognitive tests require administration and post-interview quotation by a neuropsychologist. Even though there is current research effort to measure cognitive symptoms through digital device (Stout et al., 2014; Lunven et al., 2021; R. A. Barker et al., 2013), there is no digital or remote equivalent test yet and they remain with pen and paper.

Psychiatry measures and questionnaire in HD have evolved in the past 10 years (Kingma et al., 2008). Currently, behavioural problems are assessed with the Problem Behaviours Assessment (PBA) which consists of 36 items covering nearly all behavioural problems present in HD: depression, apathy, anxiety, irritability/aggression, obsessive-compulsive behaviors, and psychosis. Each score for the separate dimension are computed by multiplying severity and frequency scores. This yields individual score for each dimension. The PBA assesses behavioural problems that occurred in the 4 weeks prior to the interview. The PBA is a semi-structured interview designed specifically for the rating of the severity and the frequency of behavioral abnormalities in HD (Craufurd et al., 2001). This interview was carried by Neurologists in our case.

Finally, the *functional* status of the patients are measured through the Total Functional Capacity (TFC) and the independence scale (IS) which are both contained the UHDRS evaluation (Kremer, Group, et al., 1996). The TFC is single score on a 13-point scale that estimates the patient's proficiency in maintaining autonomy during daily activities. It is also a primary endpoint during clinical trials. These scales can only be delivered by HD certified neurologists and it is carried through a

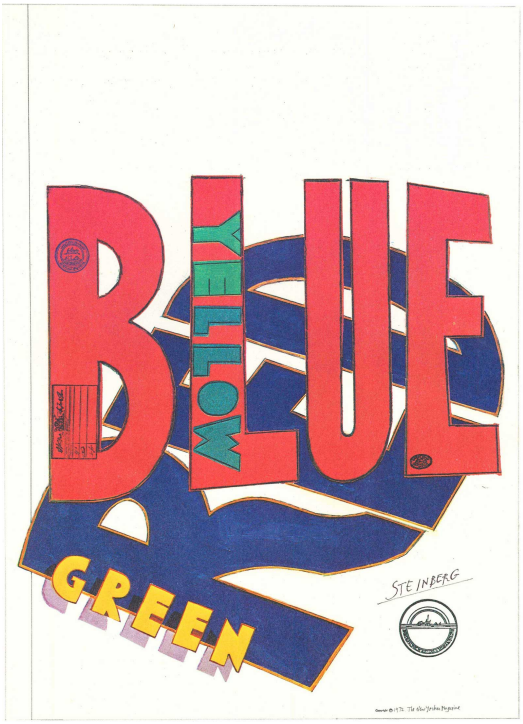


Fig. 2.2.: Illustration of the Stroop interference effect. Individuals have to inhibit the automatic reading overlearned during their school education, and denominate to color of the fonts. Saul Steinberg, Colors, 1971, © The Saul Steinberg Foundation, © photography Audrey Laurans.

questionnaire. The TFC has been partitioned into 5 stages that indicate the levels of disease severity of preHD and HD individuals.

- stage I: TFC scores from 11–13 (least severe)
- stage II: TFC scores from 7–10
- stage III: TFC scores from 3–6
- stage IV: TFC scores from 1–2
- stage V: TFC score of 0 (most severe)

However, the TFC does not evolve linearly with the years (Marder et al., 2000) and does not have the highest signal to noise ratio for the disease's measure. More recently, a multi-domain score, named composite UHDRS (cUHDRS), was proposed as a single endpoint of clinical trials in HD thanks to its greater sensitivity to disease progression (Schobel et al., 2017). The cUHDRS scale has been built to reflect the motor, cognitive, and functional declines in early HD. It relies on the combination of 4 tests of the UHDRS. The formula for the cUHDRS is:

$$\text{cUHDRS} = \left[\left(\frac{\text{TFC} - 10.4}{1.9}\right) - \left(\frac{\text{TMS} - 29.7}{14.9}\right) + \left(\frac{\text{SDMT} - 28.4}{11.3}\right) + \left(\frac{\text{SW} - 66.1}{20.1}\right)\right] + 10 \tag{2.1}$$

This formula was obtained through a Principal Component Analysis (PCA) performed on the TMS, TFC, SW and SDMT on the TRACK-HD cohort (Tabrizi et al., 2013). These clinical measures were selected for the PCA analysis as they were displayed the biggest signal-to-noise ratio for longitudinal follow-up. Authors decided to remove the apathy factor contained in the PBA even though it showed great signal-to-noise ratio. The apathy was an another component, almost independent of the other symptoms.

2.1.3 Interview protocol

Recording setup

There are different objectives for the interview protocol, and a number of practical requirements to fulfil. We wanted to obtain enough speech signal for each patient, and we were looking for speech that is clinically relevant to follow the progression of HD. However, participants in clinical study already go through a lot of different clinical tests, and it is important to keep the speech tests short and non-exhausting. In addition, the speech tests and equipment need to remain simple, that can be carried out even by non acoustic expert, especially to not increase even more the burden on the medical staff.

All participants completed a standardised speech battery, and they could interrupt the session at anytime. The interview protocol took place after neuropsychological assessments in the same room. To record participants, we decided to go with a light-weight speech equipment without an Anechoic chamber. This reduces the burden for this protocol so that the advances and findings here could be easily carried out by other neurology departments at a limited cost. There is no need of an acoustic engineer to install this setup. We provided for each neuropsychologist the equipment summarized in Figure 2.3. The main recording device is the ZOOM H4n Pro recorder and it has also two recording tracks. All the interviews are recorded at 44.1kHz with a 16-bit resolution.

All subjects were placed in the same position every time as well as the recording equipment. We illustrate the instructions for the placement in Figure 2.4. The

(a) ZOOM H4n Pro recorder

(b) Lavalier additional microphones

(c) Adapter with recorder

(d) Memory card SDHC

Fig. 2.3.: Recording device equipments used to record BasalVoice database.

recording setup was validated with Healthy controls, HD participants by the Neuropsychologists and Neurologists.

Speech tasks

The interviews were decomposed into several segments, where participants engage in different speech tasks. These speech tasks were presented to the participants in a fixed order. Speaking engage a number of cognitive resources, including working memory, to a different degree based on the situation. We refer in this thesis to the *cognitive load*, the mental demand placed by a task or situation to an individual. In our protocol, there are 13 tasks of increasing complexity, that modulate the linguistic target, the cognitive load, the emotional content, the topics and the materials of the discourse. We fixed the order of tasks for simplicity of administration and comparability between participants, at the cost of potentially impacting interpretability due interferences between tasks. For clarity, we made a distinction between the tasks with a fixed-vocabulary and the ones with an open-vocabulary.

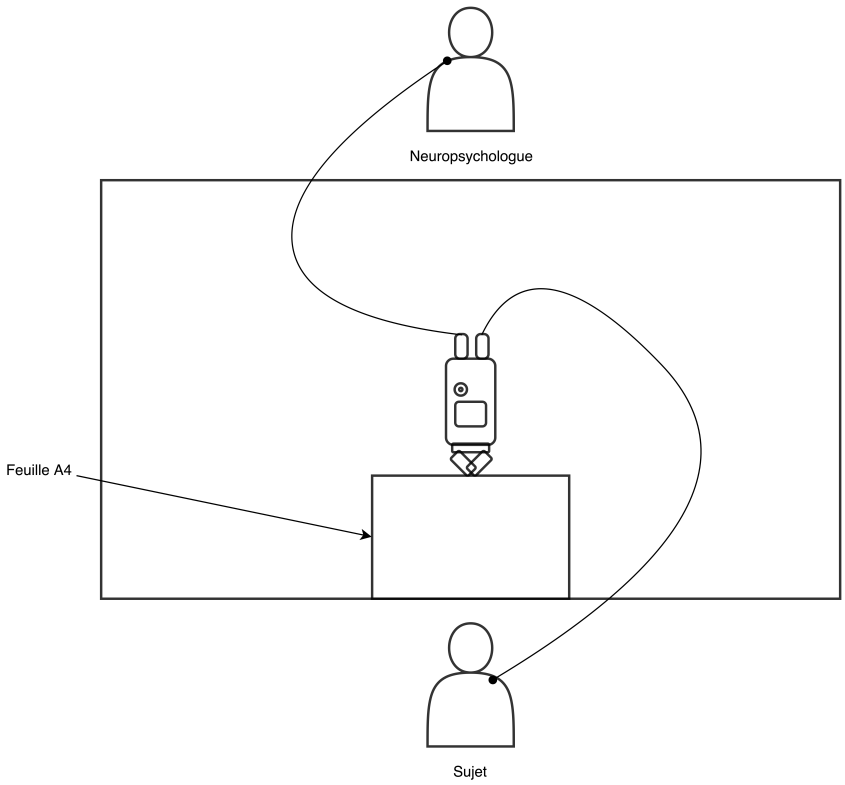


Fig. 2.4.: Placement and orientation of the recorder on the table during the interviews between Neuropsychologist and participants in BasalVoice. The main microphone is oriented towards the subject and placed at a fixed distance on the table.

Fixed-vocabulary tasks We denoted by fixed-vocabulary a speech task that has constraints concerning the vocabulary of words and phones to be produced. We summarize in table 2.3 the different components to perform each speech fixed-vocabulary task. We asked the participants to:

T0: /a/ Sustain the vowel /a/ for 2 seconds.

T1: Long /a/-/i/-/u/ Take a deep breath and sustain the vowel /a/ at a constant intensity and pitch level for as long as possible. Same for vowel /i/ and vowel /u/.

T2: heho ↗ Repeat louder and louder the two vowels "/e/ /ɔ/", by starting with a very low volume up to maximum possible: hé ho hé ho hé ho hé ho hé ho.

T3-T4: Count 1-20 ⇌ (1) count aloud numbers from 1 to 20, then (2) to count the numbers backwards from 20 to 1 while holding the hands up and closing eyes.

Tab. 2.3.: Approximate review of the invoked processing components involved in the generation of speech in the fixed-vocabulary speech tasks in the BasalVoice database. = represents minimal recruitment, + represents the presence of a specific dimension in a test and − represents the relative absence. We omit Syntax, morphology, pragmatics and the Social aspects from this Table.

Task	/a/	Long /a/-/i/-/u/	heho ↗	Count 1-20 ⇌	Months ⇌
Motor speech component					
Respiratory	+	+ +	+ +	+	+
Laryngeal	+	+	+	+	+
Vocal tract	=	=	+	+	+
Linguistic component					
Phonetics	=	=	=	+	+
Phonology	−	−	−	+	+
Semantics	−	−	−	+	+
Cognitive component					
Auditory control	=	=	+	+	+
Inhibition control	−	−	−	+	+
Task shift	−	−	−	+	+

T5-T6: Months ⇌ (1) Enumerate aloud the months of the year from January to December, then (2) to enumerate the months backwards from December to January while holding the hands up and closing eyes.

These tasks are used for different reasons: (1) a fixed-vocabulary task has more potential to be digitalized or being performed remotely via phone/zoom (Dorsey et al., 2018) (2) these tasks remain short and can test for specific aspects of speech productions (Rusz, Svihlik, et al., 2021). Indeed, simple short speech tasks can be easily transferred to other neurology departments without the need of linguistic experts and are easily translated into other languages. Especially, sustained phonation tasks have been used extensively in a number of NDDs in several languages: PD (Zhan et al., 2018; Rusz, Cmejla, et al., 2013; Tsanas et al., 2012; Bot et al., 2016) or in HD (Rusz, Saft, et al., 2014; Romana et al., 2020; Perez et al., 2021; Vogel et al., 2016). In our case, we also specifically tested respiration and auditory control during these fixed-vocabulary tasks with an increasing volume task. Besides, as shown by Vogel et al. (2012), the cognitive load could impact the performance of biomarkers extracted from speech. Therefore, sustained phonation could potentially not be sufficient, especially for dimensions of cognitive and functional symptoms. This is why, we also asked for simple fixed-vocabulary tasks with cognitive load. When participants perform the backward counting or the backward months, they need to inhibit the automatic forward recitation and disengage from the overlearned

forward sequence just previously performed. In addition, we used a dual-task (of holding hands and closing eyes) (Lo et al., 2020) which is known to increase reaction times and errors, especially when HD progresses (Mayr & Keele, 2000).

However, these tasks have some limitations. The control of these tasks removes higher social and linguistic spontaneous behaviors from speech productions. For instance, it will be hard to evaluate spontaneous emotions or conversational skills of individuals. Furthermore, there might be discrepancies in the true competence of individuals with HD during these Fixed-vocabulary tasks and the true competence of individuals in the real world (Sacks, 1998). The isolation and testing of specific components of speech production might "asepticize" the true nature of spoken language production. That is why we decided to evaluate and build the speech technologies for open-vocabulary speech tasks.

Open-vocabulary tasks Open-vocabulary tasks refer to speech tasks, where the exact words to be pronounced is an open set. We used a variety of open-vocabulary tasks, invoking different contexts and materials. In our tasks, spoken language processing components are tested with memory components, emotions, through story-telling, and visual description (as inspired by classic tests in aphasiology (Kaplan et al., 2001)). These tasks differ greatly from the fixed-vocabulary tasks. They are also referred to *connected speech* in the neurolinguistics literature (Boschi et al., 2017; Wilson et al., 2010). We tested the participants for the following speech tasks:

T8 Cookie Theft Can you describe what is happening in this image? (See image in Figure 2.5).

T9 24h Can you tell what you did in the past 24 hours?

T10 Sad story Can you tell a story that you find sad, really unhappy, deeply depressing?

T11 Angry story Can you tell a story that you find irritating, upsetting, a really annoying story?

T12 Happy story Can you tell a pleasant story, full of happiness and joy, something cheerful?

T13 Little Red Riding Hood Can you tell the story of Little Red Riding Hood that you remember? You can help yourself with these pictures which describe a version of the story (See images in Figure 2.6).

The picture description task requires minimal instructions and interactions with the Neuropsychologist, and it restricts the specific information and key elements. It can potentially help also the assessment of lexico-semantic impairments (Sajjadi et al., 2012). However, this task limits the syntactic structures used by the participants, as mostly short present tense statements are necessary to describe images. Besides, there is limited emotion and interaction involved with the Neuropsychologist during this speech task. We used the classic image used in the Cookie Theft of the Boston Diagnostic Aphasia Examination (See Figure 2.5).

To probe for memory capabilities in spontaneous language production, we also asked the participants to recall and tell their previous 24 hours. This provides more diverse syntactic structures than the image description but the scoring of specific items is harder. As for the image description task, there is practically an absence of emotion and limited interaction with the Neuropsychologist during this speech task.

These tasks do not cover at all the full spectrum of possible settings for conversations. Especially, stories are used extensively for human communication and we attach strong emotional attachments to story in books and movies that surround us during childhood (K. J. Alexander et al., 2001). Storytelling engages a number of cognitive and social processes such as Theory of Mind to imagine the point of view and to engage listeners (Symons et al., 2005). Stories are used in several neurological and psychiatric disorders to evaluate spoken language productions, such as in Autism (Kenan et al., 2019), in Alzheimer (MacWhinney, 2019), or Amyotrophic Lateral Sclerosis (Ash et al., 2014).

Even though stories and image descriptions can invoke social components to have a successful interaction, emotions are not necessarily spontaneous and can be acted. That is why we evaluated the emotional capabilities in speech production, we asked participants to share emotional stories, with three basic emotions that can be found across cultures (Ekman, 1999): anger, sadness and joy.

We summarize in table 2.4 the different components that are engaged in each speech open-vocabulary task. A detailed review for the evaluation of open-vocabulary speech tasks in neurodegenerative diseases can be found in (Boschi et al., 2017). For our database BasalVoice, we decided to not probe for reading tasks for two main reasons. First, reading capabilities and disabilities are greatly influenced by education and can become a confounding factor (Guez et al., 2021). Indeed, reading capabilities is highly explained by the socio-economic status of parents (Eriksen et al., 2013). It is of prime importance to develop markers that do not depend of such factors, in order to build inclusive healthcare (Wiens et al., 2019). The second argument to exclude reading tasks come from heterogeneous losses of neurons in

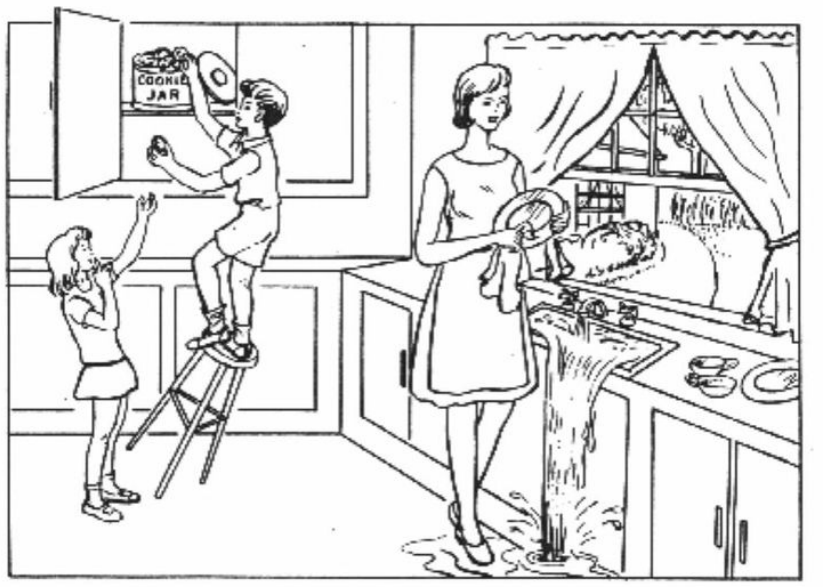


Fig. 2.5.: Image of the Cookie theft originally introduced for aphasiology studies (Kaplan et al., 2001)

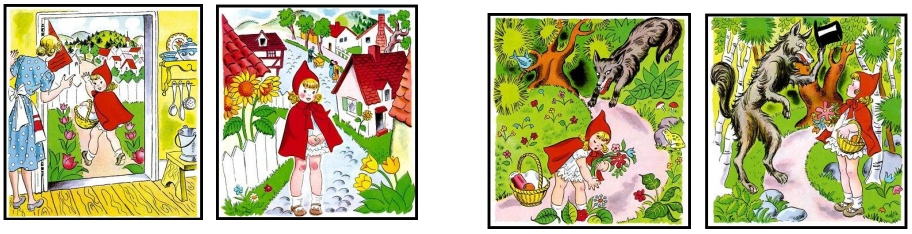


Fig. 2.6.: Sample images used for the story telling of the Little Red Riding Hood.

the Visual Cortex affecting patients with Huntington's Disease (Wiens et al., 2019), potentially affecting the reading capabilities.

As a reminder, participants in our study had the choice to interrupt and stop the interview, there is also possibility of errors in the tasks that are asked to perform by the interviewer. We do not know what interrupted the interview, but this information allowed us to measure the statistics of each speech task and the feasibility to carry them at the hospital (See Table 2.5). Each task has different administration times and different difficulty of completion. These information are important to take into account in case of deployment in the clinical setting, so that it is still easy to follow up of patients with speech tasks. The order of the tasks was fixed, so later tasks are also harder to complete as participant can be more exhausted. In the same

Tab. 2.4.: Approximate review of the invoked processing components involved in the generation of speech in the open-vocabulary speech tasks in the BasalVoice database. = represents minimal recruitment, + represents the presence of a specific dimension in a test and − represents the relative absence. We omit Motor speech components and Auditory control from this Table as they are always required.

Task	Cookie Theft	24h	Little Red Riding Hood	Emotional story
Linguistic component				
Phonetics-Phonology	+	+	+	+
Morphology-Syntax	=	=	+	+
Semantics	=	=	+	+
Pragmatics-Interaction	−	=	+	+
Cognitive/social component				
Memory	−	+	=	+
Theory of Mind	−	−	+	+
Emotion	−	−	=	+

Tab. 2.5.: Feasibility and timing of the speech tasks in BasalVoice database. We report results for 91 annotated interviews for preHD and HD individuals. # stands for 'number of'. sd stands for Standard deviation. Order refers to the position of each task during the clinical interview.

Task	Order	#Miss	%Miss	Administration times (s) Mean (sd) [min-max]
Fixed-vocabulary				
/a/	1	2.0	0.02	20.0 (14.0) [5.2-82.5]
Long /a/-/i/-/u/	2	3.0	0.03	63.9 (40.4) [13.3-205.1]
heho ↗	3	2.0	0.02	41.4 (19.5) [8.5-128.9]
Count 1-20 ⟶	4	2.0	0.02	17.0 (6.0) [7.1-42.0]
Count 20-1 ⟵	5	2.0	0.02	31.0 (12.9) [11.2-82.9]
Month ⟶	8	11.0	0.12	17.1 (8.2) [7.5-54.7]
Month ⟵	9	14.0	0.15	33.4 (19.9) [12.4-145.0]
Open-vocabulary				
24h	6	8.0	0.09	101.0 (38.3) [37.2-291.8]
Sad story	7	7.0	0.08	104.9 (59.2) [38.4-332.9]
Little Red Riding Hood	10	9.0	0.10	144.6 (52.9) [55.2-314.1]
Angry story	11	11.0	0.12	121.8 (98.3) [30.0-745.3]
Cookie Theft	12	11.0	0.12	70.9 (28.8) [24.1-162.0]
Happy story	13	10.0	0.11	101.0 (58.2) [29.2-309.3]
Overall stats	-	-	-	790.5 (309.3) [80.3-1811.4]

spirit as the HD-CAB cognitive assessment battery (Stout et al., 2014), we report the time and percentage of assessments per task for the annotated interviews (See

Table 2.5). The patient had the right to stop the interview at any moment. The neuropsychologist could decide to interrupt and not finish all the speech tasks based on the unfolding of the interview.

We did observe that the first tasks of the interview were easier to perform. Besides, there is an important variability of administration times across tasks, and large standard deviations for each single task, except for the automatic sequences of speech (Numbers and Months). The months tasks are placed inside the open-vocabulary tasks, therefore there is a higher missing rate than for other fixed-vocabulary tasks. It seems that the backward recitation of months is forgotten during the tasks by the interviewer. For the open-vocabulary tasks, story telling can take way more time than simple image description. These statistics will inform and guide the future potential choices to disseminate these tests in the clinical practice.

Audio data circuit The data in the newly introduced BasalVoice is highly sensitive, contains demographics of individuals with and without HD and needs to be protected. Indeed, even though the data we received and compiled in the BasalVoice database is anonymized, critical information are contained in the meta data (birth, profession, gender, children information) and in the voice data itself. Open data represents unique opportunity for research, but adversarial attacks and breaches can happen if the data sharing is not done carefully. For instance, to speed up research progress in Medicine and Healthcare, the Australian government decided to share anonymous clinical data with metadata of millions of its citizens. However, it was shown by researchers at the University of Melbourne (Culnane et al., 2017) that the combinations of meta data is almost unique (with number of children, dates of birth) and this allows the de-anonymization of clinical data by the cross-over with public databases such as Wikipedia. Data of government officials could be easily re-identified. This example underlines the utter importance to really think ahead before releasing clinical datasets, to avoid the release of personal information of participants.

In our case, the participation agreement for BasalVoice contained a commitment to not release the data to other researchers. Open secured sharing methods will be the object of future studies to speed up research on HD. There were a number of security requirements to fulfill to build the BasalVoice database. The participant data are owned by the *Assistance Publique - Hôpitaux de Paris* (APHP) and its access is reserved for the professionals of the AP-HP care team and the managers of the administrative, logistics and IT services of the AP-HP. The data can only be used for research purposes and the data can not be moved or stored outside the European

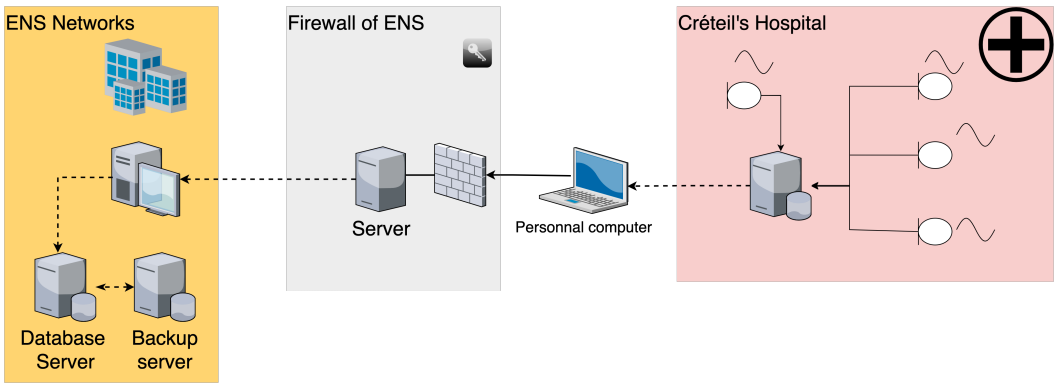


Fig. 2.7.: Circuit of audio data from the microphones to secured database server at ENS. Dashed lines represent connection without extra internet connections. Only authorized devices can connect to the firewall server of ENS.

Union or to other recipients without explicit prior consent. Besides, the data cannot circulate through the internet without proper encryption and has to land on secure servers with pre-defined accesses.

Therefore, we designed a specific system for the circuit of the audio data, illustrated in Figure 2.7. The raw audio data is first collected into 4 microphones that are given to neuropsychologists at the Créteil's hospital. We manually extract on personal computer without internet connection the audio data. Then, this data is forwarded to a Firewall server where accreditation is required to connect. The audio data is then copied to the secured server of ENS where only the researchers associated with the project can access the BasalVoice database. This strategy allowed us to avoid the storage and connection on servers outside the European Union. Besides, the connection to the ENS networks is protected and access rights are handled as required by the APHP.

In a future iteration of the BasalVoice database, the evolution of the circuit of the audio data is required. Indeed, the current system still requires manual extraction of data from the Hospital of Créteil and can not be scaled through multiple hospitals and services.

2.2 Annotation of spontaneous dysarthric speech

In this section, we summarize how we annotated the audio data (section 2.2.1), how we evaluated inter-annotator agreements for different levels of annotations for our database (section 2.2.2), and finally how we handled cohorts of annotators to ensure the database quality and speed up research process (section 2.2.3)

A large part of the work in this section has been done in close collaboration with Xuan Nga Cao and Hadrien Titeux.

2.2.1 Annotation and coding protocol

The collection of audio data can be easily performed thanks to recent advances concerning the recording devices. In a short amount of time, huge amount of data of naturalistic interactions can be collected. Yet, the collection of raw audio data is only the first step to obtain relevant, cheap and reproducible clinical biomarkers. To analyze conversational interactions, it is necessary to segment speech in terms of turns (who speaks when?), transcribe them in terms of words (what is said?) and annotate these transcripts in terms of high level linguistic marks (speech acts, word errors or disfluencies). Segmentation in turns and automatic transcription are not solved in terms of engineering for spontaneous (Goldwater et al., 2010; Riviere et al., 2021) and/or dysarthric speech (Rudzicz et al., 2012; Christensen et al., 2012). Part of the problem of current engineering systems is the scarcity and lack of good quality training data. In addition, the coding of complex phenomena is even more difficult when working from the raw waveform. This is why before developing speech processing algorithms for these new applications, we designed and implemented a specific annotation protocol, that we coined SpontaneousCHAT.

Usually, choices for an annotation protocol and its specifications are guided by three main competing goals:

Clarity All the symbols should have clear definition and have a real-world phenomena referent. The nomenclature of annotations should be consistent. There is a need for a certain balance between too coarse or too fine schema to obtain clarity.

Readability Annotation outputs should be easy to read. However, the concept of "reading" is not clear in the case of annotations. Some linguists might just want to obtain summary statistics from their annotations, or on the other side of the spectrum, each symbol should be digested automatically by a program and can be parsed in a formal manner.

Ease of data entry If an annotation protocol becomes too complex, too exhaustive, it becomes quickly very time-consuming and error-prone. There are two to alleviate this problem: (1) sacrifice some categories, and simplify the schema or (2) help with semi-automatic tools to check for annotations, or even include some automatic help.

As stressed by MacWhinney (2019), these 3 goals are competing and tradeoffs have to be made for structure of annotations, and this can be compared to the pressures that shape human languages themselves. The main difference with language is the presence of two very different audiences. On the first hand, the protocol is built to be understood and usable by an human audience of transcribers, field linguists or speech pathologists. On the other hand, the other audience is linguists, scientists, their computers and their programs.

For instance, van Gompel and Reynaert (2014) add other specifications for a schema of annotation from a computational perspective, and we list the main ones below:

Expressivity The annotation schema needs to contain enough expressiveness to capture a wide variety of linguistic phenomena, especially in spontaneous and/or dysarthric speech in our case.

Openness The annotation schema can or not commit to a specific linguistic theory during its design.

Extensibility The extensibility is the capability of the annotation format to add novel items or new phenomena not included in the first place.

A number of annotation format and protocols exist to deal with spontaneous speech and conversations: NXT-XML (Carletta et al., 2005), ODIL Syntax (Wang et al., 2020), CHAT (MacWhinney, 2019), FoLia (van Gompel & Reynaert, 2014), or GARS for French (Blanche-Benveniste et al., 1990). These annotation protocol had to negotiate during their design in the different aforementioned dimensions.

For instance, FoLia (van Gompel & Reynaert, 2014) is completely *parsable* by a computer program, however it came at the huge cost to be complicated to annotate by humans. ODIL Syntax (Wang et al., 2020) adopts a syntax approach to annotate sentences from spontaneous speech, and can be used to construct syntactic tree. Even though it is also parsable and relatively easy to annotate, only a few of spontaneous events are taken into account (revision, false starts, repetitions) and adopts the theoretical view of how disfluency occurs introduced by E. Shriberg (Shriberg, 1994). This limits the *expressivity* and *openness* of this schema of annotations and the difficulty to test alternative theories for spontaneous and disfluent speech. Besides, as we are not making assumptions concerning the way HD individuals talk, we did not want to commit to a theory that did not take into account a large number of speech phenomena.

The CHAT protocol fulfills a large number of requirements to annotate our BasalVoice database. The CHAT protocol is very exhaustive and *expressive* in terms of the number

of phenomena. The CHAT protocol has been heavily tested against multiple domain of conversations (child language acquisition, clinical database, second language database) in multiple languages (English, French, Spanish, Chinese). That is why CHAT kept growing over the years, and can be easily learned by an annotator. The CHAT protocol is not associated with a specific linguistic theory. Even though there is a broad classification among the different vocalizations, the transcriptions and analyses in CHAT take almost no assumption concerning the classification system linking the different vocalizations.

Yet, there are 3 slight limitations to the CHAT protocol: (1) it suffers from some lack of consistency and especially outputs can not be parsed in a programmatic way, (2) in the current setup of CHAT, only the software Elan is used for annotations, (3) complex speech phenomena (such as repetitions with variable insertions or multiple-level disfluencies that occur often in natural and disordered speech) can not be annotated.

To overcome these issues, we developed our own protocol of annotations SpontaneousCHAT that is derived from the CHAT protocol (MacWhinney, 2019). We focus on the utterance level [1] in this thesis. We did not include interactive/multiple utterance annotations , speech acts, morphology or syntax in the first iteration of SpontaneousCHAT. In Table 2.6 we summarize the annotation schema that we provided to the annotators, the SpontaneousCHAT protocol. In this table, we illustrate with real examples each phenomenon.

Tab. 2.6.: SpontaneousCHAT annotations schema used to annotate spontaneous face-to-face communications with and without dysarthria. All examples are drawn from real observations in the BasalVoice database.

Annotation	Name and explanation	Examples
Correct spoken words	Words that are correctly pronounced and in the vocabulary	"**donc je suis retourné à mon travail** &-euh **donc dans mon bureau** &-euh "
&+	Phonological Fragment: These are pronounced phones that are not recognisable as a known word	"on s'est tout d(e) suite couchés après donc vers **&+d** vingt-trois heures"

Continued on next page

[1] We use interchangeably *sentence* and *utterance*.

Tab. 2.6 – continued from previous page

Annotation	Name and explanation	Examples
&-{mmh, euh, em, eh, oh, n, hein, bein, bon, ah, bah, fin, ouais, mouais, alors, p}	Filler or Filled Pause	"donc j'ai fait encore différents contrats jusqu'à **&-euh** &=respiration seize heures"
&={vn, respiration, respiration+vn, coughs, sighs, laughs, sneezes, cries, groans, belches, gasps, hisses, hums, moans, mumbles, pants, grunts, sings, squeals, whines, whistles, whimpers, yawns, yells, sniffs, growls, curses}	Non-linguistic vocalizations: vn stands for vocal noises	"(mon) père m'avait préparé donc &-euh **&=respiration+vn** des tripes **&=laughs** avec des crêpres"
?	Interrogation point	"que(l)qu(e) chose que j(e) trouve triste**?**"
:	Prolongation of a segment	"y a pas grand chose **de:** vraiment gai après &-euh"
^	Pause within a word: It is necessarily after the first phone	"**pe^tit** chaperon rou:ge qui est envoyé par sa mère:"
$\neq$	Block at the beginning of a word	"y a pas de lumière là bas que $\neq$**c'est** plein d(e) trous"
$\hookleftarrow\rho$	Repetition of a segment	"**l**$\hookleftarrow\rho$ **l**$\hookleftarrow\rho$**le** fait qu'il ait donné sa vie justement pour permettre à:"
(1) Single word $< w >$ $[\times N]$ (2) Multiple words $< w_1 w_2 > [\times N]$	Repetition of a word w or group of words w_1, w_2 N times. Repetitions can become complex	(1) "Mais <**c'est**> $[\times 2]$ tout c(e) qui se passe" (2) "qu'est-ce qu'(il) y a eu <**de**> $[\times 4]$ drôle"

Continued on next page

Tab. 2.6 – continued from previous page

Annotation	Name and explanation	Examples
(3) Complex repetition $<w+a_1,...,w+a_N>[\times N]$, a_i can be a Phonological fragment, Filler, or Non-linguistic vocalization.	with insertions of fillers or other vocalisations between the repetitions.	(3) "< <**on: &=respiration , on**> [×2] a:> [//] j'ai pris une douche"
<**word or group of words**> [//]	Revisions/Retracings: the speaker decides to revise this part	"<**j'ai rien à**> [//] (il) (n')y a rien à di [: dire] [*p]"
<**words or group of words**> [_]	Perceptual abnormal prosody on a word or group of words	"Ce s(e)rait i(l) <**faut sauver**> [_] l(e) soldat Ryan"
xxx	unintelligible word, even with context	"c'est vrai que: oui l'accent est pas &+f **xxx**"
w(o)rd or (word)	Elision: Shortening of the word or Omission of a segment of the word. The word can be totally omitted.	"**(v)oilà (il n')** y a rien à dire puis Line Renaud aussi **qu(i)** a joué dedans"
+...	Trailing off. The sentence is abandoned.	"franchement c'(é)tait sympa moi je **+...**"
+..?	Trailing off of a question. If a question sentence is abandoned.	∅
+//.	Self-interruption. It is different from the Trailing off as it is not a incompletion but more a self-interruption. Material should be followed afterwards.	∅
$p_1p_2...p_m$ [: target word] [*p] or $p_1p_2...p_m$ [: target group of words] [*p] where p_j are the actual pronounced phones	Phonological errors	**pogwesifu [: progressif] [*p]** or **SHi [: je suis] [*p]**

Continued on next page

Tab. 2.6 – continued from previous page

Annotation	Name and explanation	Examples
pronounced word [: target word] [*s]	Semantic errors	**parlait [: passait] [*s]**
$p_1p_2...p_m$ [: target word] [*m] or $p_1p_2...p_m$ [: target group of words] [*m]where p_j are the actual pronounced phones	Morphological errors	**Croivent [: croient] [*m]**
$p_1p_2...p_m$ [: target word] [*n] or $p_1p_2...p_m$ [: target group of words] [*n] where p_j are the actual pronounced phones	Neologism	**bRavityd [: bravoure] [*n]**

What are the differences between our modified protocol SpontaneousCHAT with the CHAT protocol of B. MacWhinney? Our modifications brought to CHAT can be grouped into two main categories: (1) *systematization* and formalization of the protocol and (2) *augmentation* of the set of annotations.

First, as in the CHAT protocol, there is a lack of consistency and clarity for a number of phenomena, we had to enforce some choices and clarify some phenomena. We introduced the use of a well-defined phonetic alphabet in the protocol. Indeed, CHAT is still using grapheme to code phonological errors or Neologisms. CHAT leaves for subsequent analyses the transformations of these graphemes into phones, which is subject to interpretation and ambiguity. The SAMPA alphabet Gibbon et al., 1997 was used to specify the phones that are actually pronounced when there are phonological fragments, phonetic errors, neologisms. Here, we asked the speech pathologists to code phonetically these types of speech errors. We broughtchanges for the repetition of words for clarity: (1) we prohibited the use of [/] as symbol to code for a single repetition and only allowed for the coding with multiplier to be consistent [xN], we forced the use of brackets to specify if a single word or multiple words are repeated.

Besides, the CHAT was not homogeneous concerning the use of brackets <>, hooks [] and their exact relative positions. CHAT was designed to include as many naturalistic phenomena as possible and grew over time. Yet, it lost the potential to be perfectly *parsable*. In our modified protocol SpontaneousCHAT, we enforced speech pathologists to follow the pre-defined patterns thanks to regular expressions.

To do so we used regular expressions, to improve the CHAT protocol. In Table A.1 in Appendix, we display the regular expressions and the values we used to parse the annotations from our modified CHAT protocol.

These parsing methodologies allows us to construct parsed tree $\mathcal{T}$ of the vocalizations in a given stretch. Examples of these trees are shown in Figure 2.8.

```
SpontaneousSpeechStretch
 +-- [0] Filler : "&-bah"
 +-- [1] Word : "voilà"
 +-- [2] PhonologicalDistortion
 |   +-- CorrectForm
 |   |   +-- [0] Word : "je"
 |   |   +-- [1] Word : "sais"
 |   +-- DistortedForm
 |       +-- PhonemizedWord : "Se"
 +-- [3] Word : "pas"
 +-- [4] Word : "ça"
 +-- [5] Interruption : "+..."
```

```
SpontaneousSpeechStretch
 +-- [0] Word : "voulant"
 +-- [1] Word : "manger"
 +-- [2] Revision
 |   +-- [0] Word : "des"
 |   +-- [1] Word : "b(onbons)"
 +-- [3] Word : "des"
 +-- [4] Word : "gâteaux"
```

Fig. 2.8.: Example of parsed trees $\mathcal{T}$ from participants in our study. This is based on the Regex rules introduced in Table A.1

On the other hand, we also augmented the CHAT protocol, to take into account recurrent phenomena in dysarthric speech. We added the possibility to specify vocal noises, in vocalizations, and also the possibility to annotate words or group of words with abnormal prosody. We also include the possibility to code for complex repetition with small addition of phonological fragments, filler or non-linguistic vocalization.

Our derived SpontaneousCHAT protocol concerns the annotation of the speech and vocalization events during interviews, i.e. what is said during by the participant. However, this protocol does not inform about the exact timing and alignment of these events within the audio signal. There is no information about the turn-takings of the neuropsychologist and interviewee. To complete our speech annotation protocol, the speech pathologists had to also locate in time the speech events. However, a trade-off between speed of annotation/painful annotation and granularity of timings of information. On the first hand, we could ask the speech pathologists to locate in the time all the levels of the events: 'phonetic' level, 'word' level, 'stretch' level and 'sentence' level. These events could be both annotated for the participant interviewee and the neuropsychologists. In addition, to make specific task analyses, it is important to know the boundaries of the task. However, such granularity of segmentation of events and turn-takings comes at the huge cost to annotate only a small subset of the BasalVoice and reduce the *ease-of-data entry*. We decided to annotate the interviews with our SpontaneousCHAT protocol only the turn-takings of the interviewee, the boundaries of uninterrupted stretch of speech, and the sentence boundaries. In addition, we annotated the information concerning the turn-takings

of the neuropsychologist, and also ambient noises. In summary, speech pathologist had 4 tiers to annotate: the 'Task' level, the 'Patient' level, the 'Sentence' level and the 'Non-Patient' level. We show in Figure 2.9 a portion of an annotation in the Praat software with the different levels of annotations.

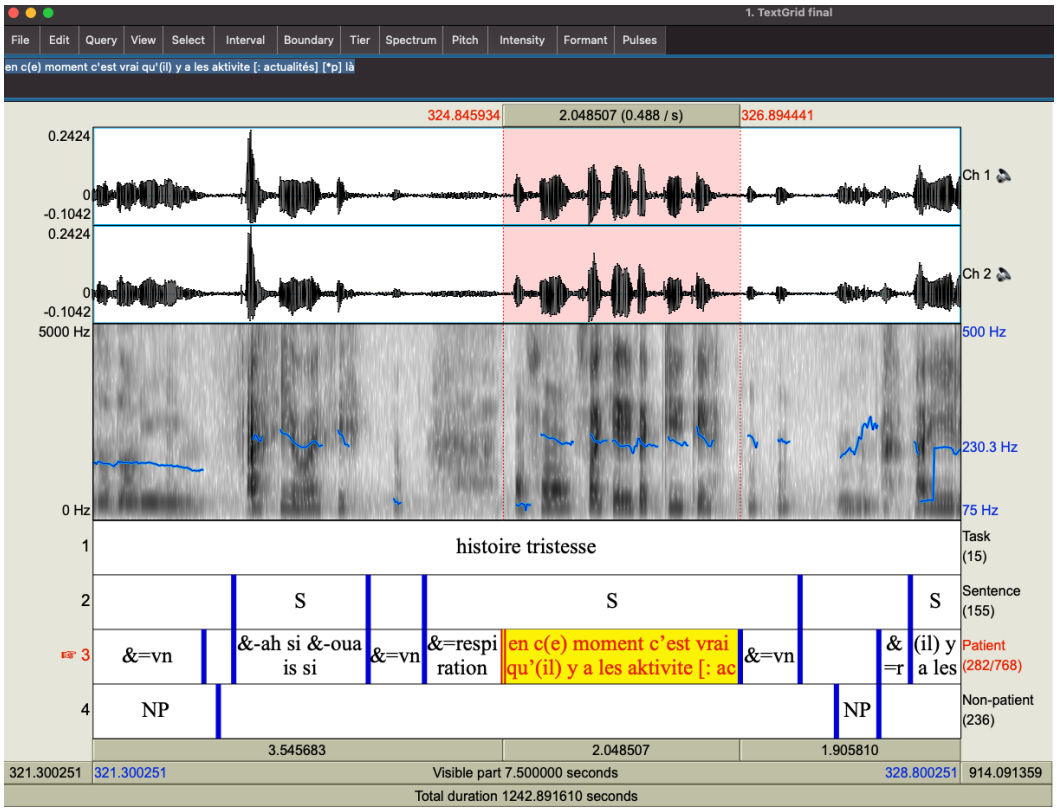


Fig. 2.9.: Example of a portion of a final annotation of in Praat. A patient stretch is highlighted.

Thanks to our annotation protocol SpontaneousCHAT, we have the phonetic information when a words is badly pronounced (elision, phonological error, semantic error, morphological error and neologism), with the actual phonetic pronunciation. Therefore, to obtain phonetic information, we made the following assumptions concerning the correctly pronounced words. We used the French metropolitan phonetization of words, text-to-phone (Bernard & Titeux, 2021). Based on the outputs of the text-to-phone and timing information at the stretch level, we force-aligned all the phones to obtain phone timings. We show in Figure 2.10 the same portion of interview as Figure 2.9, with the addition of the phonetic alignment of the patient speech.

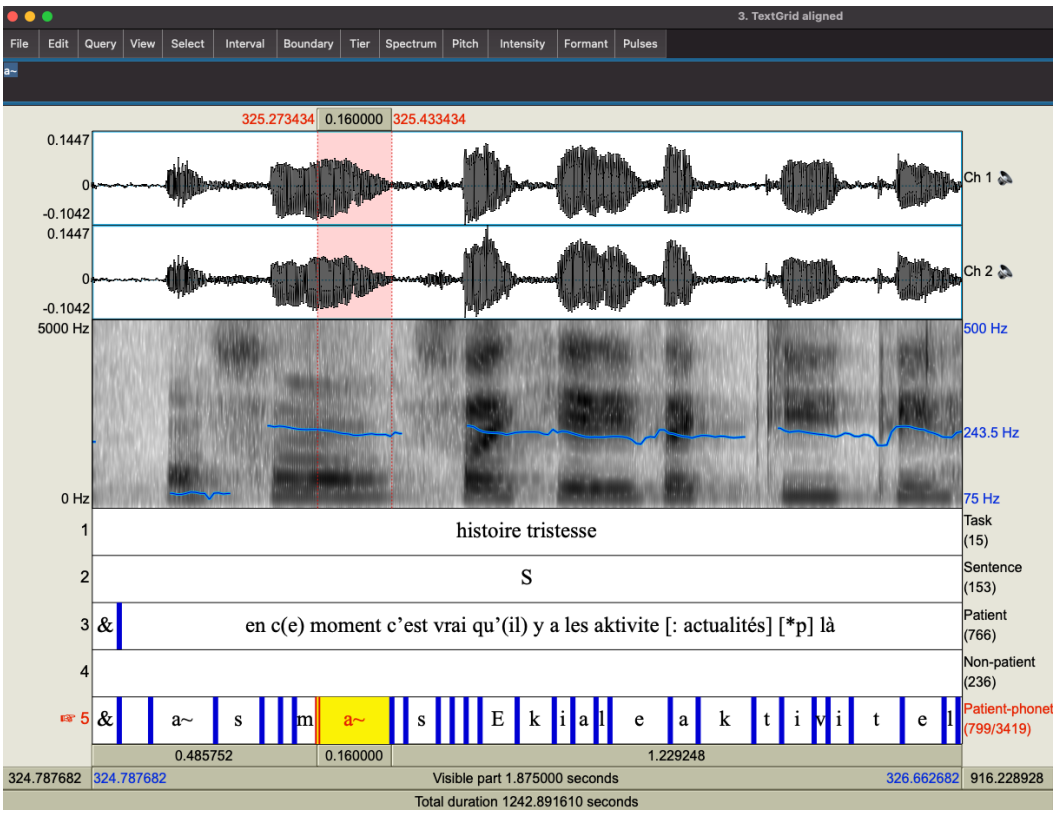


Fig. 2.10.: Example of a portion of a derived final annotation of in Praat. A phone timings is highlighted. This is a zoomed version of the portion of annotation of the shown interview in Figure 2.9. Here, we added the phonetic alignment of the interviewee's speech.

2.2.2 Annotation quality/reliability of BasalVoice

Even though, we improved the CHAT annotation protocol, we need to ensure that what is annotated by speech pathologist is correct. However, there is no perfect oracle concerning speech and vocalization annotations, there is no 'gold' or 'perfect' annotation available to measure the annotation quality produced by speech pathologists. Therefore, the strategy to measure the quality of annotation is to give the same annotation task to different speech pathologists and measure their (dis)agreements. Most coefficients (ex: α, κ) to measure agreement in psychology and natural sciences focus on the *categorization* of events.

Yet, the annotations of speech, and especially spontaneous speech represent a complex continuous phenomenon to annotate. There is not only categorization but also the localization of events, referred as *unitizing* (Krippendorff, 2018). The measure of agreement for speech conversations needs to take into account specifics of the turn-takings:

1. Sporadicity: Turn do not cover the full timeline
2. Embedding: A sentence can be decomposed into several turns.
3. Free overlap: Two different turns can overlap in time.
4. Free positioning: Turn can start and end at any time.
5. Categorization: Turn are categorized on the base of the speaker who pronounced them.

We illustrate all these characteristics in the Figure 2.11 with a real excerpt of turn-takings. In this Figure, we show two parallel annotations of the same interview of the turn-takings between the neuropsychologist and a participant interviewee. Sentence boundary annotations have similar characteristics as the turn-takings to compute agreements.

To the best of our knowledge, the only measure that fulfills these requirements is the γ agreement introduced in (Mathet et al., 2015). The authors provided a Java freeware graphical user interface for the γ agreement, which prevents a programmatical usage. In addition, the lack of open-source implementation prevents proper comparison of the algorithms and potential extensions. To fill this gap concerning the γ agreement measure, we introduce the `pygamma-agreement` a toolkit for the implement of the metric that is free and open-source (Titeux & Riad, 2021). We included the research article for this contribution that has been published in the Journal of Open Source Software.

The Journal of Open Source Software

pygamma-agreement: Gamma γ measure for inter/intra-annotator agreement in Python

Hadrien Titeux[1] **and Rachid Riad**[1, 2]

1 LSCP/ENS/CNRS/EHESS/INRIA/PSL Research University, Paris, France **2** NPI/ENS/INSERM/UPEC/PSL Research University, Créteil, France

DOI: 10.21105/joss.02989

Software

- Review
- Repository
- Archive

Editor: Arfon Smith

Reviewers:

- @faroit
- @apiad

Submitted: 25 November 2020
Published: 12 June 2021

Introduction

Over the last few decades, it has become easier to collect large audio recordings in naturalistic conditions and large corpora of text from the Internet. This broadens the scope of questions that can be addressed in speech and language research.

Scientists need to challenge their hypotheses and quantify the observed phenomenons of speech and language; this is why researchers add different layers of annotations on top of speech and text data. Some types of human intervention are used to reliably describe events contained in the corpus's content (e.g., Wikipedia articles, conversations, child babbling, animal vocalizations, or even just environmental sounds). These events can either be tagged at a particular point in time, or over a period of time. It is also commonplace to provide a categorical annotation or - in the case of speech - even precise transcriptions (Serratrice, 2000) for these events. Depending on the difficulty of the annotation task and the eventual expertise of the annotators, the annotations they produce can include a certain degree of interpretation. A common strategy when building annotated corpora is to have small parts of a corpus annotated by several annotators to be able quantify their consensus on that reduced subset of the corpus. If that consensus is deemed robust (i.e., agreement is high), we infer that the annotation task is well defined, less prone to interpretation, and that annotations that cover the rest of the corpus are reliable (Gwet, 2012). An objective measure of the agreement (and subsequent disagreement) between annotators is thus desirable.

Statement of Need

The Gamma (γ) Inter-Annotator Agreement Measure was proposed by Mathet et al. (2015) as a way to quantify inter-rater agreement for sequences of annotations. The γ-agreement measure solves a number of the shortcomings of other pre-existing measures. This quantification will have to satisfy some constraints: segmentation, unitizing, categorization, weighted categorization and the support for any number of annotators. They should also provide a chance-corrected value. Other measures, such as the κ (Carletta, 1996) or Krippendorff's α's (Krippendorff, 2011), have existed for some time to deal with these constraints, but cannot address all of them at once. A detailed comparison between metrics is available in Mathet et al. (2015).

To solve all of these constraints at once, the γ-agreement works in three steps: 1) a *disorder* (i.e., a cost) is computed for each potential alignments between the different annotators' units, using an annotation-dependent *dissimilarity* (akin to a distance between units). This *disorder* models the disagreement between annotators. 2) Using a convex optimization algorithm, a global alignment with the lowest total disorder is found. 3) By sampling random and synthetic

annotations from the original annotation and computing their own disorders, the original annotation's disorder value is chance-corrected to provide the final γ-agreement measure. The authors of Mathet et al. (2015) provided a Java freeware GUI implementation along with their paper and this implementation has already been used by some researchers (Da San Martino et al., 2019) to compute an inter-rater agreement measure on their annotations.

However, linguist and automated speech researchers today use analysis pipelines that are either Python or shell scripts. To this day, no open-source implementation allows for the γ-agreement to be computed in a programmatical way, and researchers that are already proficient in Python and willing to automate their work might be hindered by the graphical nature of the original Java implementation. Moreover, the original γ-agreement algorithm has several parameters that are determinant in its computation and cannot be configured as of now. For this reason, it would greatly benefit the speech and linguistic scientific community if a fully open-source Python implementation of the original algorithm was available – this is what we are making availble here. We have made sure that our implementation has several key features:

- It is comparatively as fast as the original implementation, taking about 10s to compute a high-confidence γ-agreement measure on a middle-range processor.
- The `pygamma-agreement` package is modular and users can easily extend one of the modules without the burden of rebuilding an optimized code base from scratch (e.g., Users can easily add a new dissimilarity measure).
- Our code allows for fine-grained constructions of an annotation Continuum and an advanced configurability of the γ-agreement's different parameters.
- We also made sure the interpretability of the input (i.e., the annotation continuum) and the output (i.e., the alignments) data structures are straightforward, as both can be visualized in a Jupyter Notebook (see Figure 1).
- Finally, we support most of the commonly used data formats and analysis pipelines in the speech and linguistic fields.

The pygamma-agreement Package

The `pygamma-agreement` package provides users with two ways to compute the γ-agreement for a corpus of annotations. The first is to use the package's Python API.

```
import pygamma_agreement as pa
continuum = pa.Continuum.from_csv("data/PaulAlexSuzann.csv")
dissimilarity = pa.CombinedCategoricalDissimilarity(categories=list(continuum.cate
gamma_results = continuum.compute_gamma(dissimilarity, precision_level=0.02)
print(f"Gamma is {gamma_results.gamma}")
```

The most important primitives from our API (the `Continuum` Figure 1 and `Alignment` Figure 2 classes) can be displayed using the `matplotlib.pyplot` backend if the user is working in a Jupyter notebook.

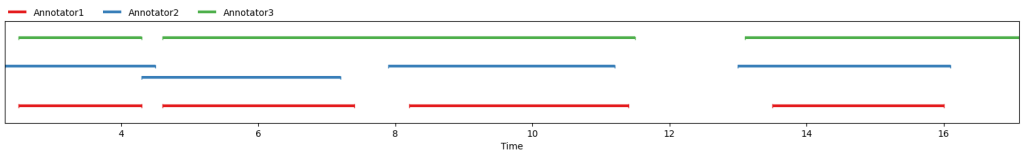


Figure 1: Displaying a Continuum in a jupyter notebook. This is a temporally accurate representation of the annotated data.

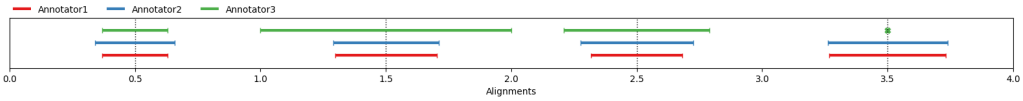


Figure 2: Displaying an Alignment in a jupyter notebook. This is a visual and schematic representation of the alignment computed between annotations of the original continuum (the order of units is respected but the annotations are scaled for visual clarity).

The second is a command-line application that can be invoked directly from the shell, for those who prefer to use shell scripts for corpus processing:

```
pygamma-agreement corpus/*.csv --confidence_level 0.02 --output_csv results.csv
```

We support a variety of commonly used annotation formats among speech researchers and linguists: RTTM, ELAN, TextGrid, CSV and `pyannote.core.Annotation` objects.

Computing the γ-agreement requires both array manipulation and the solving of multiple optimization problem formulated as Mixed-Integer Programming (MIP) problems. We thus used the *de facto* standard for all of our basic array operations, NumPy (Harris et al., 2020). Since some parts of the algorithm are fairly demanding, we made sure that these parts were heavily optimized using `numba` (Lam et al., 2015). We used `cvxpy`'s (Diamond & Boyd, 2016) MIP-solving framework to solve the optimization problem. For time-based annotations, we rely on primitives from `pyannote.core` (Bredin et al., 2020). We made sure that it is robustly tested using the widely-adopted `pytest` testing framework. We also made sure that `pygamma-agreement`'s outputs matched both the theoretical values from the original paper and those of the Java freeware. Travis CI is used to run tests to ensure package quality is maintained and most of the code is type-hinted and has descriptive docstrings, both of which can be leveraged by IDEs to ease the use of the API.

We provide a user documentation as well as an example Jupyter notebook in the package's repository. Additionally, we have used and tested `pygamma-agreement` in conjunction with the development of our own custom-built annotation platform, Seshat (Titeux et al., 2020). In Table 1, we present two use cases for our implementation of the γ-agreement measure on two corpora. These two corpora, ranging from medical interviews to child recording, allowed us to evaluate the performance of the γ-agreement on a wide panel of annotation types.

Table 1: γ Inter-rater agreement for clinical interviews (16 samples) and child-centered day-long recordings (20 samples).

Corpus	Annotation	# Classes	Mean of γ
Clinical Interviews	Turn-Takings	3	0.64
Clinical Interviews	Utterances	1	0.61
Child Recordings	Speech Activity	1	0.46
Child Recordings	Child/Adult-directed speech	2	0.27

The package is published in PyPi, thus, `pygamma-agreement` can be installed using pip.

Future Work

This implementation of the γ-agreement opens the path for a number of potential extensions:

- An obvious improvement is to add support for the "γ-cat" metric, a complement measure

(Mathet, 2017) for the γ-agreement.

- The γ-agreement's theoretical framework allows for some useful improvements such as:
  - The implementation of new dissimilarities, such as a sequence-based dissimilarity (based on the Levenshtein distance), an ordinal dissimilarity (for ordered sets of categories) and a scalar dissimilarity.
  - For categorical annotations, the support for an undefined set of categories, with annotators using different sets of categories. This would be solved using an adapted implementation of the Hungarian Algorithm. This could be useful in unconstrained diarization annotation tasks.
  - More hypothetically, the possibility to have so-called "soft alignments," where a unit has weighted alignments with other units in its continuum.

Acknowledgements

We are thankful to Yann Mathet for his help on understanding his work on the γ-agreement. We also thank Xuan-Nga Cao, Anne-Catherine Bachoux-Lévy and Emmanuel Dupoux for their advice, as well as Julien Karadayi for helpful discussions and feedback. This work is funded in part by the Agence Nationale pour la Recherche (ANR-17-EURE-0017Frontcog, ANR-10-IDEX-0001-02 PSL*, ANR-19-P3IA-0001PRAIRIE 3IA Institute) and Grants from Neuratris, from Facebook AI Research (Research Gift), Google (Faculty Research Award), Microsoft Research (Azure Credits and Grant), and Amazon Web Service (AWS Research Credits).

References

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., & Gill, M. (2020). Pyannote.audio: Neural building blocks for speaker diarization. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7124–7128. https://doi.org/10.1109/ICASSP40776.2020.9052974

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, *22*(2), 249–254. https://www.aclweb.org/anthology/J96-2004

Da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R., & Nakov, P. (2019). Fine-grained analysis of propaganda in news article. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5636–5646. https://doi.org/10.18653/v1/D19-1565

Diamond, S., & Boyd, S. (2016). CVXPY: A python-embedded modeling language for convex optimization. *J. Mach. Learn. Res.*, *17*(1), 2909–2913.

Gwet, K. Li. (2012). *Handbook of inter-rater reliability.* 6–7.

Harris, C. R., Millman, K. J., Walt, S. J. van der, Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M. H. van, Brett, M., Haldane, A., R'ıo, J. F. del, Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Krippendorff, K. (2011). *Computing krippendorff's alpha-reliability.* https://repository.upenn.edu/asc_papers/43/

Lam, S. K., Pitrou, A., & Seibert, S. (2015). Numba: A LLVM-based python JIT compiler. *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*. https://doi.org/10.1145/2833157.2833162

Mathet, Y. (2017). The Agreement Measure Gamma-Cat : a Complement to Gamma Focused on Categorization of a Continuum. *Computational Linguistics*, *43*(3), 661–681. https://doi.org/10.1162/COLI_a_00296

Mathet, Y., Widlöcher, A., & Métivier, J.-P. (2015). The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, *41*(3), 437–479. https://doi.org/10.1162/COLI_a_00227

Serratrice, L. (2000). Book reviews : The CHILDES project: Tools for analyzing talk, 3rd edition. *First Language*, *20*(60), 331–337. https://doi.org/10.1177/014272370002006006

Titeux, H., Riad, R., Cao, X.-N., Hamilakis, N., Madden, K., Cristia, A., Bachoud-Lévi, A.-C., & Dupoux, E. (2020, May). Seshat: A tool for managing and verifying annotation campaigns of audio data. *LREC 2020 - 12th Language Resources and Evaluation Conference*. https://hal.archives-ouvertes.fr/hal-02496041

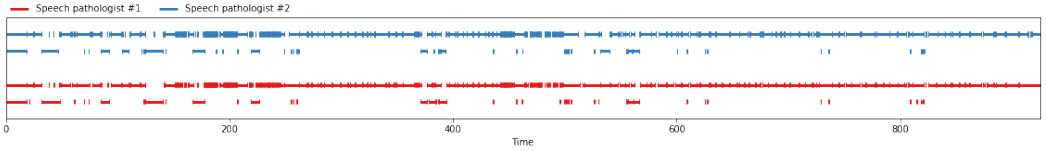


Fig. 2.11.: Visualization of turn-takings annotations of one interview by 2 different speech pathologists.

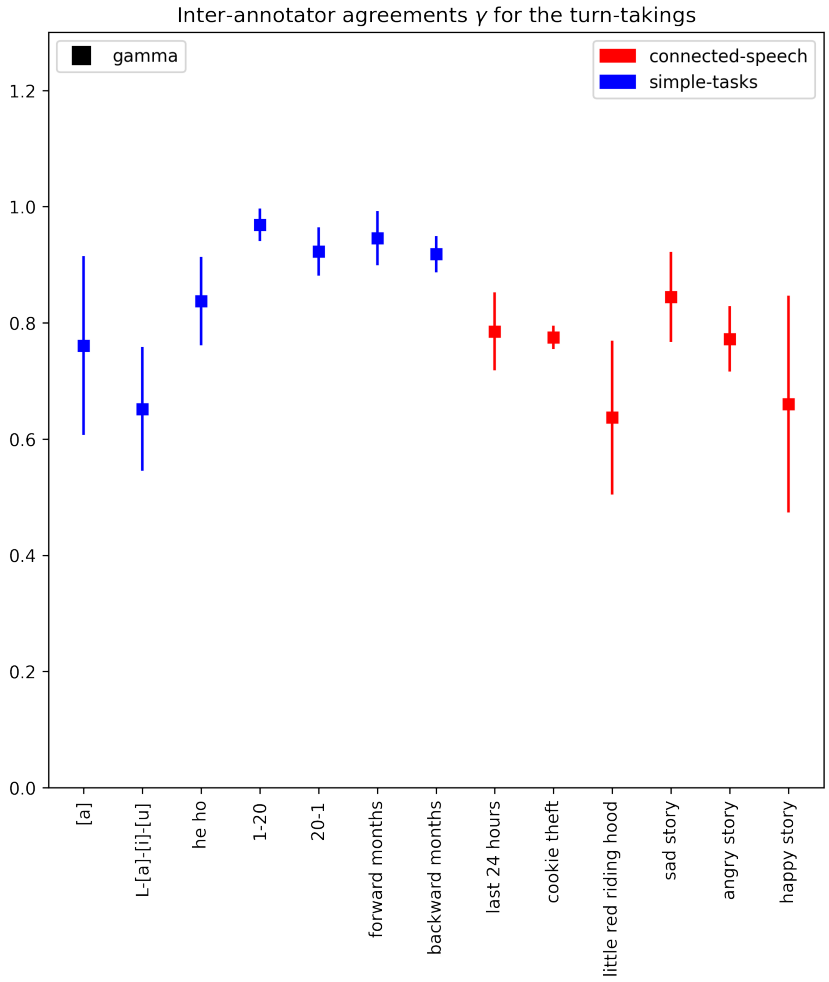


Fig. 2.12.: Inter-annotator agreements for the turn-takings. The measures are extracted from 5 interviews that have been annotated by two speech pathologists. Square and Error bars represent the mean and standard deviations respectively of the γ obtained per task.

Thanks to this contribution, we were able to apply it to BasalVoice and compute the inter-annotator agreements for the turn-takings (See Figure 2.12) and for the sentence boundaries (See Figure 2.13) for each tasks of the corpus. These agreements are computed individually for each task, for the double annotation by two

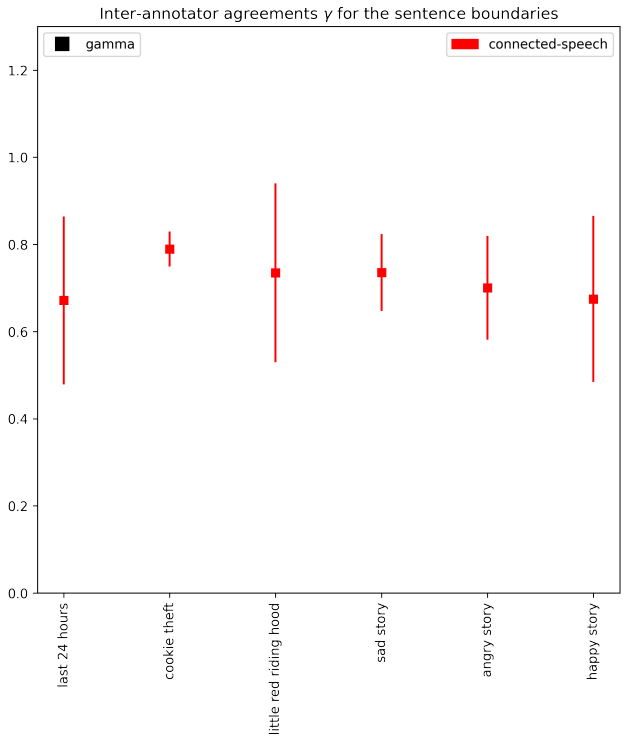


Fig. 2.13.: Inter-annotator agreements for the sentence boundaries. The measures are extracted from 5 interviews that have been annotated by two speech pathologists. Square and Error bars represent the mean and standard deviations respectively of the γ obtained per task.

different speech pathologists of 5 different files. For the turn-takings, it was reported fixed-vocabulary tasks are easier to annotate that open-vocabulary tasks (without taking into account sustained vowels) and indeed we observed a higher γ agreement than open-vocabulary tasks. One potential explanation for sustained vowels, as there are very few categories in sustained vowels, annotations precision for unitizing need to be very high.

We also notice that the overall inter-annotator γ agreements of sentence boundaries of interviewees, in Figure 2.13 is on the same level as turn-taking but there are fewer classes. This matches the speech pathologists' feedback concerning the difficulty to annotate limit sentences. Overall, the mean γ agreements for all tasks for the turn-takings and sentence boundaries were all above 0.6, which represents an acceptable value for agreement.

However, annotations of spontaneous spoken conversation are even more complicated. Indeed, the possible number of sentence produced by the SpontaneousCHAT protocol is *infinite*. Based on our knowledge, there is no way to measure agreements based on spontaneous annotations unitized in time. Comparing category mistakes is

easy while computing the distance between two annotations of disfluencies is not trivia. This represents a great line of future research, a potential avenue would be to build a distance between segment trees (See Section 2.3).

2.2.3 Seshat: A tool for managing and verifying annotation campaigns of audio data

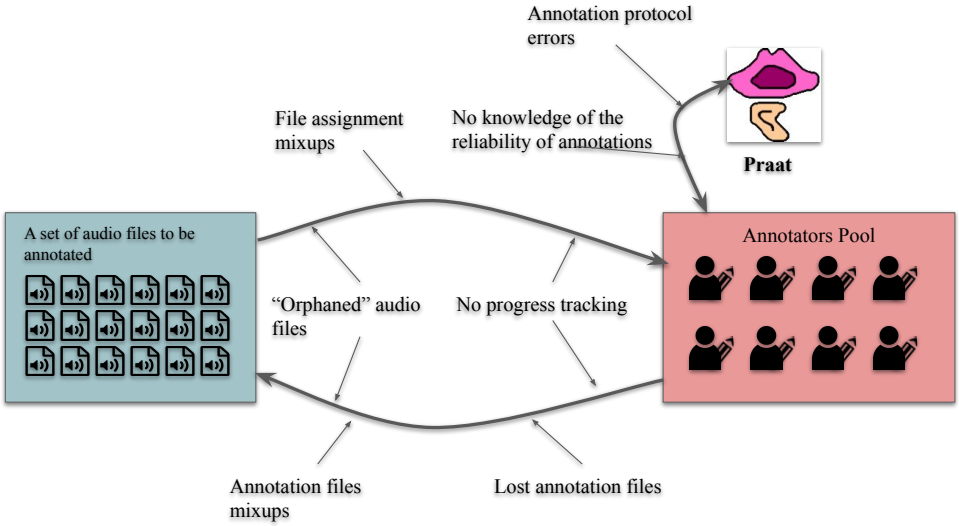


Fig. 2.14.: Major problems during the creation and annotation of speech databases such as BasalVoice. This figure is derived from a figure of Hadrien Titeux.

In previous sections, we describe *how* we collect the audio data, *how* we annotated the data, *how* we compared segmentation annotations with the γ agreement (Titeux & Riad, 2021; Mathet et al., 2015), but the construction of *reliable* and *usable* annotations of speech and language data is at-risk when the database scales. In addition, without proper annotations, the development of machine learning algorithms will be even harder (Tanno et al., 2019). As mentioned earlier in the chapter, the reliability is a critical criterion to fulfill in order to prove that speech and language markers can be used as markers in the followup of HD. Indeed, the annotations of audio and speech data are complex tasks to carry for researchers (Zue et al., 1990) and this represents now a whole type of industry with the emergence of companies as Scale or Kili technologies.

As the number of audio files to annotate exponentially grows, researchers are facing several problems that we illustrate in the Figure 2.14. To solve these issues, we need a systematic way to ensure the 'usability' of files, that they are are conform to the

protocol and that different annotators are consistent with each other. To do so, we introduced Seshat, a system to manage annotation campaigns of audio data based on the Praat software (Boersma et al., 2002). This system has been open-sourced here. This research has been done in close collaboration with Hadrien Titeux and Xuan Nga Cao. We include below the research paper presenting all the details concerning the Seshat system that has been described in LREC 2020 proceedings (Titeux* et al., 2020).

This contribution helped us to scale and speed-up our annotation efforts, without compromising the quality of annotation. This allowed up to train and manage a total of 20 speech pathologists, with up to 6 pathologists at the same time.

Seshat: A tool for managing and verifying annotation campaigns of audio data

Hadrien Titeux[⋆1] **, Rachid Riad**[⋆1,2]**, Xuan-Nga Cao**[1]**, Nicolas Hamilakis**[1]**, Kris Madden**[1]**,**
Alejandrina Cristia[1]**, Anne-Catherine Bachoud-Lévi**[2]**, Emmanuel Dupoux**[1]
1 LSCP/ENS/CNRS/EHESS/INRIA/PSL Research University, Paris, France
2 NPI/ENS/INSERM/UPEC/PSL Research University, Créteil, France
{hadrien.titeux, rachid.riad}@ens.fr,
{ngafrance, nick.hamilakis562, thekrismadden, alecristia, bachoud, emmanuel.dupoux}@gmail.com

Abstract

We introduce Seshat, a new, simple and open-source software to efficiently manage annotations of speech corpora. The Seshat software allows users to easily customise and manage annotations of large audio corpora while ensuring compliance with the formatting and naming conventions of the annotated output files. In addition, it includes procedures for checking the content of annotations following specific rules that can be implemented in personalised parsers. Finally, we propose a double-annotation mode, for which Seshat computes automatically an associated inter-annotator agreement with the γ measure taking into account the categorisation and segmentation discrepancies.

Keywords: speech transcription, speech corpora, annotations management

1. Introduction

Large corpora of speech, obtained in the laboratory and in naturalistic conditions, become easier to collect. This new trend broadens the scope of scientific questions on speech and language that can be answered. However, this poses an important challenge for the construction of reliable and usable annotations. Managing annotators and ensuring the quality of their annotations are highly demanding tasks for research endeavours and industrial projects (Zue et al., 1990). When organised manually, the manager of annotation campaigns usually faces three major problems: the *mishandling of files* (e.g., character-encoding problems, incorrect naming of files), the *non-conformity of the annotations* (Moreno et al., 2000), and the *inconsistency of the annotations* (Gut and Bayerl, 2004).

In this paper, we introduce *Seshat*, a system for the automated management of annotation campaigns for audio/speech data which addresses these challenges. It is built on two components that communicate via a Restful API: a back-end (server) written in Flask and a front-end (client) in Angular Typescript. Seshat is easy to install for non-developers and easy to use for researchers and annotators while having some extension capabilities for developers.

In Section 2., we describe the related work on annotations tools, which do not provide solutions to all the aforementioned challenges during corpus creation. In Section 3., we make an overview of the different functionalities of the software. Then, we explain, in Section 4., the architecture of the software, and also the several UX/UI design and engineering choices that have been made to facilitate the usage of the platform. We describe how to use of Seshat in Section 5. and Section 6. presents two specific use-cases. Finally, we conclude and describe future plans for Seshat in Section 7..

⋆ Equal contribution. This work was conducted while E. Dupoux was a part-time Research Scientist at Facebook AI Research. Code for Seshat is available on Github at `https://github.com/bootphon/seshat`

2. Related Work

Self-hosted annotation systems. There are many standalone solutions for the transcription of speech data that are already used by researchers: Transcriber (Barras et al., 2001), Wavesurfer (Sjölander and Beskow, 2000), Praat (Boersma and others, 2002), ELAN (MacWhinney, 2014), XTrans (Glenn et al., 2009). These systems allow the playback of sound data and the construction of different layers of annotations with various specifications, with some advanced capabilities (such as annotations with hierarchical or no relationship between layers, number of audio channels, video support). Yet, these solutions lack a management system: each researcher must track the files assigned to annotators and build a pipeline to parse (and eventually check) the output annotation files. Moreover, checking can only be done once the annotations have been submitted to the researchers. This task becomes quickly untraceable as the number of files and annotators grow. In addition, most of these transcription systems do not provide a way to evaluate consistency (intra- and inter-annotator agreement) that would be appropriate for speech data (Mathet et al., 2015).

Web-based annotations systems. There are several web-based annotation systems for the annotation of audio data. Among them we find light-weight systems, like the VIA software (Dutta and Zisserman, 2019) or Praat on the web (Dominguez et al., 2016) that allow to build simple layers of annotations. However, they do not provide a proper management system for a pool of annotators nor do they integrate annotation checking.

On the other side of the spectrum, there are more sophisticated systems with various capabilities. Camomille (Poignant et al., 2016) and the EMU-SDMS system (that can also be used offline) (Winkelmann et al., 2017) allow to work with speech data and to distribute the tasks to several annotators. But these systems require expertise in web hosting and technologies to deploy and modify them.

Finally, WebAnno (Yimam et al., 2013) and GATE Teamware (Bontcheva et al., 2013) are the tools that most closely match our main contributions regarding quality con-

trol (conformity and consistency checking), annotators' management and flexibility. WebAnno includes consistency checking with the integration of different metrics (Meyer et al., 2014). However, these tools have only been built for text data. The format and all the custom layers have been designed for Natural Language Processing tasks. Porting WebAnno to support speech data seemed a major engineering challenge. That is why it appeared necessary to develop a new and user-friendly tool addressed to the speech community.

3. Overview of Seshat

Seshat is a user-friendly web-based interface whose objective is to smoothly manage large campaigns of audio data annotation, see Figure 2. Below, we describe the several terms used in Seshat's workflow:

Audio Corpus
: A set of audio/speech files that a **Campaign Manager** wants to annotate. It is indicated either by a folder containing sound files, or by a CSV summarizing a set of files. We support the same formats as Praat so far: WAV, Flac and MP3.

Annotation Campaign
: An object that enables the **Campaign Manager** to assign **Annotation Tasks** to the **Annotators**. It references a **Corpus**, and allows the Manager to track the annotation's tasks progress and completion in real time. At its creation, a **Textgrid Checking Scheme** can also be defined for that campaign.

Annotation Task
: It is contained in an **Annotation Campaign**, it references an audio file from the campaign's designated **Audio Corpus**, and assigned to **Annotators**. It can either be a *Single Annotator Task* (assigned to one Annotator) or a Double Annotator Task (assigned to two annotators, who will annotatote the assigned task in parallel).

Textgrid Checking Scheme
: A set of rules defining the TextGrid files' structure and content of the annotations. It is set at the beginning of the **Annotation Campaign's** creation, and is used to enforce that all TextGrids from the campaign contain the same amount of Tiers, with the same names. It can also enforce, for certain chosen tiers, a set of valid annotations.

Campaign Manager
: Users with the rights to create **Annotation Campaigns** and **Annotators** user accounts, and assign **Annotation Tasks** to **Annotators**.

Annotator
: Users who are assigned a set of **Annotation Tasks**. Their job is to complete the annotation of the audio files with the Praat software.

  If the TextGrid file they submit does not comply with their **Annotation Task**'s **TextGrid Checking Scheme**, Seshat pinpoint their annotation errors with detailed messages. The annotator can resubmit the concerned file to the platform based on these different feedbacks.

Once they they connected to their instance of Seshat, *campaign managers* can access ongoing annotation campaigns or create new ones. Campaign managers are able to add *annotators*, assign *annotation tasks* and track progress. Annotator see a list of assigned tasks. The first step for them is to download the sound file with its corresponding autogenerated template TextGrid. In the current implementation, the annotation work has to be done locally with Praat. An upcoming version will use of web tools like Praat on the web (Dominguez et al., 2016). Once the task is completed, the TextGrid file is to be uploaded to Seshat via the web interface. We used the TextGrid format because of the wide acceptance of the Praat software in the speech science community (e.g., language acquisition research, clinical linguistics, phonetics and phonology).

The *Textgrid Checking Scheme* that encompasses rules on the tier's naming, file structure, and the content of the annotations, is associated with a specific campaign and defined at the creation of the campaign. Seshat back-end will automatically check that the submitted TextGrid file conforms to the *Annotation Campaign*'s *Textgrid Checking Scheme*.

Seshat allows the campaign manager to create two type of tasks: single annotator, and double annotator. Regarding the first task, one audio file is attributed to one annotator. Once the annotation is completed, Sesha automatically checks the conformity of the annotation, and only declares a tasks completed if the conformity checks is passed. Regarding the second task, one audio file is attributed to two annotators. The two annotators annotate the same file *independently*, then the two versions are merged and the annotators are guided through a *compare and review* process to agree one final version. We summarise in the Figure 1 the different steps for the double-annotator task. At each step during merging, the two annotators are provided feedbacks to focus on where are the disagreements. This process also results in the computation of an *Inter-annotator agreement* for each file. The double annotator task can be used to train new annotators alongside experts.

Annotating speech data is a joint task of segmentation and categorisation of audio events. That is why we adopted the γ measure (Mathet et al., 2015) to evaluate the inter- or intra- annotator agreement in each individual tier. Campaign manager can customise the distance used by γ by inserting a custom distance along their own parser (See short snippet of code for a parser of French Phonetics with the SAMPA alphabet in Algorithm 1).

4. Development

4.1. Engineering choices

Our utmost priority when building Seshat was to make it as easy as possible for others to deploy, use, administer and eventually contribute to. To do so, we chose the most common frameworks that are free and open-source, all of which are detailed in the following sections. Additionally, to match the current trend in web development, we decided to use the so-called "web-app" architecture for Seshat, i.e.,

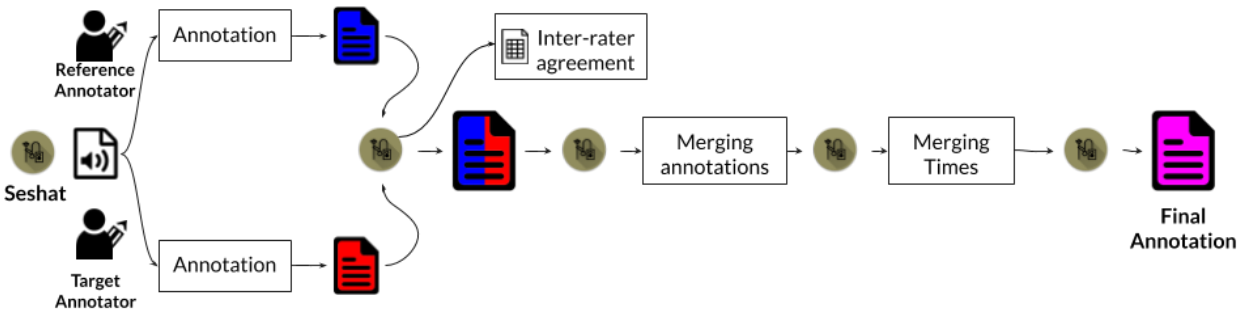


Figure 1: Double Annotator task overview. Inter-rater agreement is computed by the interface for the first independently annotated files in Red and Blue.

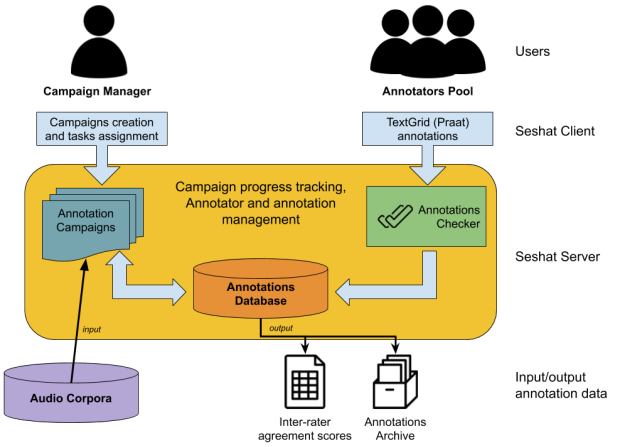


Figure 2: Seshat architecture: two different front-end, for annotators and campaign manager, a back-end with persistent data storage of the annotations and the inter-rater agreements.

we separated the application into two distinct entities: a *front-end*, running on the browser, and a *back-end*, serving data to the front-end and interacting with the database.

4.1.1. Back-end Choices

The back-end system runs on a server. It holds and updates the campaign databases and runs the annotation checking and inter-rater agreement evaluation services. We chose Python, given its widespread use in the scientific community[1], with a wide array of speech and linguistic packages. Moreover, its usage on the back-end side will allow the future integration of powerful speech processing tools like Pyannote (Bredin et al., 2019) to semi-automatize annotations. We thus went for Python3.6 for Seshat's server back-end. We used the Flask-Smorest[2] extension (which is based on Flask[3]) to clearly and thoroughly document our API, which can be exported to the popular OpenAPI 3.0.2[4] RESTful API description format.

The files and server data are stored on a MongoDB[5] database, chosen for its flexible document model and general ease of use. We used the Object-Relational Mapping (ORM) MongoEngine[6] to define our database schemas and interact with that database. MongoDB's GridFS system also allowed us to directly store annotation files (which are usually very light-weight) directly in the database, instead of going through the file system.

4.1.2. Front-end Choices

The front-end handles all of the interactions between the users (campaing manager or annotator) with the databses. It is implemented as an App within their browser. We decided to base Seshat's front-end on the Angular Typescript [7] framework. Despite its' steep learning curve, it enforces strict design patterns that guarantee that others can make additions to our code without jeopardising the stability of the App. Angular Typescript has a wide community support in the web development industry and is backed by Google and Microsoft. Moreover, the fact that it is based on TypeScript alleviates the numerous shortcomings of JavaScript, ensuring our implementation's readability and stability.

4.2. UX/UI Choices

The interface and the features we selected for our implementation are the process of a year-long iterative process involving a team of annotators, two campaign managers and software engineers. We followed some guiding principles from the recent Material[8] design language. Our goal while designing our interface (with the help of a professional designer) was to make it fully usable by non-technical people. We also put some extra care into the annotators' interface to give them a clear sense of what is to be done, how they should follow the annotation protocol, and how to correct potential errors in their annotations (See Figure 3) The goal was to reduce the number of actions to perform for annotators and enable to focus only on the annotations content.

[1] https://insights.stackoverflow.com/survey/2019

[2] https://github.com/marshmallow-code/flask-smorest

[3] https://www.palletsprojects.com/p/flask/

[4] https://github.com/OAI/OpenAPI-Specification/blob/master/versions/3.0.2.md

[5] https://www.mongodb.com/

[6] http://mongoengine.org/

[7] https://angular.io/

[8] https://material.io/design/introduction/#principles

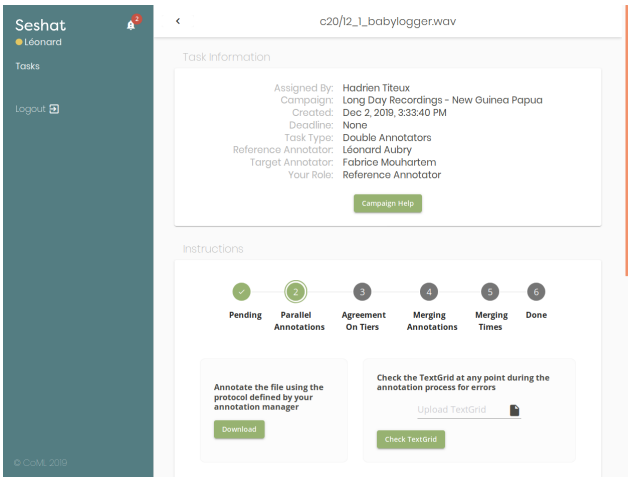


Figure 3: Assigned task from the annotator's point of view.

5. Using Seshat

5.1. Installation and Setup

Setting up a modern fully-fledged web service is a arduous task, usually requiring a seasoned system administrator as well as sometimes having very precise system requirements. Luckily, the Docker[9] virtualisation platform ensures that anyone with a recent-enough install of that software can set up Seshat in about one command (while still allowing some flexibility via a configuration file). For those willing to have a more tightly-controlled installation of Seshat on their system, we also fully specify the manual installation steps in our online documentation[10]).

Importing an audio corpus that you are willing to annotate is easy as dropping files into a default 'corpora/' folder. It is possible to either drop a folder containing audio files (with no constraints on the folder's structure), or a CSV file listing audio filenames along with their durations (in case the files are sensitive and you're not willing to risk them being hosted on the server). It is then possible to review the automatically imported files *via* the web interface.

5.2. Launching and monitoring an annotation campaign

The Campaign manager can easily define and monitor annotation campaign. As shown in Figure 5, the online form enable to choose corpora, pre-define and pre-configure the annotations scheme (tiers and parsers). There are 2 types of tiers already implemented by default: one with no check at all, and one with pre-defined categories. For the latter, these categories are pre-defined when the campaign is created.

Only Campaign managers can access and build new campaigns. If Campaign manager have several campaigns they can easily switch between them via the menu bar or get a full overview with the dashboard (See Figure 4). The campaign managers can visualise the progress of the assigned tasks at the campaign level or more precisely at the task level. They can retrieve all the intermediate files that have been created for each task. For instance, the campaign manager can examine qualitatively and quantitatively what are the annotation differences before the merge phases of the double annotator task.

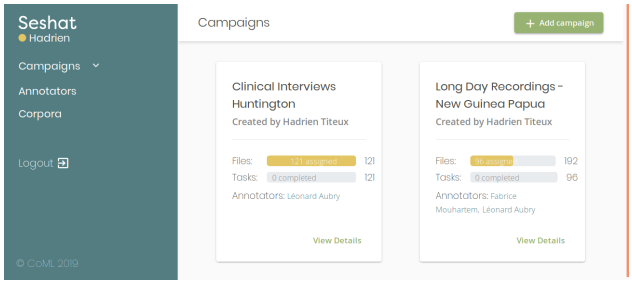


Figure 4: Dashboard for the campaign manager

5.3. Scripting API

For those willing to interact with Seshat using code, it is possible to interact with Seshat using either its RESTful API or its command-line interface (CLI). The API endpoints that can be called are all listed in a simple interface, and can be made from any programming language able to make HTTP requests. The CLI interface can be used via your terminal, and therefore can be interacted with using Bash scripts.

A typical usage of these features would be to assign annotation tasks from a large speech corpus (spoken by several speakers) to a large pool of annotators, all the while making sure each annotator has a similar number of tasks, with each speaker being evenly distributed among annotators as well. This would be tedious to do manually via the user interface, but easy to program in any scripting language.

5.4. Annotation Parser Customisation

We aimed at a reasonable trade-off between simplicity and flexibility for the TextGrid annotations checking component. However, we understand (from our own experience in particular) that sometimes annotations can follow a very specific and complex standard (for instance, parsing SAMPA phonemes strings). To allow users to define their own annotation standards, we added the possibility for users to define an annotation parser, via a simple package-based extension system (taking inspiration from pyannote's extension system). Anyone willing to create a new annotation parser has to be able to program in Python and have a minimal understanding of its packaging system.

As presented in our example French SAMPA Parser (Algorithm 1), implementing a custom annotation parsers only requires the overload of two methods from Seshat's `BaseCustomParser` class:

- `check-annotation`: takes an annotation string as input and raises an error if and only if the annotation is deemed to be invalid. It doesn't return anything.
- `distance`: takes two annotations as input and should return a float corresponding to the distance between these two annotations.

[9] https://www.docker.com/

[10] https://seshat-annotation.readthedocs.io/

```python
class FrenchSAMPAParser(BaseCustomParser):
    PHONEMES = ['&/', '2','9','9~','@','A','A/','E','E/','H',
                'N','O','O/','R','S','U~/','Z','a','a~','b',
                'd','e','e~','f','g','i','j','k','l','m','n',
                'o','o~','p','s','t','u','v','w','y','z','J']

    def __init__(self):
        # sorting (descending) phonemes by length
        self.PHONEMES = sorted(self.PHONEMES, key=len, reverse=True)

    def parse_sampa(self, pho_str: str) -> List[str]:
        """Parses a French phoneme string into a phoneme list
         ex : "septa~br" -> [s, e, p, t, a~, b, r]"""
        original_str = str(pho_str)
        pho_list = []
        while pho_str:
            parsed_phoneme = False
            # trying to match longest phoneme names first
            for phoneme in self.PHONEMES:
                if pho_str[:len(phoneme)] == phoneme:
                    pho_list.append(pho_str[:len(phoneme)])
                    pho_str = pho_str[len(phoneme):]
                    parsed_phoneme = True
                    break
            if not parsed_phoneme:
                raise AnnotationError(
                    "Can't parse phonetic form %s (stuck at %s)"
                                    % (original_str, pho_str))
        return pho_list

    def check_annotation(self, annot: str) -> None:
        self.parse_sampa(annot)

    def distance(self, annot_a: str, annot_b : str) -> float:
        """Computes the levenshtein distance between two phone sequences"""
        parsed_a = self.parse_sampa(annot_a)
        parsed_b = self.parse_sampa(annot_b)
        return levenshtein(parsed_a, parsed_b)
```

Algorithm 1: Parser Plugin Example. This parser checks the units to allow only phone sequences in the SAMPA format. The distance that can be used for the inter-rater agreement is the Levenshtein distance.

5.5. Inter-rater agreement: the γ measure

It is necessary have a measure of confidence to obtain high-quality datasets and therefore to draw valid conclusions from annotations. Annotations tasks of audio and speech data usually have some specificities. The items to annotate have to be both segmented in time and categorised. The segments can be hierarchically defined or overlapping. In addition, the audio stream may require only sparse annotations (especially in-the-wild recordings which contain a lot of non-speech segments). To evaluate speech annotations, the measure needs to take these characteristics into account. That is why we decided to re-implement and compute the γ measure (see Mathet et al. (2015) for its design and the advantages of this measure over previous agreement measures).
First, the γ software aligns (tier-wise) the annotations of the different annotators. To align the two sets of annotations the γ measure the distance between all the individual units. The difference of position of two annotated units u and v is measured with the positional distance:

$$d_{\text{pos}}(u,v) = \left(\frac{|start(u) - \text{start}(v)| + |\text{end}(u) - \text{end}(v)|}{(end(u) - \text{start}(u)) + (end(v) - \text{start}(v))} \right)^2$$

If the tiers are categorical, the distance for the content of the annotated units u and v is defined as:

$$d_{\text{cat}}(u,v) = \mathbb{1}(cat(u) == cat(v))$$

This distance can be over-written by the custom parser as mentioned above. These two distance are summed with equal weights to obtain the distance between every annotated units from 2 annotators. Then, it is possible to obtain the *disorder* $\delta(a)$ of a specific alignment a by summing the distance of all the aligned units in a. All possible alignments a are considered and the one that minimises the disorder $\delta(a)$ is kept.
To get the value of γ, the disorder is chance-corrected to obtain an expected disorder. It is obtained by re-sampling randomly the annotations of the annotators. This means that real annotations are drawn from the annotators, and one position in the audio is randomly chosen. The annotation is split at this random position and the two parts are permuted. It is then possible to obtain an approximation of the expected disorder δ_e. The final agreement measure is defined as:

$$\gamma = 1 - \frac{\delta(a)}{\delta_e}$$

This γ measure is automatically computed by the back-end server for the double-annotator tasks. The Campaign manager can retrieve these measures in Seshat by downloading a simple CSV file.

6. Use cases

We present two use cases on which Seshat was developped: clinical interviews, and daylong child-centered recordings.

6.1. Clinical interviews

Seshat was intially developped to study the impact of Huntington's Disease (Walker, 2007) on speech and language production. One hundred and fifty two interviews between a neuropsychologist and a patient with the Huntington's Disease (HD) were recorded between June 2018 and November 2019. The campaign manager created a campaign with multiple tiers to annotate the turn takings and the speech/non speech boundaries of the utterances of the patient. For both tasks, the annotations did not need to cover completely the audio (sparsity property mentioned above). For the Turn-taking annotations, there are 3 pre-defined tiers, each one with a single class ('Patient', 'Non-Patient', and 'Noise'), which results in possible overlap between these classes. For the Utterance annotations, there is only one pre-defined class ('Utterance').
To this date, a total of 67 files have been fully annotated with the help of Seshat by a cohort of 18 speech pathologist students (see Figure 5). Among these, 16 have been done by 2 different annotators independently with the Double-annotator task. The results are summarised in Table 1.
Even though there are more categories for Turn-Takings than Utterance (Gut and Bayerl (2004) reported that the more categories the more the task is difficult in speech annotations), the mean γ for the Turn-Takings $\gamma = 0.64$ is slightly higher than the one for Utterance $\gamma = 0.61$. And the range of values for the Turn-Takings is smaller than the Utterance. Indeed, the speech pathologists reported the difficulty to annotate the boundary of utterances in spontaneous speech, with several ambiguous cases due to pauses. These results will help us to redefine the protocol and be more precise on the given instructions.

6.2. In-the-wild child-centered recordings

The Seshat software is also currently used to annotate audio files in a study of day-long audio-recordings captured by

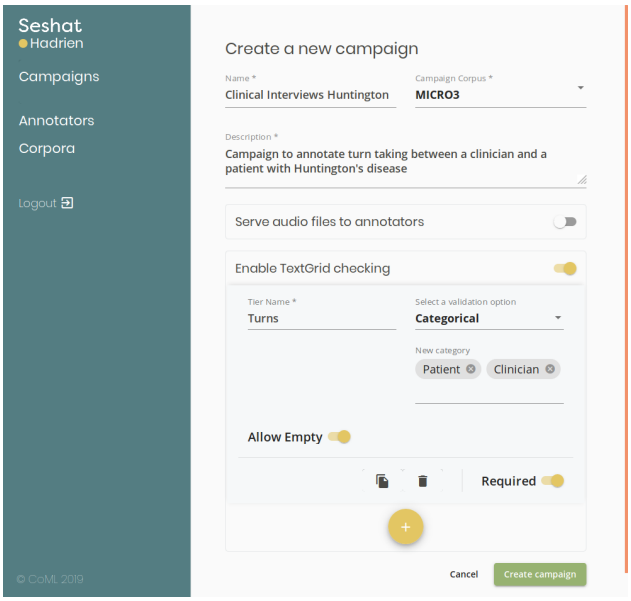


Figure 5: Annotation Campaign definition in Seshat for clinical interviews between a patient with the Huntington's Disease and a neuropsychologist

Tiers	γ		
	Mean	Range	#classes
Turn-Takings	0.64	0.18	3
Utterance	0.61	0.39	1

Table 1: γ Inter-rater agreements summary for 16 clinical interviews between a neuropsychologist and a patient with the HD.

two devices (LENA (Gilkerson and Richards, 2008), and a BabyCloud baby-logger device) worn by young children growing up in remote Papua New Guinea. The project aims at establishing language input and outcomes in this seldom-studied population. To establish reliability levels, 20 1-min files were double-annotated by 2 speech pathology students. Among the tasks given to the annotators there was: (1) locating the portions of Speech (Speech activity), (2) locating the speech produced by an adult that is directed to a child or not (*Adult-Directed Speech* versus *Child-Directed Speech*). As in the previous example, the annotations do not need to cover the full audio file. The Speech Activity task has only 1 class ('Speech') and the Addressee task has 2 classes ('ADS', 'CDS').

Tiers	γ		
	Mean	Range	#classes
Speech activity	0.46	0.60	1
ADS vs CDS	0.27	0.39	2

Table 2: γ Inter-rater agreement for 20 1-min slices extracted from child-centered day-long recordings. ADS and CDS stand for Adult-Directed Speech and Child-Directed Speech respectively.

These recordings have been done in naturalistic and noisy conditions; moreover, the annotators do not understand the language. Probably as a result of these challenges, agreement between annotators is lower than in the Clinical interviews use case. This information is nonetheless valuable to the researchers, as it can help them appropriately lower their confidence in the ensuing speech quantity estimates.

7. Conclusion and Future work

Seshat is a new tool for the management of audio annotation efforts. Seshat enables users to define their own campaign of annotations. Based on this configuration, Seshat automatically enforces the format of the annotations returned by the annotators. Besides, we also add the capability to finely tailor the parsing of the annotations. Finally, Seshat provides automatic routines to compute the inter-rate agreements that are specifically designed for audio annotations. Seshat lays some foundations for more advanced features, either for the interface or the annotation capabilities. In future work, we plan to implement an automatic task assignments and an integration of a diarization processing step to reduce human effort. Another planned feature is to add possibility for the campaign manager to design more complex annotation workflows such as, for instance, dependencies between tiers or more intermediate steps of annotations.

8. Acknowledgements

This research was conducted thanks to Agence Nationale de la Recherche (ANR-17-CE28-0007 LangAge, ANR-16-DATA-0004 ACLEW, ANR-14-CE30-0003 MechELex, ANR-17-EURE-0017, ANR-10-IDEX-0001-02 PSL*, ANR-19-P3IA-0001,) and grants from Facebook AI Research (Research Grant), Google (Faculty Research Award), and Microsoft Research (Azure Credits and Grant), and a J. S. McDonnell Foundation Understanding Human Cognition Scholar Award.

9. References

Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. (2001). Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1-2):5–22.

Boersma, P. et al. (2002). Praat, a system for doing phonetics by computer. *Glot international*, 5.

Bontcheva, K., Cunningham, H., Roberts, I., Roberts, A., Tablan, V., Aswani, N., and Gorrell, G. (2013). Gate teamware: a web-based, collaborative text annotation framework. *Language Resources and Evaluation*, 47(4):1007–1029.

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., and Gill, M.-P. (2019). pyannote. audio: neural building blocks for speaker diarization. *arXiv preprint arXiv:1911.01255*.

Dominguez, M., Latorre, I., Farrús, M., Codina-Filba, J., and Wanner, L. (2016). Praat on the web: an upgrade of praat for semi-automatic speech annotation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 218–222.

Dutta, A. and Zisserman, A. (2019). The via annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2276–2279. ACM.

Gilkerson, J. and Richards, J. A. (2008). The lena natural language study.

Glenn, M. L., Strassel, S. M., and Lee, H. (2009). Xtrans: A speech annotation and transcription tool. In *Tenth Annual Conference of the International Speech Communication Association*.

Gut, U. and Bayerl, P. S. (2004). Measuring the reliability of manual annotations of speech corpora. In *Speech Prosody 2004, International Conference*.

MacWhinney, B. (2014). *The CHILDES project: Tools for analyzing talk, Volume II: The database*. Psychology Press.

Mathet, Y., Widlöcher, A., and Métivier, J.-P. (2015). The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.

Meyer, C. M., Mieskes, M., Stab, C., and Gurevych, I. (2014). Dkpro agreement: An open-source java library for measuring inter-rater agreement. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 105–109.

Moreno, A., Lindberg, B., Draxler, C., Richard, G., Choukri, K., Euler, S., and Allen, J. (2000). Speechdat-car. a large speech database for automotive environments. In *LREC*.

Poignant, J., Budnik, M., Bredin, H., Barras, C., Stefas, M., Bruneau, P., Adda, G., Besacier, L., Ekenel, H., Francopoulo, G., et al. (2016). The camomile collaborative annotation platform for multi-modal, multi-lingual and multi-media documents. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1421–1425.

Sjölander, K. and Beskow, J. (2000). Wavesurfer-an open source speech tool. In *Sixth International Conference on Spoken Language Processing*.

Walker, F. O. (2007). Huntington's disease. *The Lancet*, 369(9557):218–228.

Winkelmann, R., Harrington, J., and Jänsch, K. (2017). Emu-sdms: Advanced speech database management and analysis in r. *Computer Speech & Language*, 45:392–410.

Yimam, S. M., Gurevych, I., de Castilho, R. E., and Biemann, C. (2013). Webanno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 1–6.

Zue, V., Seneff, S., and Glass, J. (1990). Speech database development at mit: Timit and beyond. *Speech communication*, 9(4):351–356.

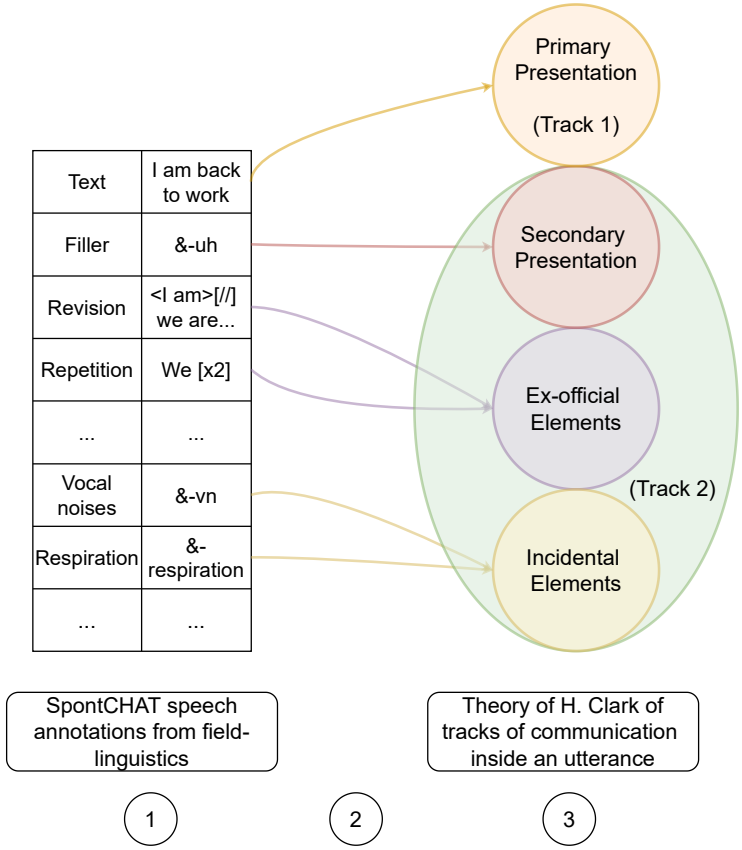


Fig. 2.15.: Schematic diagram of our annotation and analysis pipelines. 1) Introduction of SpontaneousCHAT speech annotations derived from field-linguistics protocol of Brian MacWhinney (Table 2.6) 3) Formalism and implementation of the theory of tracks of communication introduced by H. Clark . 2) Translating raw speech annotations into temporal data structures (segment trees). See below in the manuscript for the data structures.

2.3 Parsing spontaneous speech productions

Spoken natural interactions differ from written exchanges. The speech signal is evanescent, and conversation participants need to regulate floor sharing/turn-taking as the conversation unroll (Stokoe, 2018; Clark, 2002). To achieve a successful communication to do things, a participant in a conversation need to monitor and correct his/her own's speech at different linguistic levels and check for the understanding of other participants (Levelt, 1993).

To handle such complexity and issues in speech communications, human develop strategies using specific vocal items, referred as elementary particles of conversation. These particles help regulate conversations (also referred as disfluencies in speech processing): we can start and restart sentences, "repair" our own speech (Levelt,

1983), we use item such as fillers 'uh', 'um' to signal/communicate the research for the following spoken words and rare words to follow or hold the floor (Clark & Tree, 2002). Listeners can also use vocalisations and also request for more information with particles such as 'huh'? (Dingemanse et al., 2013) when context information is lacking (Piantadosi et al., 2012), or when pragmatic reasoning does not allow to infer speaker's goal (Goodman & Frank, 2016). As shown in the derived annotation protocol, Table 2.6, all these elements are present during spontaneous conversations. To reach a clear understanding on how conversation unroll and HD lose conversational capabilities through lifespan, there is a need for proper way to run data analyses on large speech corpora. Such analysis potentially can lead to a more formalized theory of speech interactions.

As mentioned in the introduction, Elizabeth Shriberg focused her modelling work of disfluencies and spontaneous speech on *how* interruption of flow in speaking occurs (Shriberg, 1994). In contrast, Herbert Clark proposed a theory for spontaneous speech by specifying the *roles* for each vocalization (linguistic or not) in communications. We focused on this theory in our work. To illustrate the complexity of vocalizations and their use, H. Clark and J.E. Fox Tree use the following example in (Clark & Tree, 2002):

> well, . I mean this . uh Mallet said Mallet was uh said something about uh you know he felt it would be a good thing if u:h . if Oscar went,

This sentence is composed of words but not only. The speaker revised previous pronounced words and also used elementary particles of communication, such as filled pauses to communicate troubles of communication to listeners. Based on all these naturalistic observations of spontaneous communications, Herbert Clark groups vocalizations in conversation in *tracks of communications* inside a given utterance (Clark, 1996). Yet, this theory developed by induction, is not completely *formalized* to properly generate hypotheses.

Clark's theory of speech performance and modelling of utterances (Clark, 1996, p. 255) states that speakers communicate using two parallel tracks. The *primary track* or *Track 1* contains the traditional linguistic content of the discourse. The *collateral track* or *Track 2* contains additional signals regulating the communication channel itself. By signal, H. Clark refers to specific actions performed by a person to mean something to another person in the Grice's terms (Grice, 2020). Among these additional signals of the collateral track we find delays, (un)filled pauses, rephrasing, mistakes, laughs, or vocal noises. The signals of the collateral track can be decomposed into three sub-types: *Secondary Presentation*, *Ex-official Elements*, *Incidental Elements*.

1. *Primary Presentation* Parts of the signal that the speaker *intends* as signals about the official business/topic of the discourse (Track 1).

2. a) *Secondary Presentation* Parts of the signal that the speaker *intends* as signals about the conversation and speech production themselves (Track 2). For instance, a filler like 'uh' in American English can help signaling delays to produce less frequent words (Clark & Tree, 2002).

   b) *Ex-official Elements* Parts of the signal that the speaker *intended* as primary presentation that the speaker later preempted thanks to other elements (Track 2). These were part of Track 1 when they were produced. Repetition and revisions and their derivatives form this category.

   c) *Incidental Elements* Those elements of the gross presentation that the speaker does not intend to be part of any communicative acts (Track 2). This category encompasses involuntary vocalizations, audible respirations, phonological fragments.

However, as underlined Clark and Tree (2002), and speech pathology studies (Amir et al., 2018), the timing and duration of these tracks is critical for the use of language. Clark's current theory of speech production is not formalized to take into account the course of time. In addition, there are no proper *data structures* to implement in practice such theory based on realistic inputs.

Our contributions are two folds for the tracks of communication (See Figure 2.15): (1) we formalized and implemented the Herbert Clark theory of tracks of communications and (2) we made it possible to obtain data structures of spontaneous utterances thanks to naturalistic inputs annotated with SpontaneousCHAT protocol and phone alignments.

First, we proposed that the theory of tracks of communications can be implemented thanks to specific data structures, *segment trees*. Segment tree is a tree data structure useful to store information concerning intervals or segments. Words are the leaves in the segment tree and contain information concerning their start and end. Their parent nodes are determined thanks to their statuses in the tracks of communication of Herbert Clark. Either words belong to the primary or collateral tracks, and collateral signals have specificities on their own. The tracks of communications are built with the union operation of segments. Finally, the root node is the full sentence with the start and end timings of the sentence.

A set of segment is described as $\mathcal{I}$ and a segment tree $S_{\mathcal{T}}$ for I is defined as:

1. $S_{\mathcal{T}}$ is a binary tree.

2. The leaves of the tree correspond to the elementary intervals of $\mathcal{I}$, in an ordered manner.

3. The internal nodes of $S_{\mathcal{T}}$ are intervals that are the union of elementary intervals, along some pre-defined values.

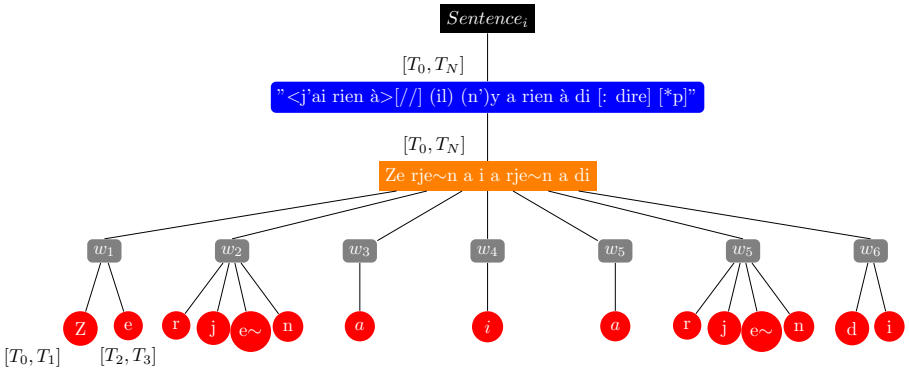
In our case, the *values* of the nodes of tracks of communication, are the categorical labels such as primary or secondary (or sub-categories of the secondary track). We denote by $\overset{+}{\leftarrow}$ the addition of a node in the tree, to the set of children. Based on the parsed trees obtain by regex rules $\mathcal{T}$ from annotations, we can derive the tree structure. To obtain the segments of node, we rely on the phonetic alignment of words. We summarize the different steps and data structures in the figure 2.16 to obtain the segment tree that represents tracks of communication. For a formal derivation of these segments tree, reader can refer to our pseudo-algorithm 1 that summarizes the steps.

Is there any reality about these data structures during spontaneous exchanges of individuals with HD? Does it inform on the disease? Can we obtain such data structures from the audio waveform? We will see in the next chapters how we started to address these questions.

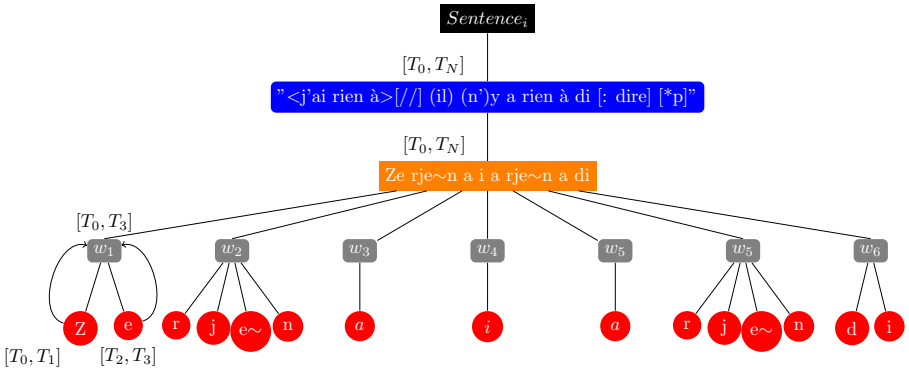
Thanks to this methodology, we can now make automatic analyses of spontaneous utterances for our BasalVoice database. All our spontaneous speech annotations could be analyzed and we could extract the segments of track of communication of each sentence of BasalVoice.

The first natural way to analyze these segment trees is just the computation of ratio per tracks of communication based total time of produced vocalizations. For instance this kind of analysis has been performed to study the fluency efficiency in stutterers (Amir et al., 2018).

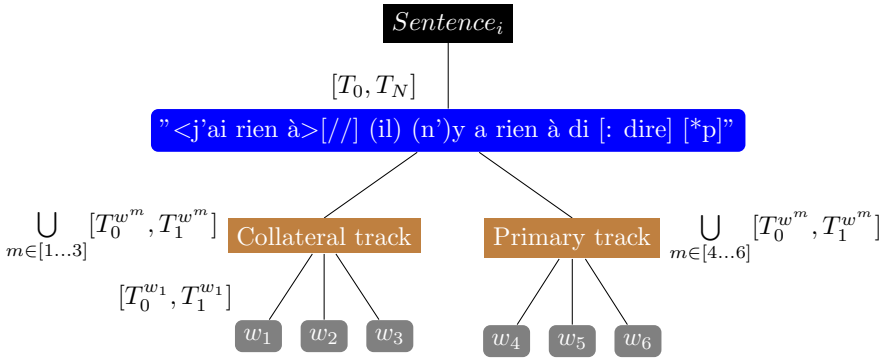
In Figure 2.17, we computed the mean of this ratio of each track of communication and for each participant in our study. This analysis is only performed for the open-vocabulary speech tasks. We can distinguish different distributions based on the group under study, with a clear diminution of efficiency, based on the ratio of Track 1, for individuals with HD. This will be verified with statistical hypothesis testing in later chapters.



(a) Input tree data structure $\mathcal{T}$ obtained from the AST parser and force phone-aligner.



(b) Construction the span of spoken words timings from phones.



(c) Segment tree representing the tracks of communication.

Fig. 2.16.: Different steps to build the Segment tree representing the tracks of communications. Sentence timings $[T_0, T_N]$ can be obtained through annotations or diarization, and set of phones timings $\{a_1 : [T_0^0, T_1^0], \ldots a_M : [T_0^M, T_1^M]\}$ can be obtained through annotations, phone recognition or force-alignment with text-to-phone system.

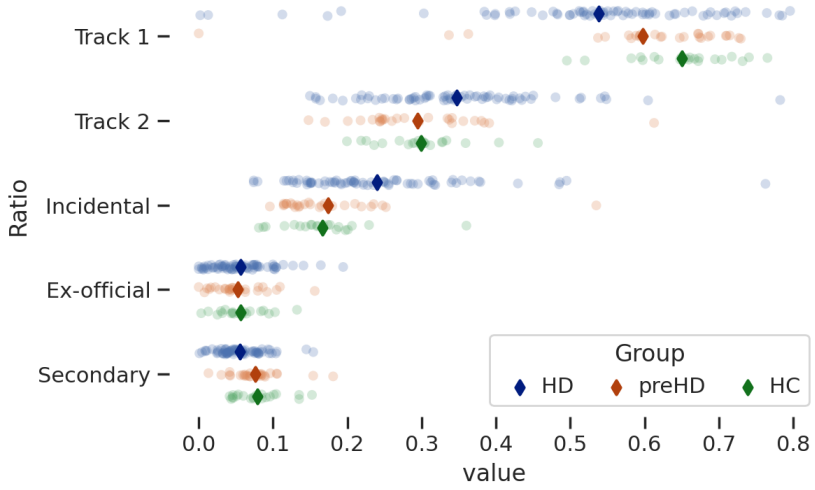


Fig. 2.17.: Distribution of ratio of tracks of communication for the different groups in BasalVoice. The measures are extracted from interviews that have been annotated by speech pathologists.

Algorithm 1: Construction of the segment tree to represent tracks of communication.

Inputs : Sentence tree data structure $\mathcal{T}$ of M spoken words w_m, Sentence timings $[T_0, T_N]$,
Set of phone timings $\{a_1 : [T_0^0, T_1^0], \dots a_M : [T_0^M, T_1^M]\}$,
Phones for each spoken word with l^m phones: $w_m : \{a_1^m, \dots a_{l^m}^m\}$.
See Figure 2.16a for inputs.

Output : Segment tree $\mathcal{S}_\mathcal{T}$ to represent tracks of communication

$w_m : [T_0^{a_1^m}, T_1^{a_{l^m}^m}] = [T_0^{w^m}, T_1^{w^m}], \forall m \in [1, M]$ ▷ Span of words timings from phones. See Figure 2.16b.

$\mathcal{S}_\mathcal{T} \leftarrow RootNode(\text{label} : w1 \dots w_M, \text{segment} : [T_0, T_N])$ ▷ Set root node.

$C_{\text{track}} \leftarrow Node(\text{label} : track(w_1), \text{segment} : [T_0^{w^1}, T_1^{w^1}])$ ▷ Set current segment.

$C_{\text{track}}.children \xleftarrow{+} Node(\text{label} : w_1, \text{segment} : [T_0^{w^1}, T_1^{w^1}])$ ▷ Add first child word.

$m_{\text{outer loop}} \leftarrow 1$

$m_{\text{inner loop}} \leftarrow m_{\text{outer loop}} + 1$

$\mathcal{S}_\mathcal{T}.children \xleftarrow{+} C_{\text{track}}$ ▷ Addition first node track to the segment tree.

while $m_{inner\ loop} \leq M$ **do** `// Iterate over children of root of` $\mathcal{T}$.

  if $track(w_{outer\ loop})$ *is* $track(w_{inner\ loop})$ **then** `// Check track.`

    $C_{\text{track}}.segment \leftarrow C_{\text{track}}.segment \bigcup [T_0^{w_{\text{inner loop}}}, T_1^{w_{\text{inner loop}}}]$ ▷ Expand current segment with union.

    $C_{\text{track}}.children \xleftarrow{+} Node(\text{label} : w_{\text{inner loop}}, \text{segment} : [T_0^{w_{\text{inner loop}}}, T_1^{w_{\text{inner loop}}}])$ ▷ Add child word.

    $m_{\text{inner loop}} \leftarrow m_{\text{inner loop}} + 1$

  else

    $C_{\text{track}} \leftarrow Node(\text{label} : track(w_{\text{inner loop}}), \text{segment} : [T_0^{w_{\text{inner loop}}}, T_1^{w_{\text{inner loop}}}])$ ▷ Set current segment.

    $C_{\text{track}}.children \xleftarrow{+} Node(\text{label} : w_{\text{inner loop}}, \text{segment} : [T_0^{w_{\text{inner loop}}}, T_1^{w_{\text{inner loop}}}])$ ▷ Add child word.

    $\mathcal{S}_\mathcal{T}.children \xleftarrow{+} C_{\text{track}}$ ▷ Addition of node track to the segment tree.

    $m_{\text{outer loop}} \leftarrow 1$

    $m_{\text{inner loop}} \leftarrow m_{\text{outer loop}} + 1$

  end

end

return $\mathcal{S}_\mathcal{T}$ ▷ Return Segment tree. See Figure 2.16c.

3

Automatic pathological speech processing

> *" There are no algorithms - yet - for producing conversation analytic transcripts. Of course, conversation analysts produce and work with more than a basic, verbatim transcript of recorded conversation. We include detail about pace, prosody, intonation, overlap, interruption, repetition, repair, interpolated laughter, the tiniest of incipient of sobs, and much, much more.".*
>
> **— Elizabeth Stokoe**
> Professor of Social Interaction
> at Loughborough University

In the previous chapter, we presented our database BasalVoice, a new resource to evaluate speech and language features as potential biomarkers for HD. We also introduced an annotation protocol digestible by a computer program and presented open-source contributions that enable to manage cohorts of annotators and verify their consistency. However, annotating conversational speech with such granularity is time consuming and *in fine*, limits the power of scientific analyses that can be performed. In addition, while the analyses of speech and language markers is promising, the use of annotations is limiting its potential to break into the clinical practice. Annotations produced by speech pathologists can be decomposed in terms of distinct computational tasks:

Diarization : Who speaks when?

Sentence boundary detection : What are the semantic boundaries of the sentences of the participants?

Identification of tracks of communications/disfluencies : Which portions of the participant's speech are primary presentations, secondary presentations, ex-official elements and incidental elements? (See Section 2.3)

Speech recognition : What are the exact words and vocalizations pronounced by the participant?

Word level categorization : What is the role of specific word in communication? What is the track of given word? Are there any specifics for this word, such as blocks or paraphasias?

Research efforts to automate these tasks by a computer program are unequal, with a large focus on Automatic Speech Recognition (ASR), despite the fact that the diarization task is an unavoidable step (Anguera et al., 2012; Park et al., 2022). In addition, there are even fewer studies trying to tackle these tasks when the setting is more spontaneous (J. Barker et al., 2018) and to build robust pipelines in challenging domains such as home dinner parties or hospitals (Ryant et al., 2020). When it comes to developing and comparing these algorithms for individuals with speech and language impairments the number of studies becomes even rarer. Usually, these tasks are studied in silo or researchers simplify the task at hand, by only focusing on sub-set of phenomena.

It is not clear to what extent the algorithms developed on public datasets from other domain will generalize to HD's spontaneous spoken language. In addition, classic methodologies in speech processing tend to overlook the expressive aspects of conversation (non verbal vocalizations, intonation and rhythm) that make spontaneous speech special and different from controlled settings such as read speech or automatic series (such as telling days of the week). As shown by Goldwater et al. (2010), disfluencies, such as filled pauses or repetitions, heavily impact and degrade the performance of ASR systems. If we want to deploy algorithms that yield speech markers for clinical applications, it is necessary to tailor them to the population under study, and properly evaluate them in real conditions.

In Section 3.1, we describe our modelling work on diarization, applied to the BasalVoice database. We introduced new methodologies to evaluate different type of speech processing approaches, and ran a number of experiments to examine which are the contributing factors for the performance. Then, in Section 3.2, we describe the experiments to identify tracks of communication in pathological speech. This represents the first attempts to identify the boundary of disfluent events from raw waveform. Based on the tracks of communication framework that we introduced in the section 2.3, we have a proper way to evaluate the different approaches. We made experiments on two datasets, an open-source dataset of stuttered speech (Bernstein & MacWhinney, 2018) and on our newly introduced BasalVoice database. This chapter is composed of different research papers, and we include additional experiments immediately after to complete these studies.

A large part of the work in this section has been done in close collaboration with Hadrien Titeux.

3.1 Who speaks when?

3.1.1 A comparison study on patient-psychologist voice diarization

The first unavoidable step to obtain speech markers is to find where the speech is in the audio signal, to answer the question 'Who speaks when?'. To assess which approach was the most suitable for the BasalVoice database on HD's speech, we compared two main families of algorithms. On the one hand, we evaluated approaches that could benefit from small amount of annotations of a given interview. These enrollment approaches can exploit the specifics of the neuropsychologist and the participant of a given interview to detect which portion of the audio belongs to each one. On the other hand, we used approaches to focus more on the 'structural roles' during the interview without any speaker enrollments: neuropsycholgist, i.e. interviewer, and interviewee in our study, HC/preHD/HD. These participants have specific roles and have specific patterns in their turn-takings and how they tackle.

To compare these approaches, we had to introduce a new type of splitting of speech corpora. In addition, as the role and number of speakers was known, we could use the Identification Error Rate (IER) and not the Diarization Error Rate (DER). Surprisingly, we found out that the best approach to detect turn-takings was a fully end-to-end neural network, not taking into account the speaker enrollment. The study reported below has been submitted to SLPAT 2022 and is included below.

A COMPARISON STUDY ON PATIENT-PSYCHOLOGIST VOICE DIARIZATION

Rachid Riad[⋆1,2] *, Hadrien Titeux*[⋆1]*, Laurie Lemoine*[2]*, Justine Montillot*[2]*,*
Agnes Sliwinski[2]*, Jennifer Hamet Bagnou*[2]*, Xuan Nga Cao*[1]*,*
Anne-Catherine Bachoud-Lévi[2]*, Emmanuel Dupoux*[1]

[1] CoML/ENS/CNRS/EHESS/INRIA/PSL Research University, Paris, France
[2] NPI/ENS/INSERM/UPEC/HD CENTER/PSL Research University, Créteil, France

ABSTRACT

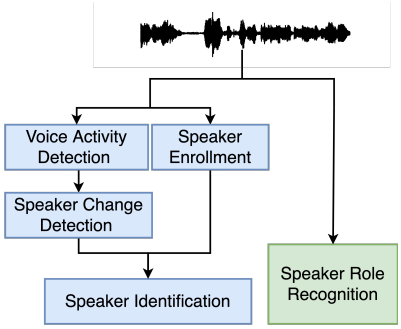
Conversations between a clinician and a patient, in natural conditions, are valuable sources of information for medical follow-up. The automatic analysis of these dialogues could help extract new language markers and speed-up the clinicians' reports. Yet, it is not clear which model is the most efficient to detect and identify the speaker turns, especially for individuals with speech disorders. Here, we proposed a split of the data that allows conducting a comparative evaluation of different diarization methods. We designed and trained end-to-end neural network architectures to directly tackle this task from the raw signal and evaluate each approach under the same metric. We also studied the effect of fine-tuning models to find the best performance. Experimental results are reported on naturalistic clinical conversations between Psychologist and Interviewees, at different stages of Huntington's disease, displaying a large panel of speech disorders. We found out that our best end-to-end model achieved 19.5% IER on the test set, compared to 23.6% achieved by the finetuning of the X-vector architecture. Finally, we observed that we could extract clinical markers directly from the automatic systems, highlighting the clinical relevance of our methods.

***Index Terms*—** Clinical Interviews, Speaker Role Recognition, Speaker Enrollment, Pathological Speech Processing, Huntington's Disease.

1. INTRODUCTION

During the last decades, it became easier to collect large naturalistic corpora of speech data. It is now possible to obtain new realistic measurements of turn-takings and linguistic behaviors [1]. These measurements can be especially useful during clinical interviews as they augment the current clinical panel of assessments and unlock home-based assessments [2]. The remote automatic measure of symptoms of patients with Neurodegenerative diseases could greatly improve the follow-up of patients and speed-up ongoing clinical trials.

Fig. 1. Two approaches for the diarization of conversational clinical interviews. The steps for the Speaker Enrollment Protocol are in Blue, and Green for the Speaker Role Recognition.



Yet, this methodology relies on the heavy burden of manual annotation to reach the necessary amount needed to draw significant conclusions. It is now indispensable to have robust speech processing pipelines to extract meaningful insights from these long naturalistic datasets [3]. Huntington's Disease represents a unique opportunity to design and test these speech algorithms for *Neurodegenerative diseases*. Indeed, individuals with the Huntington's disease can exhibit a large spectrum of *speech and language* symptoms [4] and it is possible to follow gene carriers even before the official clinical onset of the disease [5]. The first unavoidable computational tasks to extract speech and linguistic information from medical interviews is the diarization: (1) the *detection* of speaker-homogeneous portions of voice activity [6] and (2) the *identification* of speaker [7]. Speaker turns are clinically informative for diagnostic in Huntington's Disease [8, 4].

First, a number of studies are trying to solve this problem directly from the audio signal and linguistic outputs, also referred to as *Speaker Role Recognition*. They are taking advantage of the specificities (ex: prosody, specific vocabulary,

⋆ Equal contribution. We are very thankful to the patients that participated in our study. We thank Katia Youssov, Laurent Cleret de Langavant, Marvin Lavechin, and the speech pathologists for the multiple helpful discussions and the evaluations of the patients. This work is funded in part by the Agence Nationale pour la Recherche (ANR-17-EURE-0017 Frontcog, ANR-10-IDEX-0001-02 PSL*, ANR-19-P3IA-0001 PRAIRIE 3IA Institute) and Grants from Neuratris, from Facebook AI Research (Research Gift), Google (Faculty Research Award), Microsoft Research (Azure Credits and Grant), and Amazon Web Service (AWS Research Credits).

adapted language models) of each role in the different domains: Broadcast news programs [7], Meetings [9], Medical conversations [10], Child-centered recordings [11, 12].

Another approach relies on *Speaker Enrollment* [13, 14], it aims to check the identity of a given speech segment based on a enrolled speaker template. Our study differ from these studies as they are evaluating their pipelines with already segmented speaker-homogeneous speech segments. Another related approach is *Personal VAD* (Voice Activity Detection) model from [15] where they used enrolled speaker template to detect speech segments from each individual speaker.

None of these approaches have been compared under the same evaluation metric, despite prior works aiming at solving both these tasks [16] and their high degree of similarities.

Here in this paper, we aimed to *detect* automatically the portions of speech and to *identify* the speakers in medical conversation between Psychologists and Interviewees. These interviewees are either Healthy Controls (C), gene carriers without overt manifestation of Huntington's Disease (preHD) and manifest gene carriers of Huntington's Disease (HD). We introduced a novel way to split the datasets so that we are now capable to compare two different speech processing approaches to deal with these 2 problems (Figure 1): *Speaker Role Recognition* and *Speaker Enrollment Protocol*. We showed the clinical relevance of these pipelines with the extraction speech markers that have been found predictive in Huntington's Disease.

2. DATA, EVALUATION SPLITS, METRICS

2.1. Dataset

Ninety four participants were included from two observational cohorts (NCT01412125 and NCT03119246) in this ancillary study at the Hospital Henri-Mondor Créteil, France): 72 people tested with a number of CAG repeats on the Huntingtin gene above 35 (CAG > 35), and 22 Healthy Controls (C). Mutant Huntington gene carriers were considered premanifest if they both score less than five at the Total Motor score (TMS) and their Total functional capacity (TFC) equals 13 [17] using the Unified Huntington Disease Rating Scale (UHDRS). All participants signed an informed consent and conducted an interview with an expert psychologist. Therefore in the diarization setting, there are two roles in each interview: a *Psychologist* and an *Interviewee*. The speech data were annotated with Seshat [18] and Praat [19] softwares. The dataset is composed of $K = 94$ interviews $\mathcal{I}_{1\ldots K}$. We designed a new way to split of speech dataset to compare different diarization approacges: an end-to-end Speaker Role Recognition model and a Speaker Enrollment pipeline (See Figure 2). The dataset is split in three sets which we refer to *meta-train set* M_{train}, *meta-dev set* M_{dev} and *meta-test set* M_{test} with the ratio of 60%, 20%, and 20%, respectively. Interview $I \in \mathcal{I}_{1\ldots K}$ is composed of N_I segments $I = \{U_0, U_2, \ldots, U_{N_I}\}$. Each segment U_i is pronounced by a speaker s_i. We summarized the corpus statistics in Table 1.

Each interview I in the *meta-dev* and *meta-test* is split in two sets which we refer *dev set* X_{dev} and *test set* X_{test}. X_{test} is always kept fixed through all experiments, and we study the influence of the size of the X_{dev} based on T_{dev} that filters the segments (cf Figure 2).

All the data from the *meta-train* set M_{train} is used to train or fine-tune the neural network models for voice activity detection, speaker change detection, speaker role recognition, and speaker enrollment. The dev set X_{dev} of the *meta-dev* set M_{dev} and the dev set X_{dev} of the *meta-test* set M_{test} are only used for the speaker enrollment experiments, to build the template representation of each speakers. The results on the test set X_{test} of the *meta-dev* set M_{dev} are used to select all the hyper-parameters and select the best model for each experiment. The final comparison is done with the test set X_{test} of the *meta-test* set M_{test}.

Fig. 2. Illustration of the data split with 4 interviews. Each line I_i represents an interview between the Interviewee and the Psychologist. The elevation of each row indicates 'who speaks when'. The segments can overlap.

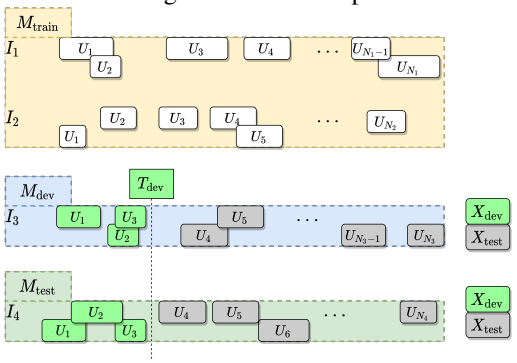


Table 1. Corpus statistics. P stands for Psychologist. IT stands for Interviewee. *Dur* stands for Duration and reported in hour.

	M_{train}	M_{dev}	M_{test}
#Interviews	57	18	19
#Segments IT	21400	7503	7788
#Segments P	4184	1381	1517
Dur Role IT (h)	7.65	3.02	3.21
Dur Role P (h)	3.54	1.14	1.15
Dur Overlap (h)	1.10	0.50	0.45
#(C/preHD/HD)	13/11/33	4/4/10	5/3/11

2.2. Metrics

To compare the final performance of the pipeline systems, we use the Identification Error Rate (IER) taking into account both the segmentation errors and confusion errors. We obtained the IER with `pyannote.metrics` [20]: IER $= \frac{T_{\text{false alarm}} + T_{\text{missed detection}} + T_{\text{confusion}}}{T_{\text{Total}}}$ The $\frac{T_{\text{confusion}}}{T_{\text{Total}}}$ component in the IER is related to the Miss-classification Rate (MR%) used in Speaker Role Recognition study [21], which is based on Frames and not duration of the turns. We compared the dif-

ferent approaches as a function of the size of the enrollment T_{dev} in Figure 3.

3. METHODS

3.1. Speaker Role Recognition

We adapted the approach from [11] for the Speaker Role Recognition. We trained on M_{train} a unique model to detect each role (Psychologist,Interviewee), and selects the best epoch on M_{dev}. This is a multi-label multi-class segmentation problem. A threshold parameter for each role is optimized on the Meta-dev set M_{dev} for the two output units of the model. Therefore the two classes can be activated at the same time, i.e. we can also detect overlapped speech.

To solve and model this task, we used SincNet filters [22] to obtain adapted speech features vectors from the audio signal. The SincNet output is fed to a stack of 2 bi-recurrent LSTM layers with hidden size of 128, then pass to a stack of 2 feed-forward layers of size 128 before a final decision layer. We used a binary cross-entropy loss and a cyclic scheduler as training procedure. The hyper-parameters to train our model can be found (here).

3.2. Speaker enrollment protocol

The Speaker enrollment protocol can be decomposed into four tasks: (1) Voice Activity Detection (2) Speaker Change Detection, (3) Enrollment, (4) Identification. We extended the speech processing toolkit from [23] `pyannote.audio` to run our experiments. Clinical laboratories can not all re-train in-domain speech processing models due to data scarcity or a lack of computational resources. Therefore, we evaluated pretrained models on open-source datasets and transfer models on our dataset to evaluate these out-of-domain performances with real clinical conversational conditions.

3.2.1. Voice Activity Detection

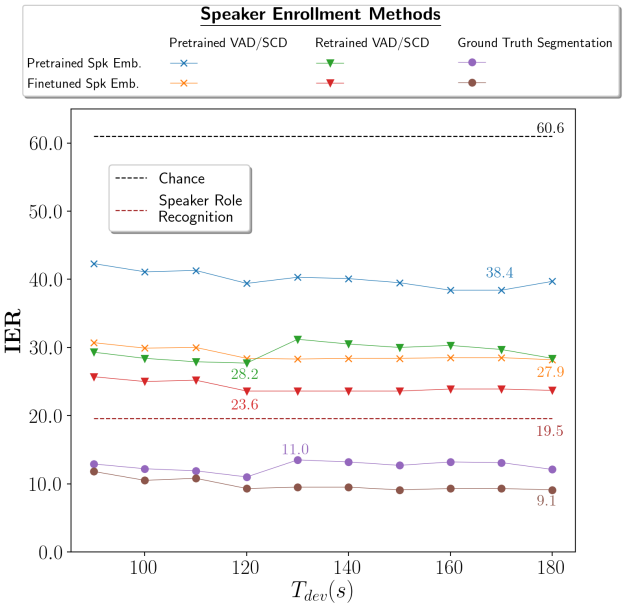
The first step is the Voice Activity Detection (VAD), i.e. obtain the speech segments in the audio signal. It can be modeled as an audio sequence labeling task. There are 2 classes (Speech or Non-Speech). The VAD labels for each interview I are the presence or not of a segment U_i at time t.

The model can be used already *Pretrained* or *Retrained* on the meta-train set M_{train} of our dataset. We choose the DIHARD dataset [24] as a potential pretrained dataset as it contains multiple source domain data (clinical interviews among them). When trained from scratch, the training is done for 200 pyannote epochs and the model is selected on the Meta-dev M_{dev}. The model is also composed of SincNet filters with 2 bi-recurrent LSTM layers and 2 feed-forward layers. The full specifications can be found here.

3.2.2. Speaker Change Detection

The second step is the Speaker Change Detection (SCD), i.e. obtain the moment when one of a speaker starts or stops talking. It can aslo be modeled as an audio sequence labeling task. There are 2 classes (Change or No-Change). The SCD labels for each interview I are the start or end of a segment

Fig. 3. Identification Error Rates for the different combination of approaches on the test set X_{test} of the meta-test set M_{test} as a function of the size of the enrollment parameter T_{dev}. *Spk Emb.*, *VAD*,*SCD* stand for Speaker Embedding, Voice Activity Detection and Speaker Change Detection. Best performance of each approach is displayed at the best T_{dev}.



U_i at time t. We also compared *Pretrained* on DIHARD and *Retrained* models. We used the same model as for the Voice Activity Detection. The full specifications can be found here.

Based on VAD and SCD outputs, we can now obtain for each Interview I a set of N'_I candidates speaker-homogeneous segments $\{\hat{U}_1, \dots \hat{U}_{N'_I}\}$.

3.2.3. Enrollment

In the enrollment stage, we need to get a Speaker Embedding function f_θ for our specific task. We combined SincNet filters and the X-vector architecture [13] as in [23]. For finetuning, we froze all layers and fine-tuned the last layer. We used the VoxCeleb2 dataset [25] as a pretraining dataset as it contains a diverse distribution of speakers and recording conditions.

Then, we used the set of segments from the dev set X_{dev} of the *meta-dev* and *meta-test* to build a template vector m_j for each speaker j in the interview I. X_{dev} contain a set of segments $U_{\text{enrollment speaker } j}$ from each speaker j. The start of each segment $U_{\text{enrollment speaker } j}$ needs to be smaller than T_{dev}. We computed the average of the representations for each speaker j:

$$m_j = \frac{1}{|U_{\text{enrollment speaker } j}|} \sum_{U \in U_{\text{enrollment speaker } j}} f_\theta(U) \qquad (1)$$

In principle, the more data you have to build template of each speaker, the easier it is to distinguish them. Thus, we studied the effect of the size of the enrollment based on the parameter $T_{dev} \in (90s, 100s, \dots, 180s)$ to build the template m_j.

3.2.4. *Identification*

For the identification stage, we use the function f_θ and the different representation m_j of the speakers from the enrollment stage. We used the following cosine distance D to build a scoring function and compare each segment $\hat{U} \in \{\hat{U}_1, \dots \hat{U}_{N_l'}\}$ to each template m_j:

$$D(\hat{U}, m_j) = \frac{1}{2}\left(1 - \frac{f_\theta(\hat{U})^\top m_j}{\left[\|f_\theta(\hat{U})\| \, \|m_j\|\right]}\right) \quad (2)$$

$$\text{argmin}_j \, D(\hat{U}, m_j) : \text{Selects Speaker } j \quad (3)$$

In addition, we analyzed topline performance of the speaker embedding models when the Ground Truth Segmentation is provided. Finally, we computed a chance baseline based on speaker Enrollment by randomly permutating all the cosine distances. Spearman correlation is computed to compare clinical markers extracted from our best system to ground truth extractions (Figures 4 and 5).

4. RESULTS AND DISCUSSIONS

Figure 3 shows results in term of IER for the different approaches. Both approaches greatly improved over chance. If we consider pipelines solving both segmentation and identification, our best performance is obtained using the Speaker Role Recognition approach with IER=19.5% while the Speaker Enrollment Protocol obtained at best IER=23.6% at $T_{dev} = 120s$, with Retrained VAD/SCD models and Finetuned Speaker Embedding. Even though, the Speaker Enrollment protocol has per-speaker templates, it is not surpassing the Speaker Role Recognition approach. The topline with Ground Truth Segmentation (IER=9.1%) indicated that Speaker Enrollment could benefit greatly from a better detection of speaker-homogeneous turns. Errors of Speaker Enrollment are accumulated through the steps and can not be recovered, while Speaker Role Recognition takes advantage of solving all steps together in an end-to-end approach.

Increasing the size of the Template Enrollment m_j for each speaker with T_{dev} lead to slight improvements to all Speaker Enrollment methods. The finetuning of the X-vector speaker embedding model with in-domain is especially crucial (ex: Based on retrained VAD/SCD the IER decreases from 28.2% to 23.6%). An additional ablation study on the size of the meta-training set M_{train} showed us that the IER goes from IER=19.5% to IER=26.5% for the Speaker Role Recognition model trained with only 10% of M_{train}. In previous studies in Huntington's Disease [4, 8], the Ratio of Silence and Statistics on utterances were informative to distinguish between classes of Individuals. These speech markers can be extracted directly from the predictions of the Speaker Role Recognition outputs. We computed the Ratio of Silence and the Standard Deviation of Duration of Utterances on the test set of the Meta-test set M_{test}. This computation was done both from the Ground Truth Segmentation and the segmentation provided by the Speaker role recognition system (Figures 4, 5. We observed that the automatic system outputs behaved differently as a function of clinical marker. The Ratio of Silence was better predicted (significant spearman correlation of $r = 0.579, p = 0.009$) than the SD of Duration of Utterances (non significant spearman correlation of $r = 0.325, p = 0.175$). Some bias of the predictive system might not hurt the IER metric but hurt the reliability of some clinical measures.

Fig. 4. Ratio of Silence from the Ground truth segmentation and from the best Speaker role recognition pipeline.

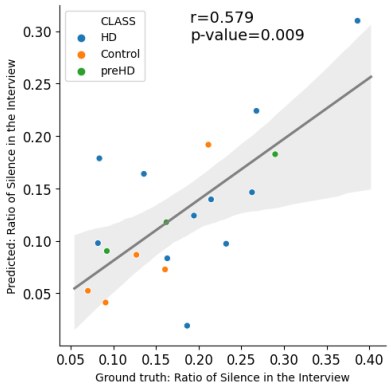
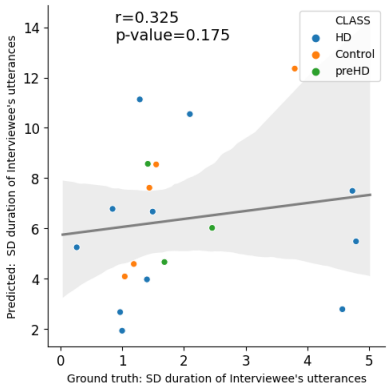


Fig. 5. Standard Deviations (SD) of the Duration of Utterances of Interviewees from the Ground truth segmentation and the best Speaker role recognition system.



5. CONCLUSION AND FUTURE WORK

Detection and Identification of speaker turns are fundamental problems in speech processing, especially in healthcare applications. While works studying these problems in isolation has provided valuable insights, in this work, we showed that Speaker Role Recognition was the most suitable approach for Interviewees at different stages of Huntington's Disease. For future work, we plan to investigate the use of these methods to derive robust biomarkers automatically and compare them to more classic approaches [26].

6. REFERENCES

[1] Sharon Ash and Murray Grossman, "Why study connected speech production," *Cognitive neuroscience of natural language use*, pp. 29–58, 2015.

[2] Katie Matton, Melvin G McInnis, and Emily Mower Provost, "Into the wild: Transitioning from recognizing mood in clinical interactions to personal conversations for individuals with bipolar disorder," *Proc. Interspeech*, pp. 1438–1442, 2019.

[3] Rimita Lahiri, Manoj Kumar, Somer Bishop, and Shrikanth Narayanan, "Learning domain invariant representations for child-adult classification from speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6749–6753.

[4] Adam P Vogel, Christopher Shirbin, Andrew J Churchyard, and Julie C Stout, "Speech acoustic markers of early stage and prodromal huntington's disease: a marker of disease onset?," *Neuropsychologia*, vol. 50, no. 14, pp. 3273–3278, 2012.

[5] Wolfram Hinzen, Joana Rosselló, Cati Morey, Estela Camara, Clara Garcia-Gorro, Raymond Salvador, and Ruth de Diego-Balaguer, "A systematic linguistic profile of spontaneous narrative speech in pre-symptomatic and early stage huntington's disease," *Cortex*, vol. 100, pp. 71–83, 2018.

[6] Simon Graf, Tobias Herbig, Markus Buck, and Gerhard Schmidt, "Features for voice activity detection: a comparative analysis," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 91, 2015.

[7] Benjamin Bigot, Julien Pinquier, Isabelle Ferrané, and Régine André-Obrecht, "Looking for relevant features for speaker role recognition," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[8] Matthew Perez, Wenyu Jin, Duc Le, Noelle Carlozzi, Praveen Dayalu, Angela Roberts, and Emily Mower Provost, "Classification of huntington disease using acoustic and lexical features," in *INTERSPEECH*, 2018, vol. 2018, pp. 1898–1902.

[9] Ashtosh Sapru and Fabio Valente, "Automatic speaker role labeling in ami meetings: recognition of formal and social roles," in *ICASSP*. IEEE, 2012, pp. 5057–5060.

[10] Nikolaos Flemotomos, Pavlos Papadopoulos, James Gibson, and Shrikanth Narayanan, "Combined speaker clustering and role recognition in conversational speech," *Proc. Interspeech 2018*, pp. 1378–1382, 2018.

[11] Marvin Lavechin, Ruben Bousbib, Hervé Bredin, Emmanuel Dupoux, and Alejandrina Cristia, "An open-source voice type classifier for child-centered daylong recordings," *arXiv preprint arXiv:2005.12656*, 2020.

[12] Nithin Rao Koluguri, Manoj Kumar, So Hyun Kim, Catherine Lord, and Shrikanth Narayanan, "Meta-learning for robust child-adult classification from speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8094–8098.

[13] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur, "Deep neural network embeddings for text-independent speaker verification," *Proc. Interspeech*, pp. 999–1003, 2017.

[14] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer, "End-to-end text-dependent speaker verification," in *ICASSP*. IEEE, 2016, pp. 5115–5119.

[15] Shaojin Ding, Quan Wang, Shuo-Yiin Chang, Li Wan, and Ignacio-Lopez Moreno, "Personal vad: Speaker-conditioned voice activity detection," in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 433–439.

[16] Paola García, Jesus Villalba, Hervé Bredin, Jun Du, Diego Castan, Alejandrina Cristia, Latane Bullock, Ling Guo, Koji Okabe, Phani Sankar Nidadavolu, et al., "Speaker detection in the wild: Lessons learned from jsalt 2019," 2019.

[17] Sarah J Tabrizi, Douglas R Langbehn, Blair R Leavitt, Raymund AC Roos, Alexandra Durr, David Craufurd, Christopher Kennard, Stephen L Hicks, Nick C Fox, Rachael I Scahill, et al., "Biological and clinical manifestations of huntington's disease in the longitudinal track-hd study: cross-sectional analysis of baseline data," *The Lancet Neurology*, vol. 8, no. 9, pp. 791–801, 2009.

[18] Hadrien Titeux*, Rachid Riad*, Xuan-Nga Cao, Nicolas Hamilakis, Kris Madden, Alejandrina Cristia, Anne-Catherine Bachoud-Lévi, and Emmanuel Dupoux, "Seshat: A tool for managing and verifying annotation campaigns of audio data," in *LREC*, Marseille, May 2020, * Equal contribution.

[19] Paul Boersma et al., "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, 2002.

[20] Hervé Bredin, "pyannote.metrics: a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems," in *Interspeech*, Stockholm, Sweden, August 2017.

[21] Nikolaos Flemotomos, Panayiotis Georgiou, David C Atkins, and Shrikanth Narayanan, "Role specific lattice rescoring for speaker role recognition from speech recognition outputs," in *ICASSP*. IEEE, 2019, pp. 7330–7334.

[22] Mirco Ravanelli and Yoshua Bengio, "Speaker recognition from raw waveform with sincnet," in *Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[23] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill, "Pyannote. audio: neural building blocks for speaker diarization," in *ICASSP*. IEEE, 2020, pp. 7124–7128.

[24] Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, and Mark Liberman, "The second dihard diarization challenge: Dataset, task, and baselines," *Proc. Interspeech*, pp. 978–982, 2019.

[25] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: a large-scale speaker identification dataset," *Telephony*, vol. 3, pp. 33–039, 2017.

[26] Rachid Riad, Hadrien Titeux, Laurie Lemoine, Justine Montillot, Jennifer Hamet Bagnou, Xuan Nga Cao, Emmanuel Dupoux, and Anne-Catherine Bachoud-Lévi, "Vocal markers from sustained phonation in huntington's disease," *arXiv preprint arXiv:2006.05365*, 2020.

Tab. 3.1.: Speaker Role Recognition Ablation study: Identification Error Rates on the test set X_{test} of the meta-test set M_{test} as a function of the percentage of interview in the meta-train set M_{train}. MD stands for Missed detection, FA for False Alarm and Conf. for Confusion

Percentage of M_{train}	MD	FA	Conf.	IER
10%	8.0	14.5	3.9	26.5
20%	7.8	12.4	3.8	24.0
50%	7.5	10.4	2.5	20.7
100%	7.1	10.2	2.3	19.5

In addition to the results in the submitted study, we ran an ablation experiment (Table 3.1) for the Speaker Role Recognition to measure the amount of data necessary. This ablation study informed us on the necessary amount of data to reach certain level of performance. Even though models are better than Chance, we found out that at least 50% of our dataset (28 Interviews) is necessary to outperform the Speaker Enrollment Protocol pipeline (IER of 20.7% vs 23.6%).

The analysis of the pattern of errors showed that the most important component is the False Alarm (FA), and a tenfold increase in dataset size allows to gain 4 points of FA. Therefore, most of the errors come from the voice activity detection part of the system. One of our hypothesis is that the system is confused by too much ambient noises from the hospital environment and thus potentially trigger too much positive presence of speech.

3.2 Identification of primary and collateral tracks

The detection of portions of speech is the first step, but it does not yield the full richness of the annotations produced by the speech pathologists. Especially, as seen in previous sections, vocalizations during spontaneous conversations can be decomposed into several broad categories, the tracks of communication proposed by Clark (1996). In this thesis, we propose to formalize this identification of tracks of communication as a computational task in speech processing: based on speech audio data, can we identify directly the tracks of communication? To solve this task, the algorithm has to digest speech audio date and locate in time these tracks of communication and categorize them for their specific roles. In fact, this task is related to the task of disfluency detection in the field of Natural Language Processing (NLP). However, in NLP, to the best of our knowledge there is no study wich starts from ASR outputs from audio signal before evaluating the NLP models, and evaluate their NLP models afterwards (Ferguson et al., 2015; Zayats et al., 2016). We propose

a new evaluation framework that allows direct comparison between text-based methods and audio-based methods.

3.2.1 Identification of primary and collateral tracks in stuttered speech

As a proof-of-concept of this methodology, we ran a complete study on a open-source database of stuttered speech (Bernstein & MacWhinney, 2018). In addition, we introduced new types of audio features, which can be directly detected from the audio waveform disfluencies, the Audio Span Features. We found that text-based approaches obtained the best performance overall, and that our newly introduced Audio Span Features outperformed the speech-based baselines. This study has been published in LREC 2020 proceedings (Riad, Bachoud-Lévi, et al., 2020b) and is included below.

Identification of primary and collateral tracks in stuttered speech

Rachid Riad[1,2*], **Anne-Catherine Bachoud-Lévi**[2], **Frank Rudzicz**[3], **Emmanuel Dupoux**[1]
1 CoML/ENS/CNRS/EHESS/INRIA/PSL Research University, Paris, France
2 NPI/ENS/INSERM/UPEC/PSL Research University, Créteil, France
[3] University of Toronto/Vector Institute/St Michaels Hospital/Surgical Safety Technologies Inc, Toronto, Canada
rachid.riad@ens.fr, bachoud@gmail.com, frank@spoclab.com, emmanuel.dupoux@gmail.com

Abstract

Disfluent speech has been previously addressed from two main perspectives: the clinical perspective focusing on diagnostic, and the Natural Language Processing (NLP) perspective aiming at modeling these events and detect them for downstream tasks. In addition, previous works often used different metrics depending on whether the input features are text or speech, making it difficult to compare the different contributions. Here, we introduce a new evaluation framework for disfluency detection inspired by the clinical and NLP perspective together with the theory of performance from (Clark, 1996) which distinguishes between primary and collateral tracks. We introduce a novel forced-aligned disfluency dataset from a corpus of semi-directed interviews, and present baseline results directly comparing the performance of text-based features (word and span information) and speech-based (acoustic-prosodic information). Finally, we introduce new audio features inspired by the word-based span features. We show experimentally that using these features outperformed the baselines for speech-based predictions on the present dataset.

Keywords: evaluation metrics, disfluency, stuttering, speech processing, audio features

1. Introduction

Around 6% percent of spoken words in non-pathological speech are categorised as disfluent (Tree, 1995) and this increases with the cognitive load of the speaker (Bortfeld et al., 2001; Lindström et al., 2008). Speaking in real time is a demanding activity, subject to cognitive constraints and pragmatic settings. Under time pressure, a word may not be retrieved, part of a sentence may be revised, unfilled and filled pauses may be inserted, words or part of words may be repeated. Some of these deviations can be viewed as the symptoms of sentence planning problems (McRoberts and Clark, 1996) or as the results of some *strategies* (Clark and Wasow, 1998) unfolded under speaker's control to signal something.

Stuttering is a severe case of speech pathology, that interrupts at a higher rate the flow of speech than in typical Speech Production. Indeed, *in addition* to the 'classic' disfluencies, stutterers can produce other forms of disfluency that are both quantitatively and qualitatively distinguishable from typical forms (Lickley, 2017).

Speech pathologists need to quantify all the disfluency events for clinical screening but also to assess potential treatments (Yaruss, 1997). A large number of factors and speaking settings influence stuttering behaviours and occurrences of disfluencies: interlocutor's characteristics (ex: age, relationship with the speaker), conversational settings (at home, at the hospital, at work), speaking tasks (ex: reading, dialogues, descriptions of scenes). The clinical assessment still rely heavily on subjective and one on one evaluation (Yaruss and Quesal, 2006). An automatic, reliable procedure would provide Speech Pathologists an objective comparison between clinical facilities and treatments. Besides, detecting automatically disfluencies and stuttering symptoms from speech, in different settings, could unlock in-home assessments and more frequent trainings for patients.

Two main issues arise from the literature around disfluency detection. The first one is the lack of public pathological annotated datasets. The second issue is the absence of a clear evaluation protocol for the automatic detection of disfluencies. This might be due to the fact that different communities have different applications in mind. Speech pathologists are interested in the automatic classification of type and duration of disfluencies (Yaruss, 1997). Most used proprietary datasets, with patients performing reading tasks and where disfluent and non disfluent parts have been balanced (Nöth et al., 2000; Yildirim and Narayanan, 2009; Oue et al., 2015). Researchers in the NLP community are interested in modelling the disfluencies for several reasons (Shriberg, 2001): text normalisation for downstream tasks such as Dependency Parsing or Semantic Role Labeling. In addition, NLP researchers are also interested in disfluency detection for affective computing (Tian et al., 2015) applications. They use features derived from transcribed text (Honal and Schultz, 2003) using Shriberg's formalism (Shriberg, 1994), and focus on the detection of disfluencies in non-pathological speech in telephonic conversation using datasets like Switchboard (Godfrey et al., 1992). Yet, the work from (Goldwater et al., 2010) demonstrated that words preceding disfluent interruption points also have high error rates for speech recognition systems. Finally, psycholinguists and clinicians are interested in the distribution and type of disfluencies, which could inform on speech and language production systems (Jackson et al., 2015) (Fromkin, 1971) as well as diagnosis. Obviously, for this kind of application, running interview-like speech with minimal annotations would be preferable. Since several hybrid text/speech systems have been proposed (Tran et al., 2018; Yildirim and Narayanan, 2009), we believe that a common evaluation method would be beneficial to bridge the gap between these research communities.

*Part of the work done at the Vector Institute during RR's summer internship

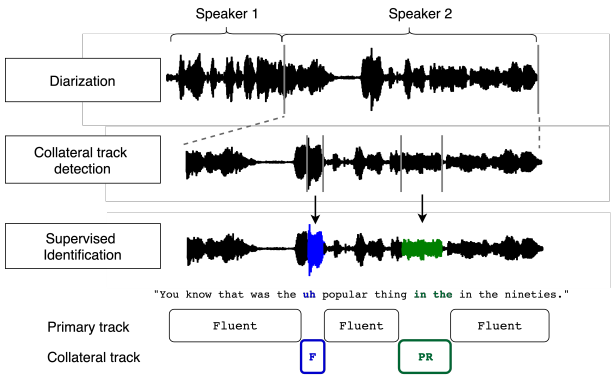


Figure 1: Schematic diagram of identification of the primary and collateral tracks of communication. In black the primary track (fluent), in other colors the words in the collateral track, in blue a filler, in green a phrase repetition.

First, we introduce a new framework for the evaluation of disfluency detection which would be relevant both for spontaneous and/or pathological speech using metrics combining insights from both NLP and Speech Technologies (ST) communities (Section 2). Second, we test these metrics on a new dataset obtained by force-aligning an annotated corpus of pathological speech (Section 3). All annotations in Praat format and code for evaluation will be released on the GitHub repository of the first author RR [1]. Third, we compare the performance of different baseline systems across textual and speech inputs on this dataset that were usually used in NLP and ST (Section 4). Four, to bridge the gap in performance between NLP and ST methods, we introduce new audio features that improve on the different frame-based baselines (Section 5).

2. Metrics: primary and collateral tracks

We take inspiration from the H. Clark's theory of speech performance (Clark, 1996, p. 255), which states that speakers communicate using two parallel tracks. The *primary track* contains the traditional linguistic content of the discourse while the *collateral track* contains additional signals regulating the communication channel itself. Among these signals we find delays, (un)filled pauses, rephrasing, mistakes, laughs, or vocal noises. The extraction of these two tracks from continuous speech can be decomposed in several engineering tasks (see Figure 1). First, a diarization (Anguera et al., 2012) component involves assigning stretches of signal (turns) to each speaker. This component is well studied and existing evaluation metrics can be used (see (Bredin, 2017) for a suite of diagnostic tools). Second, each turn is analysed in terms which sub-part contains collateral information (collateral detection task). This allows us to quantify how well the collateral track (only disfluencies in this work) is detected without specifying their category. The parsed segments can be evaluated in terms of a gold lexicon and a gold alignment. For this, we report the Detection Error Rate and the Detection F1-score. Third, each collateral sub-part is categorised into a small number of categories, which we restrict here to disfluency types

[1] https://github.com/Rachine/

Metrics	Formula
Detection Precision	$\frac{T_{\text{True Positive}}}{T_{\text{True Positive}}+T_{\text{false alarm}}}$
Detection Recall	$\frac{T_{\text{True Positive}}}{T_{\text{True Positive}}+T_{\text{missed detection}}}$
Detection F1-score	$2\frac{\text{detection precision}\times\text{detection recall}}{\text{detection precision}+\text{detection recall}}$
Detection Error Rate	$\frac{T_{\text{false alarm}}+T_{\text{missed detection}}}{T_{\text{Collateral Track}}}$
Identification Precision	$\frac{1}{5}\sum_i \text{Precision}_i$
Identification Recall	$\frac{1}{5}\sum_i \text{Recall}_i$
Identification F1-score	$2\frac{\text{identification precision}\times\text{identification recall}}{\text{identification precision}+\text{identification recall}}$
Identification Error Rate	$\frac{T_{\text{false alarm}}+T_{\text{missed detection}}+T_{\text{confusion}}}{T_{\text{Collateral Track}}}$

Table 1: Metrics used for the detection and identification of the collateral track. $T_{\text{false alarm}}$ is the duration of false alarm (e.g. primary track classified as collateral), $T_{\text{miss detection}}$ is the duration of missed detection (e.g. collateral track classified as primary), $T_{\text{Collateral Track}}$ is the total duration of the collateral track in the reference, $T_{\text{confusion}}$ is the total duration of the confusion between disfluency labels. Detection F1-score is computed as there are only two classes (primary and collateral). Precision_i and Recall_i are computed as the detection formula where the positive class is the i-th disfluency Table 2

(identification task). For that purpose we use the Identification F1-score and the Identification Error Rate. The formula to obtain these metrics are summarised in the Table 2.. The two error rate metrics are defined with the similar formula to that used in the Diarization Error Rate, which means that they can go over 100% (the denominator being restricted here to the collateral track). All these metrics were coded using the python toolkit *pyannote.metrics* (Bredin, 2017).

Another motivation for this framework comes from speech pathology research on Stuttering evaluation. Indeed, from these timing prediction of the primary and collateral track, it can be computed automatically the Speech Efficiency Score (SES) introduced in (Amir et al., 2018). This study demonstrated that this score, which is based on a time-domain analysis is closely equivalent to stuttering severity ratings done by speech pathologists. By solving the diarization task, and the disfluency detection task mentioned above, it is possible to obtain an estimation of the SES (see below the formula for the equivalence between our framework and their notations).

$$\begin{aligned} SES &= \frac{T_{\text{Primary Track}}}{T_{\text{Primary Track}}+T_{\text{Collateral Track}}} * 100 \\ &= \frac{T_{\text{Efficient time}}}{T_{\text{Total time}}-T_{\text{Silence}}} * 100 \end{aligned}$$

3. Dataset

We built on FluencyBank, a large-scale open source audiovisual dataset primarily used by clinical researchers to study fluency (Bernstein and MacWhinney, 2018), from which we selected and forced-aligned a consistent subset focused on stuttering. FluencyBank contains a collection of sub-datasets collected by different research groups to study typical and disordered fluency in infants and adults.

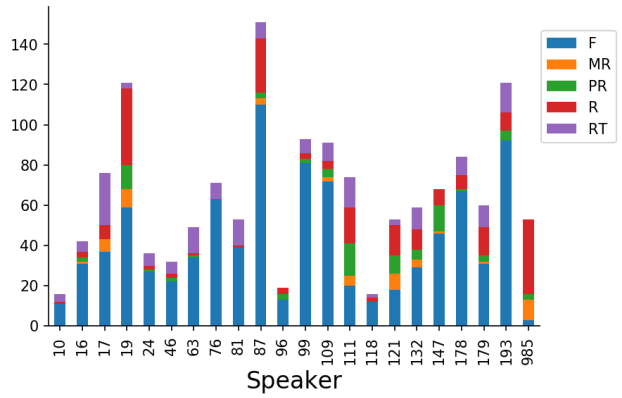


Figure 2: Distribution of disfluency in the FluencyBank-AWS per speaker. See table 2 for definitions of disfluencies.

We selected the Adult-who-Stutter(AWS) sub-dataset of [2]FluencyBank, which contains video recordings focused on patients. We excluded the recordings where the annotation was lacking, and we obtained 22 speaker video interviews (1429 utterances and 24693 words). The original recordings were done while the participants answered questions of the OASES elicitation protocol (Yaruss and Quesal, 2006), and transcriptions and disfluencies were done in the original dataset at the sentence level. Table 2 provides the five classes of disfluencies that we consider here. We provide some examples for each class of patients answering some questions of the protocol. In this work, we did not consider blocks, syllable repetitions or prolongations. Yet, our formulation with the primary and collateral tracks can easily be extended to these disfluencies.

We obtained the timings of the primary and collateral tracks by force-alignment at the phone level with the kaldi toolkit (Povey et al., 2011) with a HMM-GMM model. Figure 2 shows the distribution of disfluencies per speaker in the dataset. The total number of disfluencies and their types vary greatly across speakers.

Table 2: Collateral signals taxonomy (usually called disfluency) under consideration here in the FluencyBank dataset: *Italic* for the primary track and **Bold** for the collateral track.

Disfluency	**Example**
Filled pause (F)	*I was primarily* **uh** *focused on fluency.*
Single word Repetition (R)	**I** *I don't like switch word.*
Multi-Repetition (MR)	*I'm fortunate to* **be be be** *be in graduate school.*
Phrase Repetition (PR)	**they are** *they are so sweet.*
Retracing or Revision (RT)	**I ended when I was** *it ended when I was seventeen.*

4. Baselines: text versus speech predictions

Here, we provide two different kind of baseline systems with the purpose of comparing textual and acoustic approaches on the same metrics: (1) word-based systems, which assume that the input speech has been segmented into words, and aggregate textual and/or acoustic features over the entire span of each word (2), frame-based system which make decision on a frame-by frame basis from raw speech. Obviously, the latter kind of system cannot use textual features. All evaluations are performed with leave-one-speaker-out cross-validation, so that we can assess the generalisation to unseen speaker.

4.1. Detection from aligned speech: word-based systems

Word-based systems can incorporate both textual and acoustic features (See Table 3). As for textual features, we used token and span features which are common in the NLP community (Shriberg, 1994; Charniak and Johnson, 2001). As for acoustic/prosodic, we use summary statistics on duration energy and F0. All statistics and pooling are done using the timing alignment of each word w_i. The semantic representation and part-of-speech tags are extracted with (Honnibal and Montani, 2017). The number of syllables and phones are extracted with (Bernard et al., 2019).

Table 3: List of features in the word-based prediction.

Type	Core features and dimension	
Token	semantic representation	384
	part-of-speech (pos): p_i	19
	word position	1
Span	$w_i == w_{i+k}, k \in [-15, +15]^*$	30
	$p_i == p_{i+k}, k \in [-15, +15]^*$	30
	$w_i, w_{i+1} == w_{i+k}, w_{i+k+1}$ $k \in [-4, +4]^*$	8
	$p_i, p_{i+1} == p_{i+k}, p_{i+k+1}$ $k \in [-4, +4]^*$	8
Acoustic Prosodic	word duration (s)	1
	number of syllables	1
	number of phones	1
	high, low and total energy in filterbanks	3
	F0 mean, std, median, min, max 5%, 25%, 75% & 95% percentiles	9
	surrounding pause times (s)	2
	pitch breaks inside and around word limits	3

Such hand-crafted features have been used previously in the literature (Zayats et al., 2016; Ferguson et al., 2015) and have shown to improve the prediction performance. Indeed, neighbour words and prosodic cues are very informative about the disfluency events (Shriberg, 1994). In (Yildirim and Narayanan, 2009), they obtained that interrupting points in disfluencies are 98% associated with a pitch break. We obtained a similar result in the FluencyBank AWS dataset with 95% of the disfluent boundaries that match with a pitch break in a 100 ms vicinity. All these features are normalized with the $MaxAbsScaler$ from scikit-learn (Pedregosa et al., 2011) to avoid the loss of sparsity (specially in the span features).

We compared 5 different models (see Table 4). The latest work is focusing on sequence tagging prediction with Recurrent Neural Network architectures (Zayats et al.,

[2]https://fluency.talkbank.org/access/Voices-AWS.html

2016; Tran et al., 2018). We compared Forward and Bi-Directional architectures with Long-Short-Term-Memory (LSTM) Networks for disfluency detection and identification. The hidden dimension of the recurrent networks is set at 20. All the experiments are carried using the Adam training procedure with the default parameters (Kingma and Ba, 2014) and early-stopping on a held-out validation set of 20% of spoken utterances of the training speakers. In addition, we used a discriminative approach with classical machine learning classifiers. Every word is supposed to be independent: the goal is now to predict each word individually, without its neighbour's representation or prediction. The information listed in Table 3 is already obtained by the aggregation of local and long range information and could be sufficient to make predictions. We compare a standard classic Support Vector Machine with linear kernel (SVM) and a penalty term of $C = 0.025$, a L2 regularised logistic regression and a classic deep forward neural network (DNN) with 2 hidden layers with 100 and 50 hidden units.

In the results, we also report the Token F1 score 4 at the word level to compare the predictions. This differs from the F1-score reported for disfluency detection in the NLP community (Godfrey et al., 1992). Usually, NLP models tries to detect the disfluency and identify the subparts of each disfluency (reperandum, interegnum, repair), not the different categories of disfluencies.

4.2. Detection from raw speech: frame-based systems

Frame-based systems have no other information than speech. In principle, they are closer to what would be useful for clinical purposes, but obviously, the task is much harder. For comparison purposes we propose 3 baseline frame-based systems. Here, we evaluate frame-level prediction for the disfluency detection and identification as in (Oue et al., 2015). The patterns of the disfluencies in the speech signal can range from very local phenomenon (filled pauses) to long time-scales (retracing). Here, we investigate the predictions are made every 10 ms, in a bottom-up manner, using only local features.

Speech represented using a bank of 40 log-energy Mel-scale filters representing 25 ms of speech (Hamming windowed) every 10 ms. The Mel features are mean-variance normalised per file, using the VAD information. Besides, we extract prosodic features with the F0 trajectory and its first derivatives in a 50 ms window (obtaining a 56-dimension vector as F0 is computed every 1.8ms). These spectral and prosodic representation are concatenated to obtain a final 106-dimension vector representation every 10 ms. All the frame-based systems use a window of 7 stacked frames (Oue et al., 2015).

Based on these representations directly extracted from the signal, we follow a similar procedure as in the word-based predictions: we compare a standard classic Support Vector Machine with linear kernel (SVM) and a penalty term of $C = 0.025$, with a classic deep forward neural network (DNN) with 2 hidden layers with 100 and 50 hidden units. These approaches have been previously used in stuttering detection literature (Chee et al.,).

As in many machine learning problems (Lemaître et al., 2017), disfluency datasets have the attribute to be very imbalanced. The number of frames that are labelled fluent exceeds by a large margin all the others classes (92.7% of the frames are labelled as fluent). We evaluate a random undersampler technique (Lemaître et al., 2017) that discards randomly a large number of the majority class (here fluent) before training each model. This undersampling strategy has been used in Speech Technologies, yet, systems had not been evaluated on running speech datasets.

5. Audio Span Features

We want to improve the frame-based system using information over a long time span and replace the textual features with equivalent ones directly from the raw speech. We introduce here our Audio Span Features. The goal of these features is to obtain similar information as the span features from the word-based systems (Table 3).

Our main assumptions for disfluency events are: (1) Repetition-like disfluency events exhibit a common underlying structure property in the frequency domain, (2) filled pauses exhibit specific acoustic correlates with a steady frequency signature (Gabrea and O'Shaughnessy, 2000), (3) these filled pauses have usually adjacent unfilled pauses/silences (Daly, 1994).

That is why we posit that local neighbour-similarities in the frequency domain can approximate the span features for the word comparisons from Table 3. Besides, different chunk size can also inform on the different type of disfluencies (close and more local similarities are triggered by fillers versus spaced and long range similarities are triggered by repetitions of words).

Therefore, for every time-step t, for a given window-scale s, we compute the similarity $\psi(t, s, i)$ of the frequency representation x_t centered on t with its i-th closest neighbours. We compute this similarity with the N previous neighbours and the N next. The frequency representation $x_t \in \mathbb{R}^{40}$ is still the bank of 40 log-energy Mel-scale filters computed every $\delta = 10ms$. These neighbours are centered centered every $t_i^s = t + s \cdot i \cdot \delta$. The scale s is the (odd) number of stacked frames. So we denote by $x_t^s \in \mathbb{R}^{40 \cdot s}$ the concatenation of the s frames around t:

$$x_t^s = \begin{pmatrix} x_{t-(\frac{s-1}{2})\cdot\delta} \\ x_{t-(\frac{s-1}{2}+1)\cdot\delta} \\ \dots \\ x_{t+(\frac{s-1}{2}-1)\cdot\delta} \\ x_{t+(\frac{s-1}{2})\cdot\delta} \end{pmatrix}$$

Finally, our Audio Span Features can be computed:

$$\forall i \in [\![-N, N]\!]^*, \psi(t, s, i) = x_t^s \cdot x_{t_i^s}^s \cdot \frac{1}{n_s} \qquad (1)$$

We divided the similarity by $n_s = 40 \cdot s$ to normalise in the scale dimension and not privilege the bigger stacked frames similarities. We computed this similarity for 8 different scales with a logarithmic spacing for the different scales between 30 ms and 1 s ($s \in [101, 61, 37, 23, 13, 9, 5, 3]$). We choose these numbers to capture different orders of magnitude that characterise disfluent segments of speech: phones ($\sim 30ms$), words ($\sim 100ms$), sentences ($\sim 1s$).

Table 4: Results of the evaluation of detection and identification of primary and collateral track for the different approaches described in Section 4. The best scores for each metric for each condition (word vs frame based) are in **bold**, best metrics overall are underlined. For the evaluation of the Audio Span Features, we report the performance with a DNN model trained with the Standard sampler.

Model	NLP	Detection				Identification			
	Token F1	P	R	F1	Error Rate	P	R	F1	Error Rate
Word-based 4.1 (all features)									
Forward LSTM	0.416	0.823	0.595	0.691	0.623	0.717	0.518	0.601	0.701
Bi-LSTM	0.417	0.786	0.605	0.684	0.731	0.701	0.537	0.608	0.799
SVM-Linear	**0.569**	**0.966**	0.642	**0.771**	**0.381**	**0.905**	**0.599**	**0.721**	**0.424**
Logistic Regression	0.544	0.846	**0.645**	0.732	0.513	0.762	0.576	0.656	0.581
DNN	0.485	0.958	0.611	0.746	0.417	0.855	0.544	0.665	0.484
Frame-based 4.2 (Baseline Signal features only)									
Standard sampler + DNN	–	0.312	0.014	0.026	1.005	0.182	0.010	0.020	1.008
Undersampler + DNN	–	0.073	**1.000**	0.136	15.286	0.038	**0.520**	0.069	15.766
Standard sampler + SVM	–	0.150	0.086	0.109	1.502	0.116	0.067	0.086	1.520
Undersampler + SVM	–	0.077	0.838	**0.140**	12.529	0.025	0.288	0.048	13.079
Frame-based 5 (Audio Span Features)									
Audio Span Features + Standard Sampler + DNN	–	**0.864**	$\leq$ 1e$-$3	$\leq$ 1e$-$3	1.000	**0.818**	$\leq$ 1e$-$3	$\leq$ 1e$-$3	1.003
Audio Span Features + baselines features + Standard Sampler + DNN	–	0.488	0.063	0.112	**0.986**	0.450	0.059	**0.105**	**0.990**

We finally chose $N = 4$ for the number of neighbours ($i \in [-4, -3, -2, -1, 1, 2, 3, 4]$). Now, for every time-step t we concatenate the neighbour cross-similarities at different scales and obtain the final vector $\psi(t) \in \mathbb{R}^{4 \cdot 2 \cdot 8 = 64}$. See Figure 3 for a schematic representation of the computations. We evaluate these new Audio Span Features alone and along the acoustic and prosodic representation described in subsection 4.2. We report the evaluation with the Standard Sampler and the DNN model as in 4.2.

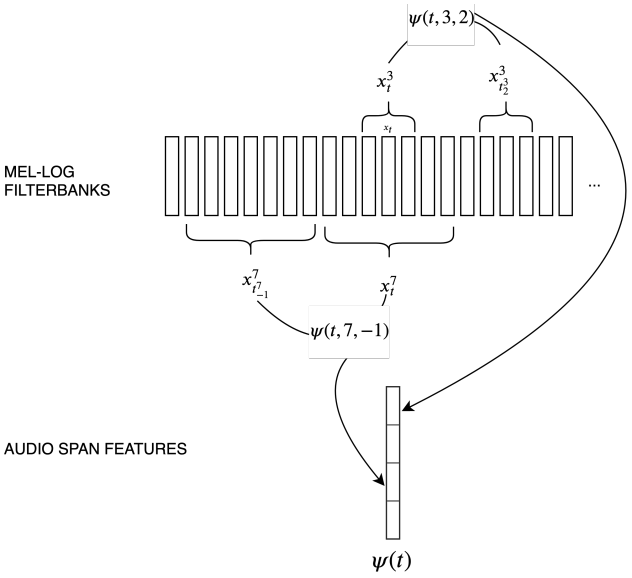


Figure 3: Audio Span Features: It is the concatenation of similarities between the current representation with the $2 \cdot N$ closest neighbours filterbank representations at different scales s.

6. Results and Discussions

Table 4 shows performances on the detection and identification of primary and collateral tracks. We first review the results from the word-based predictions methods when we take all features as input. Overall, we observed that Sequence-to-sequence models underperform compared to more classical machine learning classifiers. We hypothesise that if the data gets larger the LSTMs architectures might catch on compared to the classifiers. With respect to the F1-score in the detection and identification tasks, the LSTMs architecture are actually not that far from classifiers. Yet, there is an important drop in performance on Error Rates. This highlights the importance to take into account more than one composite score. Among these classifiers, good old SVM-linear model yields the best performances in almost all metrics (except in the Detection recall for the Logistic Regression).

Now, we turn to results from the frame-based methods. The results show a sharp drop in performance for all the systems in comparison to the word-based predictions. With the standard sampler, both the DNN and SVM are missing a large number of the disfluency events (Detection Recall at 0.014 and 0.086 respectively). The undersampling technique improves by a large margin the Detection Recall. By contrast, the Precision metrics, the Detection Error Rate and Identification Error Rate are way above 100%. This shows that with extreme class imbalance, frame-based methods that were previously shown reasonable performance in balanced datasets fail in a spectacular fashion. This highlights the importance of addressing the issue of the detection of disfluencies using running speech rather than artificially bal-

anced datasets or read-speech.

The new Audio Span Features alone demonstrate really poor performance and are missing almost all the disfluency events (Detection Recall and F1 lower than 0.001). However, the Audio Span Features along the acoustic and prosodic representation show the best performance on the frame-based system, especially in the identification task (Identification F1 0.118 and Error Rate below 1). The system misses a number of disfluent events (Low Detection Recall 0.063), but maintain a good precision level in comparison to the other frame-based baselines (Detection Precision 0.488 and Identification Precision 0.450). The Audio Span Features do not capture all the necessary information and can be improved. Especially, our Audio Span Features do not have grammatical information captured by the word-based span features. One of our hypothesis is that the Audio Span Features fail also to detect the revision/retracing disfluent events.

Table 5: Results of the evaluation of detection and identification of primary and collateral track for the different input features in the word-based predictions with a SVM model with linear kernel. The best error rates for each metric overall are in **bold**.

Features	Detection		Identification	
	F1	Error Rate	F1	Error Rate
Word-based (SVM-Linear)				
Token only	0.639	0.537	0.622	0.550
Span only	0.313	0.826	0.263	0.856
Acoustic only	≤ 0.001	1.000	≤ 0.001	1.000
Acoustic+Span	0.314	0.825	0.269	0.852
Acoustic+Token	0.646	0.529	0.628	0.543
Token+Span	**0.772**	0.382	**0.720**	0.427
All (line 3 from Table 4)	0.771	**0.381**	0.721	**0.424**

To better understand the impact of the input features, we ran an ablation study for the word-based predictions, see Table 5. We compare the different combinations of features as defined in Table 3. First, the acoustic/prosodic features are not informative on their own to predict disfluencies. Span are better but still not reach the full model performance. They might be more suitable to detect the repetition-like disfluencies but not necessarily for the Filled pauses. Obviously, the token based representations have a clear advantage especially for Filled pauses. The acoustic/prosodic features provide a little gain for the span and token representation, but the combination of span and token is already sufficient on its own and gets very close to the combination of all features.

This study could orient future work to bridge the gap in performance between our frame-based predictions and word-based predictions. Indeed, the semantic and grammatical part-of-speech features play an important part in the good results of the word-based systems. To obtain such features from the signal, we could build an Automatic-Speech-Recognition pipeline suited for Stutterers and obtain the word2vec representations (Mikolov et al., 2013). Or we can obtain such semantic information directly from the signal (Chung and Glass, 2018).

7. Conclusions

In this work, we investigated a framework to evaluate the disfluencies detection system in stuttered speech. First, we prepared and adapted an open dataset of Adult-Who-Stutter used by clinical researchers, for the task of disfluency detection from running speech. We provided a suite of metrics based on the forced alignment, that enables to compare word-based predictions and frame based-predictions. This allows the direct comparison between different type of approaches. Finally, we compared different baselines systems with textual or acoustic input features, and using word- or frame based pooling of information. The word-based systems show superior performance, illustrating the need (1) to improve frame-based aggregation of information over a long time span and (2) replace textual features with equivalent ones that can be derived automatically from raw speech. Finally, we introduced new Audio Span Features that show the best performances for the frame-based methods.

8. Acknowledgements

This work is funded through a Facebook AI Research grant, and supported by INRIA, as well as grants ANR-10-IDEX-0001-02 (PSL*) and ANR-17-EURE-0017 and grants from FacebookAI Research (Research Grant), Google (Faculty ResearchAward) and Microsoft Research (Azure Credits and Grant).

9. Bibliographical References

Amir, O., Shapira, Y., Mick, L., and Yaruss, J. S. (2018). The speech efficiency score (ses): A time-domain measure of speech fluency. *Journal of fluency disorders*, 58:61–69.

Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., and Vinyals, O. (2012). Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356–370.

Bernard, M., Isn0gud, and Benjumea, J. (2019). bootphon/phonemizer: phonemizer-1.0.1, March.

Bernstein, R. N. and MacWhinney, B. (2018). Fluency bank: A new resource for fluency research and practice. *Journal of fluency disorders*, 56:69.

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., and Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and speech*, 44(2):123–147.

Bredin, H. (2017). Pyannote. metrics: a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems. In *Annual Conference of the International Speech Communication Association*.

Charniak, E. and Johnson, M. (2001). Edit detection and parsing for transcribed speech. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–9. Association for Computational Linguistics.

Chee, L. S., Ai, O. C., and Yaacob, S.). Overview of automatic stuttering recognition system.

Chung, Y.-A. and Glass, J. (2018). Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech. *Proc. Interspeech 2018*, pages 811–815.

Clark, H. H. and Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive psychology*, 37(3):201–242.

Clark, H. H. (1996). *Using Language*. Cambridge University Press.

Daly, N. A. (1994). Acoustic study of verbal hesitations, filled pauses, and unfilled pauses in spontaneous speech. *The Journal of the Acoustical Society of America*, 95(5):2949–2949.

Ferguson, J., Durrett, G., and Klein, D. (2015). Disfluency detection with a semi-markov model and prosodic features. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 257–262.

Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances. *Language*, pages 27–52.

Gabrea, M. and O'Shaughnessy, D. (2000). Detection of filled pauses in spontaneous conversational speech. In *Sixth International Conference on Spoken Language Processing*.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520. IEEE.

Goldwater, S., Jurafsky, D., and Manning, C. D. (2010). Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 52(3):181–200.

Honal, M. and Schultz, T. (2003). Correction of disfluencies in spontaneous speech using a noisy-channel approach. In *Eighth European Conference on Speech Communication and Technology*.

Honnibal, M. and Montani, I. (2017). spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Jackson, E. S., Yaruss, J. S., Quesal, R. W., Terranova, V., and Whalen, D. (2015). Responses of adults who stutter to the anticipation of stuttering. *Journal of Fluency Disorders*, 45:38–51.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5.

Lickley, R. (2017). Disfluency in typical and stuttered speech. *Fattori sociali e biologici nella variazione fonetica-Social and biological factors in speech variation*.

Lindström, A., Villing, J., Larsson, S., Seward, A., Åberg, N., and Holtelius, C. (2008). The effect of cognitive load on disfluencies during in-vehicle spoken dialogue. In *Ninth Annual Conference of the International Speech Communication Association*.

McRoberts, G. W. and Clark, H. H. (1996). *The role of lexical access in spontaneous speech disfluencies*. Ph.D. thesis, ASA.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Nöth, E., Niemann, H., Haderlein, T., Decher, M., Eysholdt, U., Rosanowski, F., and Wittenberg, T. (2000). Automatic stuttering recognition using hidden markov models. In *Sixth International Conference on Spoken Language Processing*.

Oue, S., Marxer, R., and Rudzicz, F. (2015). Automatic dysfluency detection in dysarthric speech using deep belief networks. In *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, pages 60–64.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, 12:2825–2830.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glem-

bek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit. Technical report, IEEE Signal Processing Society.

Shriberg, E. E. (1994). *Preliminaries to a theory of speech disfluencies*. Ph.D. thesis, Citeseer.

Shriberg, E. (2001). To 'errrr'is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31(1):153–169.

Tian, L., Moore, J. D., and Lai, C. (2015). Emotion recognition in spontaneous and acted dialogues. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 698–704. IEEE.

Tran, T., Toshniwal, S., Bansal, M., Gimpel, K., Livescu, K., and Ostendorf, M. (2018). Parsing speech: A neural approach to integrating lexical and acoustic-prosodic information. In *Proceedings of NAACL-HLT*, pages 69–81.

Tree, J. E. F. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of memory and language*, 34(6):709–738.

Yaruss, J. S. and Quesal, R. W. (2006). Overall assessment of the speaker's experience of stuttering (oases): Documenting multiple outcomes in stuttering treatment. *Journal of fluency disorders*, 31(2):90–115.

Yaruss, J. (1997). Clinical measurement of stuttering behaviors. *Contemporary Issues in Communication Science and Disorders*, 24(24):33–44.

Yildirim, S. and Narayanan, S. (2009). Automatic detection of disfluency boundaries in spontaneous speech of children using audio–visual information. *IEEE transactions on audio, speech, and language processing*, 17(1):2–12.

Zayats, V., Ostendorf, M., and Hajishirzi, H. (2016). Disfluency detection using a bidirectional lstm. *arXiv preprint arXiv:1604.03209*.

3.2.2 Identification of primary and collateral tracks in dysarthric speech

In the previous sections, we described the first attempts to identify the turn-takings in interviews with patients with HD. Then, we presented the proof-of-concept for the identification of tracks of communication in population demonstrating a lot of these events, pathological or not, stuttering individuals.

Yet, in the section 3.1, we assumed the turn-takings were already provided. But as we observed in previous section, getting these turn-takings automatically from waveform is not easy. Therefore, here we made the first attempts to retrieve turn-takings *and* the tracks of communication from the raw waveform at the same time.

The extension of tracks of communication framework to dysarthric speech and application to the BasalVoice database is still ongoing, and the results in this section are preliminary. In this part we focused only on the spontaneous tasks, which are harder for the identification of turn-takings , but also present spontaneous utterances with different tracks of communication.

Based on the turn-takings and the segment tree that model the tracks of communication inside an utterance, we can obtain a segmentation of the audio waveforms into three classes: Speech of the neuropsychologist (NP), Track 1 of communication of the interviewee and Track 2 of communication of the interviewee (Secondary, Ex-official, and Incidental elements). For more information on these tracks see Section 2.3 and Figure 2.16. The identification task here is to obtain from the raw waveform these three classes and their boundaries.

Tab. 3.2.: Identification Error Rates on the test set. Metrics are computed thanks to (Bredin, 2017). Best results per metric are represented in **bold**.

Meta-class	Window(s)	Best epoch	F1			IER
			Track1	Track2	NP	
No	0.50	40.00	0.85	0.59	0.76	0.39
No	2.00	370.00	0.80	0.42	0.70	0.68
No	4.00	390.00	0.82	0.47	0.71	0.55
No	6.00	380.00	0.79	0.46	0.68	0.64
No	8.00	390.00	0.82	0.48	0.73	0.53
No	10.00	330.00	0.79	0.44	0.65	0.68
Yes	0.50	70.00	0.85	0.58	0.76	0.39
Yes	1.00	70.00	**0.86**	**0.60**	0.78	0.38
Yes	2.00	140.00	**0.86**	0.59	**0.78**	**0.37**

To solve and model this task, we used SincNet filters (Ravanelli & Bengio, 2018) that adapt to the task at hand. The SincNet output is fed to a stack of 2 bi-recurrent Long short-term memory (LSTM) layers with hidden size of 128, then pass to a stack of 2 feed-forward layers of size 128 before a final decision layer. We used a binary cross-entropy loss and a cyclic scheduler as training procedure.

The current task is modelled as a multi-label classification problem, this means that multiple classes can be activated at the same time. This allows two things: (1) the detection of overlapping speakers is possible, (2) we can introduce additionnal classes as objectives, that group sub-classes to boost performance and training of the neural network. These additional classes are referred as *meta-class*. The natural way to group the three classes (Speech of the interviewer, Track 1 of the interviewee, Track 2 of the interviewee) and to add a meta class is to group (Track 1 of the interviewee, Track 2 of the interviewee) as (Speech of the interviewee). In this scenario, with the addition of the meta-class, there are 4 classes to identify (Speech of the interviewer, Speech of the interviewee, Track 1 of the interviewee, Track 2 of the interviewee). Therefore, we started to evaluate different window size of analysis and also the inclusion of the meta-class of speech of the interviewee.

In addition, as underlined by (Oue et al., 2015), the choice of the window size for the identification of disfluency is crucial to obtain good performance. That is why investigate several size of the window given to the bi-recurrent LSTM layers. We report in Table 3.2 the preliminary results for these experiments. First, we found out that the window size asymmetrically impact the identification of Track 1 and Track 2, when we do not use the meta-class of interviewee. In addition, we observed that the use of meta-class sped up the training and also obtained the best performance for all metrics. Performances to obtain the Track 1 are relatively high, with a F1=0.86 for the current best approach. On the other Track 2 is more difficult to obtain. Statistics from Track 1 will be in theory more reliable and this gives hope to derive automatically speech markers from our model.

4 Vocal and linguistic markers in HD

> *Signs can be either signals or symptoms.*
>
> — **Herbert Clark**
> Using Language

In previous chapter, we described the first algorithms to detect turn-takings and different tracks of communications in conversations with individuals with HD. However, the data obtained through annotations or algorithms along the speech signal is unstructured, and there is the need to capture the relevant part for the measure of HD symptoms and discard irrelevant ones. In addition, there is a need of *formatting* of these spoken language markers, so that they can be fed to classic statistics and machine learning methods. This formatting refers to processes to summarize the information about speech and language to make inferences about our data.

Our literature review around speech and language in the introductory chapter revealed that several steps of spoken language production are impaired in HD. In this section, we present the first analyses to confirm or infirm these previous studies, but also extend the knowledge concerning spoken language markers as potential biomarkers in HD. The validation of biomarkers can be performed at different levels of analysis, group vs individual level, and there are also different levels of validity. This validity "ladder" of biomarkers can be formulated as specific machine learning tasks, and we list them from lower to higher importance for Clinical Neurology:

Diagnostic Detection of the official clinical onset of HD: Classification of HD vs preHD (or HD vs HC). Detection of future HD: Classification of future manifestation of the disease, preHD vs HC.

Disease evaluation Measurement of the severity of the different symptoms of HD through time: Regression of HD's clinical scores, such as cUHDRS, TFC or TMS.

Prognostic Prediction of disease onset of HD and future phenotype: Regression of the date of the future official clinical onset of preHD, Regression of future clinical scores of preHD or HD.

If possible, the accuracy on these tasks needs to remain high even if individuals go through potential treatments, and therefore track the success of clinical trials. In addition, it is important to evaluate the timing to perform and annotate a given task in order to evaluate the feasability of a speech task in clinical conditions.

In Section 4.1, we tried to replicate and extend previous results obtained in other HD cohorts concerning the study of sustained phonations and vowel distortion. Then, we evaluated extensively the potential of the combination of two tasks with minimal and high cognitive load, that remain short to perform and annotate, to predict different clinical scores. Then, we report preliminary results concerning the different speech events that characterize spontaneous speech in HD. Finally, we also developed novel methodologies to investigate to which extent emotional spoken language production is impaired in HD (See Section 4.2). These methods extends the current knowledge around emotions and spoken language production in HD. This chapter is also composed of different research papers, and we include additional experiments immediatly after to complete these studies.

4.1 Validation of spoken language markers

As underlined in Section 1.2.1 in the introductory chapter, the validation of biomarkers can be performed at the group or individual level. Here, we extracted different spoken language markers from the different tasks of the BasalVoice database and tested their potential as biomarkers. In this thesis, we propose when it was possible to go even further in the validation of biomarkers. First, we proposed to validate the validation of spoken language markers to predict the clinical scores and not only their clinical status. The clinical status in HD can be performed with a 100% accuracy with the genetic testing. The current burdensome evaluation of the different symptoms is more difficult and require the frequent visits of HD at the hospital and long meetings with Neurologists and Neuropsychologists (See Figure 2.1).

4.1.1 Vocal markers from sustained phonation in Huntington's Disease

Markers from sustained phonation have been shown promising results in HD, especially due to the progression of dysarthric symptoms in HD (Rusz, Saft, et al., 2014; Rusz, Klempir, Baborova, et al., 2013; Romana et al., 2020; Vogel et al., 2012; Rusz, Klempir, Tykalova, et al., 2014). Such a simple task is appealing due to the potential of remote collection and the fact there is no need of complex annotations to extract features. Here, we ran a replication of studies that have reported very high performance to distinguish HD from HC (Rusz, Klempir, Baborova, et al., 2013), and preHD from HC (Rusz, Saft, et al., 2014). We also propose the use of different features, measuring the spectro-temporal modulations of sounds, that have been shown useful in the assessment of intelligibility (Elhilali et al., 2003) and in dysarthria (Moro-Velazquez et al., 2015; Kodrasi & Bourlard, 2020; Janbakhshi et al., 2019). We add to implement all the speech features from scratch as there were no available official implementation, and we open-sourced the code to replicate our results https://github.com/bootphon/sustained-phonation-features.

We evaluated these features at the group and individual level to separate the different groups HC, preHD and HD. In addition, we measured the capabilities of the different set of features to predict at the individual level motor, functional and global measures in HD. The results of this study have been published in Interspeech 2020 proceedings (Riad, Titeux, et al., 2020) and is included below.

Vocal markers from sustained phonation in Huntington's Disease

Rachid Riad[1,2], *Hadrien Titeux*[1], *Laurie Lemoine*[2,3], *Justine Montillot*[2,3], *Jennifer Hamet Bagnou*[2,3], *Xuan Nga Cao*[1], *Emmanuel Dupoux*[1], *Anne-Catherine Bachoud-Lévi*[2,3]

[1] CoML/ENS/CNRS/EHESS/INRIA/PSL Research University, Paris, France
[2] NPI/ENS/INSERM/UPEC/PSL Research University, Paris, France
[3] Huntington's Disease National Reference Center, Neurology Department, Henri-Mondor Hospital, APHP, Créteil, France

rachid.riad@ens.fr

Abstract

Disease-modifying treatments are currently assessed in neurodegenerative diseases. Huntington's Disease represents a unique opportunity to design automatic sub-clinical markers, even in premanifest gene carriers. We investigated phonatory impairments as potential clinical markers and propose them for both diagnosis and gene carriers follow-up. We used two sets of features: Phonatory features and Modulation Power Spectrum Features. We found that phonation is not sufficient for the identification of sub-clinical disorders of premanifest gene carriers. According to our regression results, Phonatory features are suitable for the predictions of clinical performance in Huntington's Disease.

Index Terms: Huntington's disease, Phonation, Pathological speech processing, Dysarthria, Modulation Power Spectrum.

1. Introduction

Huntington's disease is an autosomal dominant neurodegenerative disease with a complete penetrance [1]. The disease is characterised by a triad of symptoms (motor, cognitive and psychiatric [2]) that leads to the progressive disability functioning of the individual with HD. The age of the clinical onset is highly variable (mean around 45 year-old) and the symptoms gradually worsen over 15 to 20 years until the death. The motor assessment allows defining a cut-off score that splits between premanifest and manifest gene carriers and thus the disease onset. Therefore, some of the deficits cognitive or psychiatric impairments or sub-clinical disorders may appear long before the disease manifests.

The automatic identification of pre-symptomatic carriers of the gene of Huntington's Disease and the prediction of the clinical scores are of particular interest for the neurological practice [3]. Indeed, having ecological markers of these clinical endpoints may: (1) help to prevent detrimental and harmful life events, (2) speed-up clinical trials, (3) increase the understanding of the disease in ground truth condition. Here, by using speech analysis of sustained phonation, we aim at distinguishing, in individuals, stage of the disease and predicting their clinical scores, as assessed by neurologists and neuropsychologists.

We rely on the fact that individuals with Huntington's Disease can exhibit different disorders during speech [4, 5] and language production [6, 7]. Motor speech disorders in Huntington's Disease are commonly referred to as *hyperkinetic* dysarthria [8]: variable rate, abnormal prosody, imprecise consonants and distorted vowels, phonation deviations, and sudden forced breath. Yet, there is some recent evidence that the speech motor impairments in Huntington's Disease are highly heterogeneous [9].

Here, we considered the speech features collected from simple recordings of the French vowel /a/ uttered in a sustained manner for as long as possible, in regular clinical conditions. We collected production of these vowels for Healthy Controls (C), gene carriers without overt manifestation of Huntington's Disease (preHD) and manifest gene carriers of Huntington's Disease (HD). We modelled each vowel with several features, which are impaired in hyperkinetic dysarthria. We investigated two sets of features: (1) Phonatory features already proven useful to distinguish gene carriers from control (preHD vs Controls [4] and HD vs Controls [10]), (2) Modulation Power Spectrum features to measure the modulations that characterise speech intelligibility [11] and roughness [12]. First, we conducted a statistical analysis at the group level of these features. Then, based on these sets of features, we used regularised linear models to predict the group and clinical scores of the patients.

Our statistical analyses and classification results showed that HD patients distinguished from controls, whereas the boundaries around preHD are more blurred. The Modulation Power Spectrum features complemented the Phonatory features to help identify preHD and improve the F1-score. In contrast, we observed that the Phonatory features have the best predictive capabilities of the clinical scores.

2. Related work

Previous studies have used sustained phonation for the assessment of various neurodegenerative disorders such as Parkinson's disease [13], Amyotrophic Lateral Sclerosis [14] and also Huntington's Disease [10]. These studies have extracted a number of hand-crafted features from the sustained phonation and built discriminative models to distinguish patients from controls. Yet, collecting pre-symptomatic speech data on other acquired neurodegenerative diseases is difficult, and can only be done retrospectively [15].

Rusz et al. [4] provided the closest study to ours using

Table 1: *Participants demographics and clinical scores*

	Controls	Huntington's disease Gene carriers	
Sub-groups	C	PreHD	HD
N	24	16	45
Gender	12F/12M	9F/7M	27F/18M
Age (years)	54.1 (8.9)	50.2 (12.2)	53.9 (11.3)
CAG Triplets	≤ 35	41.4 (1.5)	44.3 (3.4)
cUHDRS [16]	—	17.6 (1.4)	9.0 (3.6)
TFC [17]	—	13.0 (0.0)	10.3 (2.2)
TMS	—	0.38 (1.0)	36.0 (15.9)

sustained phonation for the automatic identification of preHD. They show great discriminative performance between preHD and controls.

Perez et al. [18] trained a Speech Recognition system on Huntington's speech and extracted a number of language features ranging from speech rate, and goodness of pronunciation to utterance length. The goal of this study differed from ours as they grouped together the preHD and HD patients in a single group to distinguish from control. They also pointed out to the difficulty to identify preHD.

We did not find any study attempting to predict the clinical scores in Huntington's disease from speech. Our strategy, which may allow following up individuals remotely, is likely to apply to other neurodegenerative diseases, such as in the Parkinson's Disease [19] or Alzheimer Disease [20].

3. Voice Database

Eighty five participants were included from two observational cohorts (NCT01412125 and NCT03119246) in this ancillary study at the Hospital Henri-Mondor Créteil, France): 61 people tested with a number of CAG repeats on the Huntingtin gene above 35 [1] (CAG $>$ 35), and 24 Healthy Controls (C) (See Table 1). All participants signed an informed consent. Mutant Huntingtin gene carriers were considered premanifest if both they score less than five at the Total Motor score (TMS) and their Total functional capacity (TFC) equals 13 [21] using the Unified Huntington Disease Rating Scale (UHDRS) [22].

All participants completed a standardised speech battery. The data were annotated with Seshat [23] and Praat [24] softwares. The annotators were second-year graduate students in speech pathology, all French native speakers.

Each participant was asked to take a deep breath and to sustain the vowel /a/ at a constant intensity and pitch level for as long as possible. The recordings were done in the same condition for all participants, with a ZOOM H4n Pro recorder, sampled at 44.1 kHz with a 16-bit resolution.

Table 2: *List of Phonatory features based on [4]. SD stands for Standard Deviation. †The vocal tremor features could not be computed on all samples.*

Dimension	Features
Airflow insufficiency	Maximum Phonation Time First Occurrence of Voice Break
Aperiodicity	Number of Voice Breaks Degree of Pitch Breaks Degree of Vocal Arrests
Irregular vibration of vocal folds	F0 SD Recurrence Period Density Entropy
Signal perturbation	Jitter (local) Shimmer (local)
Increased noise	Harmonics to Noise Ratio Detrended Fluctuation Analysis
Vocal tremor†	Frequency Tremor Intensity Index Amplitude Tremor Intensity Index
Articulatory deficiency	Mean of SD of MFCC Mean of SD of Delta MFCC

4. Features

4.1. Phonatory Features

We used the Phonatory features from [10, 4] to measure dimensions that can impede the correct sustained production of the vowel /a/ for HD and preHD gene carriers: airflow insufficiency, aperiodicity, irregular vibration of vocal folds, signal perturbation, increased noise, vocal tremor, articulatory deficiency. These features are summarised in the Table 2.

To automatically extract these features, we used the Praat software [24], the Parselmouth wrapper [25] and the tremor package [26]. The fundamental frequency (F0) and MFCCs were obtained with the Kaldi toolkit [27]. Besides, we implemented the Detrended Fluctuation Analysis and the Recurrence Period Density Entropy features from [13], in Python. To compute these two features and replicate [13], we down-sampled the audio to 22.5 kHz, otherwise the audio is down-sampled to 16 kHz. Due to instability for the extraction of the tremor features, there are missing data points for this dimension (See Table 3).

To overcome the limitations of heterogeneity of acoustic methodologies across studies and libraries, we provide an open-source version of the code (link) to reproduce our results and to extract each dimension of the sustained phonation of the vowel /a/.

4.2. Modulation Power Spectrum features

To complete the Phonatory features, we used the Modulation Power Spectrum (MPS) extracted for each vowel /a/ of each individual. Different perceptual attributes occupy distinct areas of the MPS: roughness, gender and size characteristics of the speaker, and linguistic meaning [12]. Besides, the MPS captures the spectral and amplitude modulations of the pitch and its harmonics [1]. The MPS representation is the amplitude spectrum of the 2D Fourier Transform of a time-frequency representation obtained from the sound waveform [11, 28]. This time-frequency representation is a log-scaled amplitude of a spectrogram computed every 1 millisecond (ms) with a spacing between each frequency bin of 50Hz using a Gaussian window. The linear spacing in the frequency axis [11] better describes sounds that have harmonic structure, like long steady vowels. The MPS is computed every 10 ms with a 100 ms window of the spectrogram, then the final representation is averaged over the full duration of the sound. Upward and downward temporal modulations are kept between -200 Hz and +200 Hz, and spectral modulations between 0 and 9.5 Cycles/kHz (99% of the energy was found between these intervals).

5. Methods

5.1. Group level: Statistical analysis

We followed the process for statistical comparison and correction from Rusz et al., 2014 [4]. Given the non-normality of the data, we tested the differences between groups with the non-parametric Kruskal-Wallis test. The Bonferroni correction is applied to correct for the seven types of speech deficits (see Table 2). For post-hoc analyses, we applied the Kolmogorov-Smirnov test to all features to check for normality within each group, and the Levene's test for homoscedasticity between groups. If normality and homoscedasticity requirements were fulfilled, we applied an independent t-test otherwise; we applied a non-parametric equivalent (Mann-Whitney U-test). The p-values were Bonferroni corrected for multiple comparisons per feature. We also estimated the effect size with the Cohen's d. Results are summarised in Table 3. Statistical analyses for the MPS features are displayed in the arxiv version of this paper.

[1] Readers can refer to [11, 12] for in-depth study of the different perceptual areas

Table 3: *Results of the statistical analyses between the three groups for the sustain of the vowel /a/: Controls (C), asymptomatic genetic carrier of Huntington's Disease (preHD), symptomatic genetic carrier of Huntington's Disease (HD). The p-values significativity of the tests (Kruskal-Wallis tests H-statistic (H stat) and post-hoc pairwise tests on the Cohen's d) are reported with *:* $0.01 < p \leq 0.05$, ***:* $0.001 < p \leq 0.01$, ****:* $p \leq 0.001$. *The post-hoc statistics are corrected for multiple comparison feature wise. SD stands for standard deviation.* † *23 HD, 6 preHD, 6 Controls data points could not be computed for the Frequency Tremor Intensity Index.* ‡ *15 HD, 6 preHD, 3 Controls data points could not be computed for the Amplitude Tremor Intensity Index.*

	Mean (SD)			H stat	Effect size Cohen's d		
	C	preHD	HD		HD/PreHD	HD/C	preHD/C
Phonatory Features							
Maximum Phonation Time (s)	17.0 (8.7)	15.7 (4.9)	9.1 (5.1)	23.6^{***}	-1.27^{***}	-1.19^{***}	-0.18
First Occurrence of Voice Break (s)	13.2 (7.6)	10.3 (5.7)	6.3 (5.1)	16.6^{***}	-0.77^{**}	-1.14^{***}	-0.41
Number of Voice Breaks	3.6 (9.2)	7.2 (12.3)	3.8 (7.4)	2.9	-0.37	0.02	0.34
Degree of Pitch Breaks (%)	1.1 (3.1)	4.8 (9.4)	6.3 (10.6)	5.7	0.15	0.59^{**}	0.59
Degree of Vocal Arrests (%)	1.6 (2.8)	3.8 (5.9)	7.3 (10.1)	6.8	0.39	0.68^{**}	0.49
F0 SD (Hz)	7.3 (10.8)	13.8 (15.2)	16.8 (15.3)	13.1^{**}	0.20	0.68^{***}	0.51
Recurrence Period Density Entropy	0.60 (0.18)	0.57 (0.16)	0.56 (0.16)	0.52	-0.09	-0.23	-0.13
Jitter (local) (%)	0.73 (0.47)	1.1 (1.7)	1.0 (1.0)	0.70	-0.10	0.32	0.36
Shimmer (local) (%)	8.5 (3.6)	7.9 (4.0)	8.4 (4.5)	0.37	0.14	0.02	0.15
Harmonics to Noise Ratio (dB)	16.1 (4.1)	15.8 (5.2)	15.3 (5.1)	0.52	-0.11	-0.19	-0.07
Detrended Fluctuation Analysis	0.70 (0.04)	0.70 (0.04)	0.68 (0.06)	2.5	-0.52	-0.41	0.16
Frequency Tremor Intensity Index †	4.0 (4.3)	7.1 (6.4)	7.4 (8.1)	3.5	0.07	0.52	0.60
Amplitude Tremor Intensity Index ‡	24.2 (7.6)	29.4 (14.7)	27.2 (14.4)	13.7^{**}	0.57	1.10	0.50
Mean of SD of MFCC	6.6 (1.2)	6.6 (0.79)	8.5 (1.7)	25.6^{***}	1.23^{***}	1.21^{***}	0.01
Mean of SD of Delta MFCC	1.3 (0.19)	1.2 (0.25)	1.5 (0.38)	14.7^{**}	0.80^{**}	0.77^{***}	-0.13

5.2. Individual level: Machine Learning

To assess the performance of each set of features for classification and regression, we conducted 100 repeated learning-testing with 20% of the data left out as test set [29]. We used scikit-learn for all our models and data processing [30].

5.2.1. Group classification

We first compared the predictive power of the different input features to discriminate between the three groups: Controls (C), preHD, and HD patients. To do so, we trained a logistic regression regularised with ElasticNet (e.g. L1 and L2 combined with $C = 1$ and $ratio = 0.5$). Then, we computed the mean and standard deviation of the Accuracy and the F1 score (See Table 4). We also reported the chance level by selecting randomly the class based on the prior distribution.

5.2.2. Regression of the clinical scores

Second, we assessed the predictive capabilities of the features to predict the composite score cUHDRS [16] (currently used in international clinical trials [31]), the TFC and the TMS. We trained a Linear regression model a with ElasticNet regulariser (e.g. L1 and L2 combined with $C = 1$ and $ratio = 0.5$). The phonation from Controls are not used in the regression. Then, we computed the mean and standard deviation of the Mean Absolute Error and the coefficient of Determination R^2 (See Table 4) as well as the chance level by predicting the mean of each outcome on the train set.

6. Results and discussions

Statistical Analyses in Table 3 showed that the most affected dimensions in HD are those related to airflow insufficiency and articulatory deficiency. We did not find any significant difference between PreHD and Controls, while both groups differed from HD. Post-hoc analyses revealed that HD differed from from Controls but not from preHD for pitch/F0 related features: the Degree of Pitch Breaks, the Degree of Vocal Arrests, and Standard deviation of the F0. This suggests an impairment of the vocal cords control prior to the clinical onset. None of the Phonatory features were sufficient to distinguish preHD from Controls. Tremor features displayed the strongest effect size when comparing preHD and Controls. But the unequal level of data loss in each group (51% HD, 38% preHD, and 25% Controls) suggests defining estimation methods that incorporate tremor instability and better tremor tracking methods [14] for future work.

The classification results (Table 4) for the Phonatory features yielded lower discrimination than the MPS features. Phonatory features performances are lower than the studies [4, 10]. This can be explained by the difference in the definition of the preHD group: the preHD population were based on a previous definition (the Diagnostic Confidence Level, item 17 of the UHDRS Motor Assessment). Some of their preHD genetic carrier showed some motor deficits (TMS than was up to 8 in their population). These differences may be due to the small sample size, inter-individual variability. The differences can also be attributed to the difficulty to tease apart the 3 groups in comparison to 2 groups comparison only. Besides, the MPS features are more suitable to identify the preHD. We also show the type of errors made by each set of features with the Confusion Matrices (see Figure 2). HD are the most identifiable group. Clearly, the preHD are often confused by the model as HD or Controls. Yet the types of errors differ. With the Phonatory features the preHD are even more classified as Controls than the Controls themselves (0.68 versus 0.45), which suggests a compensation mechanism.

The weights of the logistic regression trained to classify the sub-group based on the MPS features are interpretable. These weights can be visualised in the Figure 1. Even though the models had no prior how close the features are in the MPS space, we saw the emergence of patterns. The HD sub-classifier showed an area of activations for spectral modulation around 4 cycles/kHZ, which can be associated with temporal modula-

Figure 1: *Averaged weights of the Logistic Regression regularised with ElasticNet applied on the Modulation Power Spectrum Features to discriminate between each sub-group. Mean Sparsity* $= 37.1\%$

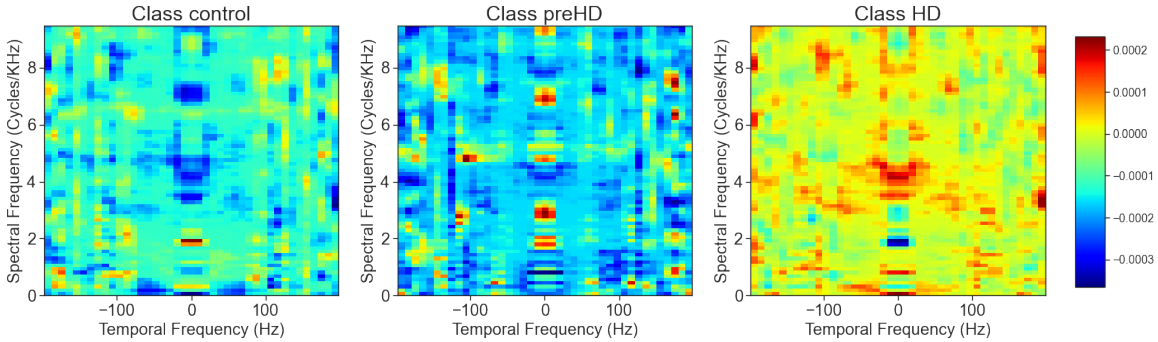
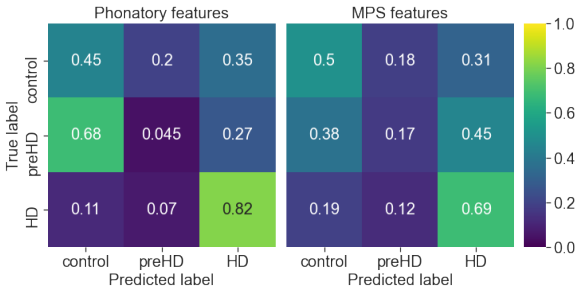


Table 4: *Results of the machine learning experiments based on the set of features obtained from the sustain of the vowel /a/. Mean and Standard Deviation are reported for each metrics. Classification performance is reported with Accuracy and F1-macro score. Regression performance is reported with the Mean Absolute Error (MAE) and the Coefficient of Determination (R^2). Best score for each metric is reported in* ***bold****. † The Phonatory features do not include the tremor features as they could not be computed for each subject.*

	Classification Results		Regression Results					
	C vs preHD vs HD		cUHDRS		TFC		TMS	
	Accuracy	F1-macro	MAE	R^2	MAE	R^2	MAE	R^2
Random Prior	0.38 (0.1)	0.31 (0.1)	4.21 (0.6)	0.0 (0.0)	1.96 (0.3)	0.0 (0.0)	19.07 (2.3)	0.0 (0.0)
Phonatory feat.†	**0.56 (0.1)**	0.40 (0.1)	**2.78 (0.5)**	**0.58 (0.2)**	**1.64 (0.3)**	**0.37 (0.2)**	**13.14 (1.7)**	**0.53 (0.2)**
MPS	0.54 (0.1)	0.43 (0.1)	3.68 (0.5)	0.28 (0.1)	1.83 (0.3)	0.15 (0.1)	17.87 (3.2)	0.26 (0.2)
MPS+Phonatory feat.†	**0.56 (0.1)**	**0.46 (0.1)**	3.10 (0.5)	0.49 (0.1)	1.70 (0.3)	0.25 (0.1)	15.34 (3.1)	0.29 (0.2)

Figure 2: *Confusion matrices for the Logistic regression based on the Phonatory features (left) and on the MPS features (right) averaged across all the repeated learning-testing experiments.*



tions between -45 and 45 Hz. We found the strongest activation at the origin (0.0, 0.0), which relates to voice breaks during the phonation. Even though the classifier is not perfect for preHD, we saw several specific activations along the Spectral modulation at 2.5 Cycles/kHZ and 7 Cycles/kHZ at Temporal Frequency equal to 0. This might suggest Frequency modulations specific to preHD, trying to avoid the zone around 4 Cycles/kHz of HD. The combination of the set of MPS and Phonatory features improved the classification performances up to an Accuracy of 0.56 and F1-macro of 0.46.

In contrast, Phonatory features better reflected the clinical scores cUHDRS, TFC, and TMS (Table 4) The composite measure cUHDRS, currently used in the assessment of international clinical trial [31], is the best predicted among the scores, if we rank them based on the coefficient of determination R^2. This means that the Phonatory features are a better indicator of the severity of the disease, once the clinical onset is declared.

7. Conclusions

Here, we combine the data from three groups for the study of the vocal markers of sustained phonation in Huntington's Disease patients: symptomatic, pre-symptomatic and control. We applied a statistical analysis, a classification study and assessed the capabilities to predict clinical scores. In addition, we introduced Modulation Power Spectrum features, to more traditional Phonatory features. Airflow insufficiency and articulatory deficiency measures distinguished HD patients from both preHD and Controls. However, Modulation Power Spectrum features provided more hope of distinguishing preHD from Controls. They allowed a three-fold reduction in misidentification of preHD. When replicated in a larger scale and in another population, this suggest that speech phonation might replace long traditional assessments, considering that the the sustained vowel task takes less than 1 minute and UHDRS takes a minimum of 30 minutes when ran by experts. It may allow repetitive testing with limited retest effect and recordings could also be blindly scored and analysed. This points to speech as a major future tool in the clinical panel of assessments.

8. Acknowledgements

We are very thankful to the patients that participated in our study. We thank Agnes Sliwinski, Katia Youssov, Laurent Cleret de Langavant for the multiple helpful discussions and the evaluations of the patients. This work is funded in part by the Agence Nationale pour la Recherche (ANR-17-EURE-0017 Frontcog, ANR-10-IDEX-0001-02 PSL*, ANR-19-P3IA-0001 PRAIRIE 3IA Institute) and Grants from Neuratris, from Facebook AI Research (Research Gift), Google (Faculty Research Award), Microsoft Research (Azure Credits and Grant), and Amazon Web Service (AWS Research Credits).

9. References

[1] J. F. Gusella, N. S. Wexler, P. M. Conneally, S. L. Naylor, M. A. Anderson, R. E. Tanzi, P. C. Watkins, K. Ottina, M. R. Wallace, A. Y. Sakaguchi *et al.*, "A polymorphic dna marker genetically linked to huntington's disease," *Nature*, vol. 306, no. 5940, pp. 234–238, 1983.

[2] M. J. Novak and S. J. Tabrizi, "Huntington's disease," *Bmj*, vol. 340, p. c3109, 2010.

[3] E. J. Wild and S. J. Tabrizi, "One decade ago, one decade ahead in huntington's disease," *Movement Disorders*, vol. 34, no. 10, pp. 1434–1439, 2019.

[4] J. Rusz, C. Saft, U. Schlegel, R. Hoffman, and S. Skodda, "Phonatory dysfunction as a preclinical symptom of huntington disease," *PloS one*, vol. 9, no. 11, 2014.

[5] A. P. Vogel, C. Shirbin, A. J. Churchyard, and J. C. Stout, "Speech acoustic markers of early stage and prodromal huntington's disease: a marker of disease onset?" *Neuropsychologia*, vol. 50, no. 14, pp. 3273–3278, 2012.

[6] W. Hinzen, J. Rosselló, C. Morey, E. Camara, C. Garcia-Gorro, R. Salvador, and R. de Diego-Balaguer, "A systematic linguistic profile of spontaneous narrative speech in pre-symptomatic and early stage huntington's disease," *Cortex*, vol. 100, pp. 71–83, 2018.

[7] M. Teichmann, E. Dupoux, S. Kouider, P. Brugières, M.-F. Boissé, S. Baudic, P. Cesaro, M. Peschanski, and A.-C. Bachoud-Lévi, "The role of the striatum in rule application: the model of huntington's disease at early stage," *Brain*, vol. 128, no. 5, pp. 1155–1167, 2005.

[8] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential diagnostic patterns of dysarthria," *Journal of speech and hearing research*, vol. 12, no. 2, pp. 246–269, 1969.

[9] S. K. Diehl, A. S. Mefferd, Y.-C. Lin, J. Sellers, K. E. McDonell, M. de Riesthal, and D. O. Claassen, "Motor speech patterns in huntington disease," *Neurology*, vol. 93, no. 22, pp. e2042–e2052, 2019.

[10] J. Rusz, J. Klempır, E. Baborová, T. Tykalová, V. Majerová *et al.*, "Objective acoustic quantification of phonatory dysfunction in huntington's disease." *PloS one*, vol. 8, no. 6, pp. e65 881–e65 881, 2013.

[11] T. M. Elliott and F. E. Theunissen, "The modulation transfer function for speech intelligibility," *PLoS computational biology*, vol. 5, no. 3, 2009.

[12] L. H. Arnal, A. Flinker, A. Kleinschmidt, A.-L. Giraud, and D. Poeppel, "Human screams occupy a privileged niche in the communication soundscape," *Current Biology*, vol. 25, no. 15, pp. 2051–2056, 2015.

[13] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. Costello, and I. M. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *Biomedical engineering online*, vol. 6, no. 1, p. 23, 2007.

[14] J. Peplinski, V. Berisha, J. Liss, S. Hahn, J. Shefner, S. Rutkove, K. Qi, and K. Shelton, "Objective assessment of vocal tremor," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6386–6390.

[15] V. Berisha, J. Liss, T. Huston, A. Wisler, Y. Jiao, and J. Eig, "Float like a butterfly sting like a bee: Changes in speech preceded parkinsonism diagnosis for muhammad ali," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2017, 2017, pp. 1809–1813.

[16] S. A. Schobel, G. Palermo, P. Auinger, J. D. Long, S. Ma, O. S. Khwaja, D. Trundell, M. Cudkowicz, S. Hersch, C. Sampaio *et al.*, "Motor, cognitive, and functional declines contribute to a single progressive factor in early hd," *Neurology*, vol. 89, no. 24, pp. 2495–2502, 2017.

[17] I. Shoulson, "Huntington disease: functional capacities in patients treated with neuroleptic and antidepressant drugs," *Neurology*, vol. 31, no. 10, pp. 1333–1333, 1981.

[18] M. Perez, W. Jin, D. Le, N. Carlozzi, P. Dayalu, A. Roberts, and E. M. Provost, "Classification of huntington disease using acoustic and lexical features," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2018, 2018, pp. 1898–1902.

[19] K. M. Smith, J. R. Williamson, and T. F. Quatieri, "Vocal markers of motor, cognitive, and depressive symptoms in parkinson's disease," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 71–78.

[20] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge," in *Proceedings of INTERSPEECH 2020*, Shanghai, China, 2020. [Online]. Available: https://arxiv.org/abs/2004.06833

[21] S. J. Tabrizi, D. R. Langbehn, B. R. Leavitt, R. A. Roos, A. Durr, D. Craufurd, C. Kennard, S. L. Hicks, N. C. Fox, R. I. Scahill *et al.*, "Biological and clinical manifestations of huntington's disease in the longitudinal track-hd study: cross-sectional analysis of baseline data," *The Lancet Neurology*, vol. 8, no. 9, pp. 791–801, 2009.

[22] K. Kieburtz, J. B. Penney, P. Corno, N. Ranen, I. Shoulson, A. Feigin, D. Abwender, J. T. Greenarnyre, D. Higgins, F. J. Marshall *et al.*, "Unified huntington's disease rating scale: reliability and consistency," *Neurology*, vol. 11, no. 2, pp. 136–142, 2001.

[23] H. Titeux*, R. Riad*, X.-N. Cao, N. Hamilakis, K. Madden, A. Cristia, A.-C. Bachoud-Lévi, and E. Dupoux, "Seshat: A tool for managing and verifying annotation campaigns of audio data," in *Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, May 2020, * Equal contribution.

[24] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, 2002.

[25] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[26] M. Brückl, "Vocal tremor measurement based on autocorrelation of contours," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[27] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 2494–2498.

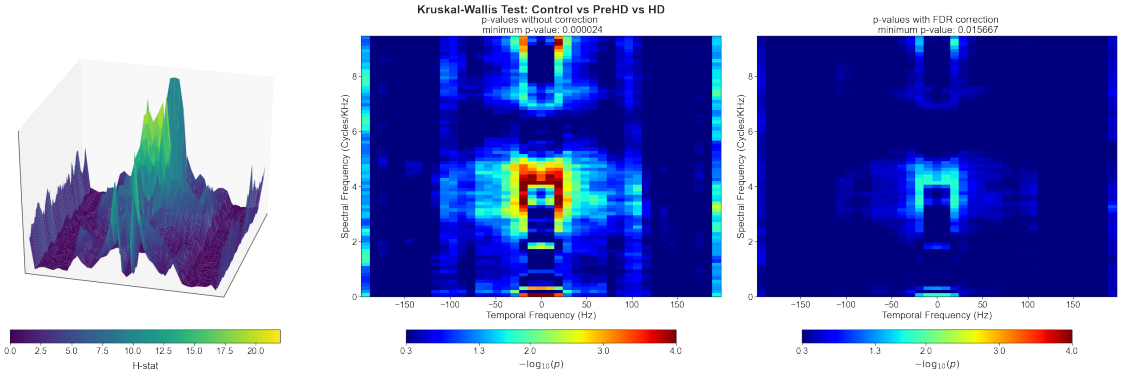
[28] J. E. Elie and F. E. Theunissen, "Zebra finches identify individuals using vocal signatures unique to each call type," *Nature communications*, vol. 9, no. 1, pp. 1–11, 2018.

[29] G. Varoquaux, P. R. Raamana, D. A. Engemann, A. Hoyos-Idrobo, Y. Schwartz, and B. Thirion, "Assessing and tuning brain decoders: cross-validation, caveats, and guidelines," *NeuroImage*, vol. 145, pp. 166–179, 2017.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[31] S. J. Tabrizi, B. R. Leavitt, G. B. Landwehrmeyer, E. J. Wild, C. Saft, R. A. Barker, N. F. Blair, D. Craufurd, J. Priller, H. Rickards *et al.*, "Targeting huntingtin expression in patients with huntington's disease," *New England Journal of Medicine*, vol. 380, no. 24, pp. 2307–2316, 2019.

Figure 3: *Results of the statistical analyses for the Modulation Power Spectrum Features between the three groups for the sustain of the vowel /a/: Controls (C), asymptomatic genetic carrier of Huntington's Disease (preHD), symptomatic genetic carrier of Huntington's Disease (HD). Left figure is the distribution of the H-statistic. Middle figure is the distribution of the uncorrected p-values. Right figure is the distribution of the FDR corrected p-values.*



10. Appendix

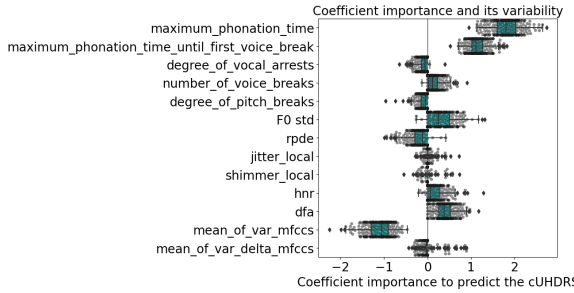
10.1. Additional Statistical analysis for the Modulation Power Spectrum Features

We tested the differences between groups with the non-parametric Kruskal-Wallis test. The False-Discovery-Rate correction is applied to correct for the multiple comparisons ($77 \times 41 = 3157$). We reported the statistical results in the Figure 3. 1.7% (14.6%) of the MPS features are found significant with (without) the FDR correction. We found significant area to separate the three groups of activation, especially around 4 cycles/kHz (very similar to the area found to identify the symptomatic HD patients).

10.2. Interpretation of coefficients of linear models for the regression of clinical scores based on the Phonatory features

As we used linear models, each target value is modelled as a linear combination of the input features. We followed the methodology analysis from the example of scikit-learn [30] (link). The stability of the predictors is shown through the different coefficients across folds. As we used the ElasticNet Regulariser (L1+L2 regularisation of the coefficients), we also observed the selection of variables based on the Coefficient Importance analysis. The results for the regression results for the clinical measures cUHDRS, TFC, TMS are reported respectively in the Figure 4, Figure 5, Figure 6.

Figure 4: *Coefficient Importance of the different Phonatory Features across the different cross-validation folds to predict the cUHDRS*



The Maximum Phonation Time and First Occurrence of Voice Break are the features the most used for all 3 regression tasks. The Mean of SD of MFCC also contribute to the prediction of the cUHDRS and TMS. The Detrended Fluctuation Analysis is also a feature useful for the prediction of the TMS (coefficient never set to 0).

Otherwise, the contributions of the other features for the other tasks are more blurred or often set to 0. These coefficient importance give also the direction associated with the progress of the scores and then the disease.

Figure 5: *Coefficient Importance of the different Phonatory Features across the different cross-validation folds to predict the TFC*

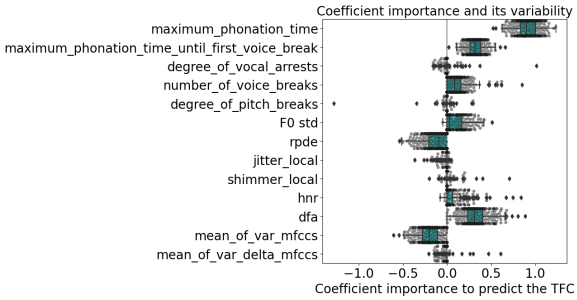
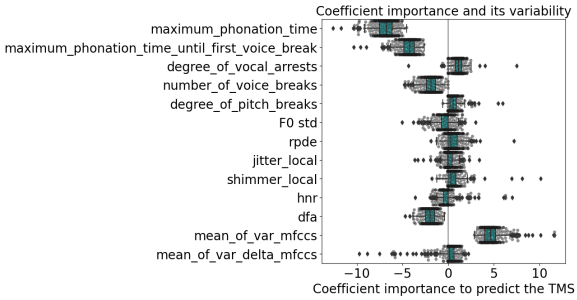


Figure 6: *Coefficient Importance of the different Phonatory Features across the different cross-validation folds to predict the TMS*



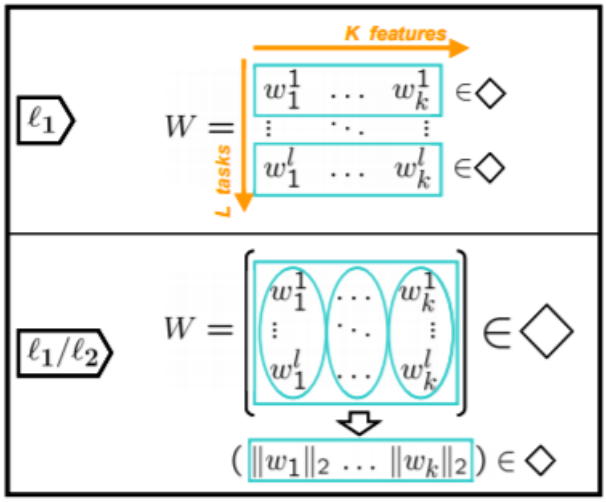


Fig. 4.1.: The ℓ_1/ℓ_2 vs the ℓ_1 regularization schemes. The illustration figure is extracted from (Obozinski et al., 2006).

In addition to this study, in collaboration with the Master student Reda Arab, we ran a number of additional analyses on the same data of the Interspeech 2020 paper (Riad, Titeux, et al., 2020). We evaluated what was the best modelling strategy for the prediction of clinical scores.

In the paper, we considered a single task prediction setup, where we built models independently for each single score alone. However, the task to predict each clinical score are not completely independent, as their goals is to reflect the cognitive, motor or global measures of HD. As motor and cognitive symptoms both contribute to the global decline in HD, we wanted to take advantage of this information in the construction of the models. Especially, as the number of data points is limited, correlations among tasks can be helpful to construct better models (Obozinski et al., 2006).

We can exploit the correlation among tasks to select and weights of the input features, so that they use the same set (Obozinski et al., 2006). This approach of selection of features has been shown to be especially helpful and low-data regime. In the Interspeech 2020 paper, we selected the variables based on the lasso regularization and regularized the features using the ridge regularization. The combination of these 2 regularizations is the ElasticNet regularization.

For clarity, we describe how the multi-task model works in the setting of Lasso Regularization. See Figure 4.1 for illustration comparing single and multi-task regularization.

Solving the l-th task with linear models is performed through the learning a set of weights $w^l \in \mathbb{R}^K$ based on input training data $\left\{\left(x_i^l, y_i^l\right) \in \mathbb{R}^K \times \mathbb{R}, i = 1 \cdots N_l, l = 1 \cdots L\right\}$ thanks to the Lasso regularization ℓ_1:

$$\min_{w^l} \frac{1}{N_l} \sum_{i=1}^{N_L} J^l \left(w^l, x_i^l, y_i^l\right) + \lambda \|w^l\|_1 \tag{4.1}$$

Each row of w^l represents the weights to predict the l-th task.

This is equivalent to solve the following global problem, by taking into account all L tasks:

$$\min_{W} \sum_{l=1}^{L} \frac{1}{N_l} \sum_{i=1}^{N_l} J^l \left(w^l, x_i^l, y_i^l\right) + \lambda \sum_{l=1}^{L} \|w^l\|_1 \tag{4.2}$$

The solution to this global problem yields sparse independent coefficient vectors w^l to predict the l-th task.

Multi-task Lasso is using the penalty ℓ_1/ℓ_2, and will enforce sparsity along the other dimension, along the features.

$$\min_{W} \sum_{l=1}^{L} \frac{1}{N_l} \sum_{i=1}^{N_l} J^l \left(w^l, x_i^l, y_i^l\right) + \lambda \sum_{k=1}^{K} \|w_k\|_2 \tag{4.3}$$

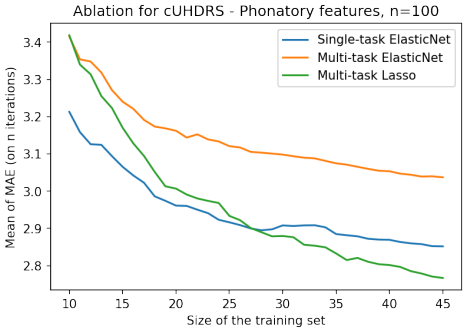
This strategy helps to selects the features based on all the tasks and share the same support set.

The Multi-task ElasticNet is an extension of the Multi-task Lasso, by adding the Froebenius norm to regularize the overall magnitude of the weights.

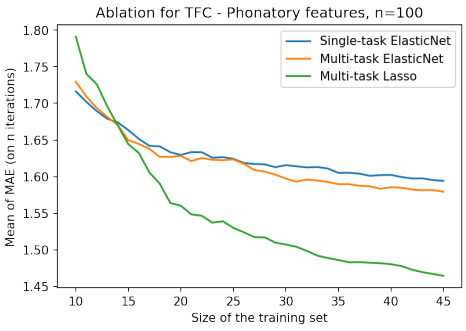
$$\min_{W} \sum_{l=1}^{L} \frac{1}{N_l} \sum_{i=1}^{N_l} J^l \left(w^l, x_i^l, y_i^l\right) + \lambda \sum_{k=1}^{K} \|w_k\|_2 + \gamma \|W\|_F \tag{4.4}$$

We compared the two multi-task modelling approaches to the single task modelling from the original paper. We ran an ablation study on the phonatory features to select for the best ones for each regression of clinical score under study (See Figure 4.2).

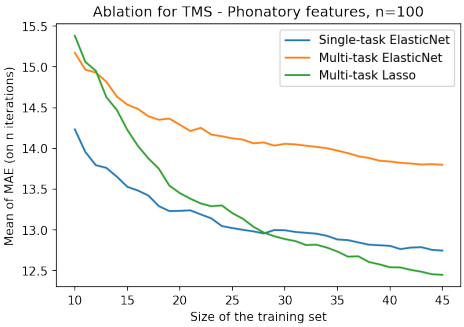
First, we can observe that in the very low data regime all models perform poorly close to the chance for the 3 clinical markers. In all cases, the selection of variables thanks to the multi-task Lasso improves the predictive performance for each clinical



(a) Prediction scores of the cUHDRS.



(b) Prediction scores of the TFC.



(c) Prediction scores of the TMS.

Fig. 4.2.: Ablation study on the phonatory features to extract the clinical scores from sustained phonations. Comparison of single and multi-task models. The X-axis of each graph is the size of training set. The experiments for each size of the training set are repeated 50 times as in the previous study. The reported metric is the Mean Absolute Error (MAE) averaged over the 50 repetitions.

marker as the size of the training set increases. This method is especially useful to predict the TFC score in comparison to single score prediction.

4.1.2 Imprecise vowel articulation in Huntington's disease

The distortion of vowels is often reported in HD, and represent one of the features that characterize the hyperkinetic dysarthria (Darley et al., 1969; Diehl et al., 2019; Rusz, Tykalova, et al., 2021). Yet, speech motor impairments in Huntington's Disease are highly heterogeneous. Each individual might cope individually with the disease's progression and there is potential different profiles (Diehl et al., 2019). To study the distortion of vowels, we considered voice features collected from (1) simple recordings of the French vowel /a/, /i/ and /u/ uttered in a sustained manner for as long as possible, and from (2) the vowels /a/, /o/, /e/, /i/, /u/ extracted from spontaneous speech during the open-vocabulary tasks not involving emotional content. To summarize vowel distortion along a single dimension is complicated, even though researchers tends to use the Vocalic Space Area (VSA) as the most classic way to measure it. It has been used several times in HD, as underlined in the review of speech features in HD (J. C. Chan et al., 2019).

The vowel distortion is usually measured, by first computed the first two formants based on vowels (F1, F2), and measuring statistics of these values and all observation of vowels, while still taking into account their categories. We used measures from (Rusz, Tykalova, et al., 2021) and (Audibert & Fougeron, 2012) to measure the distortion of the vowel space. All the measures are summarized in Table 4.1.

In addition, as mentioned in (Vogel et al., 2012; Rusz, Cmejla, et al., 2013), there is a growing body of evidence that dysarthria in neurodegenerative diseases vary as the function of the task, especially as function of the cognitive load. That is why, we examined which of fixed vocabulary and open vocabulary speaking tasks are more sensitive to vowel distortion in HD. We also analyzed the performance of pre-symptomatic individuals, if these vowel imprecision start even before the official clinical onset of the disease. Here, these questions are addressed at the group level, using statistical methodologies like in the sustained phonation paper (Riad, Titeux, et al., 2020).

Our results concerning vowel distortion features are *mixed* and do not reproduce all the results from (Rusz, Tykalova, et al., 2021). Based on sustained phonations, we did not observe any effects that allows to conclude that the 3 group differ (See Table 4.2). In the vowels from spontaneous speech, we observed for the first time,

Tab. 4.1.: List of Distortion features based on (Audibert & Fougeron, 2012), (Rusz, Cmejla, et al., 2013) and (Huet & Harmegnies, 2000). †The distortion features requiring all vowels could not be computed on sustain phonation tasks.

.

Description	Formula
Triangular vowel space area (tVSA)	$\frac{1}{2}\sum_{i\in\{/a/,/i/,/u/\}}(F1_iF2_{i+1} - F1_{i+1}F2_i)$
Pentagon vowel space area (pVSA)†	$\frac{1}{2}\sum_{i\in\{/a/,/e/,/i/,/u/,/o/\}}(F1_iF2_{i+1} - F1_{i+1}F2_i)$
F1 range ratio (F1RR)	$2\frac{F1_a}{F1_i+F1_u}$
F2 range ratio (F2RR)	$\frac{F2_i}{F2_u}$
Vowel Articulation Index (VAI)	$\frac{F1_a+F2_i}{F1_i+F1_u+F2_a+F2_u}$
Mean square between vocalic clouds in the vocalic space. CMinter †	See (Huet & Harmegnies, 2000)
Mean square within vocalic clouds in the vocalic space. CMintra †	See (Huet & Harmegnies, 2000)
Index of system organization (Φ)†	$\frac{CMinter}{CMintra}$

significant difference between preHD and HC for the VAI measure from (Roy et al., 2009). The VAI is a measure that have been developed to minimize the inter-speaker variability and sensitivity to dysarthria perceptual severity. We also observe that two features separate HD (CMinter and the Φ) from other populations. This underlines the potential need for cognitive load to observe effect in HD speech (See next Section).

4.1.3 Predicting clinical scores in Huntington's Disease: a lightweight speech test

Previously, we examined only simple vowel features, extracted either from sustained phonation or vowels from spontaneous speech. However, it was observed in previous studies that timings of the speech units, and prosody might be affected in HD. Here, we proposed to use the combination of two tasks that modulates cognitive load and that remain short and simple tasks to perform and annotate. Yet, we kept in mind the need to measure different dimensions which can not all be extracted from simple sustained phonations: (1) articulatory and phonatory deficiencies, (2) Rhythm and temporal statistics, (3) sequence errors and perseverations and (4) Collateral track additions.

Tab. 4.2.: Results of the statistical analyses for vowel distortion measure between the three groups for the sustain of the vowel /a/, vowel /i/ and vowel /u/: Healthy Controls (HC), asymptomatic genetic carrier of Huntington's Disease (preHD), symptomatic genetic carrier of Huntington's Disease (HD). The p-values significativity of the tests (Kruskal-Wallis tests H-statistic (H stat) and post-hoc pairwise tests on the Cohen's d) are reported with *: $0.01 < p \leq 0.05$, **: $0.001 < p \leq 0.01$, ***: $p \leq 0.001$. The post-hoc statistics are corrected for multiple comparison feature wise. SD stands for standard deviation.

	Mean (SD)			H-stat	Effect size Cohen's d		
	HC (N=18)	preHD (N=20)	HD (N=45)		HD/PreHD	HD/HC	preHD/HC
Feature							
tVSA	7.4 (4.9)	4.9 (4.1)	4.6 (3.4)	4.7	-0.08	-0.72	-0.55
F1RR	1.8 (0.2)	1.8 (0.3)	1.8 (0.4)	3.3	-0.06	-0.15	-0.12
F2RR	1.6 (0.4)	1.5 (0.5)	1.4 (0.4)	3.1	-0.12	-0.43	-0.28
VAI	0.78 (0.1)	0.74 (0.1)	0.72 (0.1)	3.9	-0.11	-0.64	-0.46

Tab. 4.3.: Results of the statistical analyses for vowel distortion measure between the three groups for the 30ms - center of vowels /a/,/e/,/i/,/o/,/u/ extracted from spontaneous speech: Healthy Controls (HC), asymptomatic genetic carrier of Huntington's Disease (preHD), symptomatic genetic carrier of Huntington's Disease (HD). The p-values significativity of the tests (Kruskal-Wallis tests H-statistic (H stat) and post-hoc pairwise tests on the Cohen's d) are reported with *: $0.01 < p \leq 0.05$, **: $0.001 < p \leq 0.01$, ***: $p \leq 0.001$. The post-hoc statistics are corrected for multiple comparison feature wise. SD stands for standard deviation. †The distortion features requiring all vowels could not be computed on sustain phonation tasks.

	Mean (SD)			H-stat	Effect size Cohen's d		
	HC (N=22)	preHD (N=23)	HD (N=59)		HD/PreHD	HD/HC	preHD/HC
Feature							
tVSA	2.5 (1.5)	1.6 (1.19)	2.3 (1.5)	4.3	0.43	-0.14	-0.60
pVSA†	2.7 (1.8)	2.0 (1.6)	2.6 (1.6)	3.3	0.38	-0.08	-0.45
F1RR	1.34 (0.12)	1.27 (0.15)	1.31 (0.17)	2.2	0.26	-0.21	-0.55
F2RR	1.37 (0.14)	1.28 (0.10)	1.32 (0.14)	4.0	0.24	-0.36	-0.67
VAI	0.65 (0.03)	0.62 (0.04)	0.63 (0.04)	5.8	0.15	-0.5	-0.76*
CMintra†	3.1 (1.1)	3.7 (1.0)	3.6 (1.9)	3.2	-0.04	0.28	0.52
CMinter†	152.1 (54.3)	161.8 (96.6)	118.1(80.3)	11.1*	-0.51*	-0.45*	0.12
Φ †	54.0 (27.1)	49.2(34.7)	38.7 (30.1)	8.3	-0.33	-0.51*	-0.15

The dimensions (1)-(2)-(4) have been propely quantitavitly studied in the past, but the perseveration have not been properly investigated, even though perseverations are frequently observed in HD (Oosterloo et al., 2019). They are inappropriate repetition of a given behaviour, while a new type of behavior is expected. There are different level of complexity for these repeating behaviors (Cohen & Dehaene, 1998): (1) the highest-level, 'stuck-in-set' perseverations, an inability to switch of task or adopt new strategy (this can be measured with the Wisconsin Card Sorting Test) (2) the lowest level, 'continuous' perseveration, compulsive repetition of the same elementary motor schema (3) intermediate level of complexity, with 'recurrent', repetition of a given unit, such as the repetition of a word.

That is why, in the counting study, we developed a number of features to account for the sequence error and perseverations features. Here, we describe briefly how we measured these aspects. As the sequence to be pronounced by each participant, we proposed to compare quantitatively what has been actually pronounced. We used edit-distance and gestalt-similarity features. This comparison is performed both at the *word* and *phone* levels of the sequence to measure lower and higher levels of perseveration behaviors in the counting tasks.

For, the edit-distance we used the Damerau–Levenshtein distance. To compute this edit-distance, we need to define the function $d_{a,b}(i,j)$ that computes the distance between the prefix of the two strings *a* and *b* indexed by an *i* index and a *j* index. We can compute recursively the value of this function:

$$d_{a,b}(i,j) = \min \begin{cases} 0 & \text{if } i = j = 0 \\ d_{a,b}(i-1,j) + 1 & \text{if } i > 0 \\ d_{a,b}(i,j-1) + 1 & \text{if } j > 0 \\ d_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} & \text{if } i,j > 0 \\ d_{a,b}(i-2,j-2) + 1 & \text{if } i,j > 1 \text{ and } a_i = b_{j-1} \text{ and } a_{i-1} = b_j \end{cases} \tag{4.5}$$

Finally, the value of the edit-distance between a and b is $d_{a,b}(|a|,|b|)$ where $|a|$ is the length of a.

As the edit-distance is only based on *local* errors in the sequence, we also computed the gestalt-similarity, that takes into account more general patterns in the sequence. The gestalt similarity is computed as:

$$S_{\text{Gestalt}}(a,b) = \frac{2K_m}{|a| + |b|} \tag{4.6}$$

K_m represents the biggest number of matching characters. They are those in the longest common sub-string plus, recursively, matching characters in the unmatched region on either side of the longest common sub-string.

Based on all the features, we measured the capabilities to predict at the individual level motor, cognitive, functional and global scores in HD. We extended our analyses to correlation with striatal measures and we pre-registered all our methods before running any analyses on the BasalVoice database.

The results of this study are acceptable subject to some minor revisions for publication in the Journal of Neurology (Riad, Titeux, et al., 2020) and the research publication is included below. We include in the Appendix A.2 the pre-registration methodology built with aspredicted.org.

Journal of Neurology
https://doi.org/10.1007/s00415-022-11148-1

ORIGINAL COMMUNICATION

Predicting clinical scores in Huntington's disease: a lightweight speech test

Rachid Riad[1,2,3,4,5,6,7] · Marine Lunven[1,2,3,4] · Hadrien Titeux[5,6,7] · Xuan-Nga Cao[5,6,7] · Jennifer Hamet Bagnou[1,2,3,4] · Laurie Lemoine[1,2,3,4] · Justine Montillot[1,2,3,4] · Agnes Sliwinski[1,2,3,4] · Katia Youssov[3,4] · Laurent Cleret de Langavant[1,2,3,4] · Emmanuel Dupoux[5,6,7] · Anne-Catherine Bachoud-Lévi[1,2,3,4]

Received: 12 October 2021 / Revised: 19 January 2022 / Accepted: 18 April 2022

Abstract

Objectives Using brief samples of speech recordings, we aimed at predicting, through machine learning, the clinical performance in Huntington's Disease (HD), an inherited Neurodegenerative disease (NDD).

Methods We collected and analyzed 126 samples of audio recordings of both forward and backward counting from 103 Huntington's disease gene carriers [87 manifest and 16 premanifest; mean age 50.6 (SD 11.2), range (27–88) years] from three multicenter prospective studies in France and Belgium (MIG-HD (ClinicalTrials.gov NCT00190450); BIO-HD (ClinicalTrials.gov NCT00190450) and Repair-HD (ClinicalTrials.gov NCT00190450). We pre-registered all of our methods before running any analyses, in order to avoid inflated results. We automatically extracted 60 speech features from blindly annotated samples. We used machine learning models to combine multiple speech features in order to make predictions at individual levels of the clinical markers. We trained machine learning models on 86% of the samples, the remaining 14% constituted the independent test set. We combined speech features with demographics variables (age, sex, CAG repeats, and burden score) to predict cognitive, motor, and functional scores of the Unified Huntington's disease rating scale. We provided correlation between speech variables and striatal volumes.

Results Speech features combined with demographics allowed the prediction of the individual cognitive, motor, and functional scores with a relative error from 12.7 to 20.0% which is better than predictions using demographics and genetic information. Both mean and standard deviation of pause durations during backward recitation and clinical scores correlated with striatal atrophy (Spearman 0.6 and 0.5–0.6, respectively).

Interpretation Brief and examiner-free speech recording and analysis may become in the future an efficient method for remote evaluation of the individual condition in HD and likely in other NDD.

Keywords Huntington's disease · Speech · Machine learning

✉ Anne-Catherine Bachoud-Lévi
anne-catherine.bachoud-levi@aphp.fr

[1] Département d'Études Cognitives, École Normale Supérieure, PSL University, 75005 Paris, France

[2] Faculté de Médecine, Université Paris-Est Créteil, 94000 Créteil, France

[3] Inserm U955, Institut Mondor de Recherche Biomédicale, Équipe E01 NeuroPsychologie Interventionnelle, 94000 Créteil, France

[4] Centre de Référence Maladie de Huntington, Service de Neurologie, AP-HP, Hôpital Henri Mondor-Albert Chenevier, 51 avenue du Maréchal de Lattre de Tassigny, 94000 Créteil, France

[5] Laboratoire de Sciences Cognitives et Psycholinguistique, CNRS 8554, PSL University, 29 rue d'Ulm, 75005 Paris, France

[6] INRIA, Cognitive Machine Learning Team, 2 Rue Simone IFF, 75012 Paris, France

[7] EHESS, 54 boulevard Raspail, 75006 Paris, France

Published online: 14 May 2022

Introduction

Huntington's disease (HD) is a rare severe inherited neurodegenerative disease (NDD) whose natural history is well known and well characterized. It combines all complexity of NDDs by associating motor, psychiatric, and cognitive disorders resulting in functional impairment [1]. Despite the development of innovative and promising clinical therapies, a major challenge is the identification of markers sensitive to disease progression, even in the premanifest stage (preHD), before the appearance of motor symptoms.

Current clinical assessments are carried out with the Unified Huntington's Disease Rating Scale (UHDRS) [2], the worldwide reference scale for HD studies. This is done once or twice a year, during face-to-face examinations performed by trained experts from different specialties (neurologists, neuropsychologists, psychiatrists, and nurses). Each clinical domain is evaluated separately using lengthy, and often subjective scales [3, 4]. Recently, a multi-domain score, named cUHDRS, was proposed as a single endpoint of clinical trials in HD thanks to its greater sensitivity to disease progression [5]. As it combines various scales of the UHDRS, it still requires trained experts and multiple scale assessments. Cognitive batteries with time-dependent tasks [6], brain imaging with striatal volumes [7] or biofluids with Human Cerebrospinal Fluid (CSF) Neurofilament level [8] have also been evaluated as potential markers. These three types of markers have been considered as candidate biomarkers to follow the evolution of HD. However, they all require the presence of the patient at the hospital and a high level of expertise or equipment. In particular (1) cognitive batteries are carried out face-to-face by an expert neurologist/neurologist; (2) high quality brain imaging requires visits of the patient to the neuroimaging center with expensive equipment; (3) analysis of biofluids such as CSF imposes an invasive procedure, which additionally cannot be performed outside hospital under clinical surveillance.

This calls for objective, cost-effective tests to measure the symptoms in a unified approach [9–11]. Neurodegenerative disorders are complex and heterogeneous at the individual level. It is very unlikely that a single marker/measure would have all the good properties for diagnostic and severity assessments of different types of symptoms and truly help for real life clinical decisions. Yet, the combination of complementary biomarkers appears to be a more promising path to predict accurately the different clinical symptoms. Traditional methodologies used in Neurology, Inferential or Bayesian statistics, cannot handle and properly digest very high dimensional data, especially when the number of markers is on par or outnumber the number of data points in the cohort. Making accurate predictions at the individual level becomes possible with machine learning methods. These methods are designed to detect subtle patterns, taking into account a large number of variables, potentially with non-linear interactions [12, 13]. Thanks to increasing computing power, machine learning models now provide an effective methodology to analyze the high-dimensional output of sensors, such as microphones or smartwatches, yielding a patient-tailored approach. This could lead to improved efficiency of the screenings and evaluations of disease modifying therapies by capturing the different clinical dimensions of HD [11].

In this context, speech and language offer an appealing alternative unlocking potential remote evaluation and offering a relevant multi-domain approach. Speaking invokes complex motor abilities [14], cognitive control, and planning at multiple linguistic levels [15]. HD participants are impaired during different steps of spoken language production: phonetics and prosody [16–22], syntax and morphology [23], semantic [24, 25] as well as timings and pauses [26–28], making spoken language a good candidate for clinics. Significant differences were found between healthy controls and HD groups for acoustic markers [16, 27] and language markers [26]. Among these markers, it was found that the speech rate correlates with disease burden score, probability of disease onset, the estimated years to onset, and cognitive score [19, 27]. In addition, speech analysis combined with machine learning models allowed the discrimination of manifest HD and PreHD individuals from controls [29, 30]. However, some of these speech tasks suffer some drawbacks, such as the requirement of fastidious annotation by linguistic experts or language adaptation difficulties, which make their use not suitable for clinical practice; and their sensitivity to the various HD symptoms remain unknown [31].

To fill this gap, we test the capacity of speech to predict the main clinical variables of the UHDRS (cUHDRS, motor, functional, and cognitive) in carriers of the mutant Htt gene. Participants performed a quick speech test consisting of counting forward and backward numbers. We developed a method to quantify articulation, rhythm, perseveration, and vocalization additions. Machine learning models were trained and assessed on different sets of participants to ensure generalization of our results. Finally, the clinical value of speech features was further substantiated by their correlations with the striatal atrophy, the anatomical hallmark of HD [1].

Methods

We pre-registered all the methods before running the analyses to ensure its reliability and avoid inflated results (https://aspredicted.org/blind.php?x=/66K_66C). We developed the

methods with a first cohort (the Multicentric intracerebral grafting cohort, MIG-HD, NCT00190450) and then pre-registered. This first cohort is only used for training models, but the validation was only performed with independent cohorts (see Fig. 1).

Participants

French native speakers ($N = 103$) individuals with at least 36 CAG repeats on the mutant Htt gene of HD were included in this study (Table 1). One visit refers to one visit to the hospital for a given participant. All assessments were performed on the same visit. Participants were enrolled from three prospective studies: 36 manifest HD from MIG-HD prior to any intervention in 6 centers in France and Belgium from Stage I to Stage III, as defined by the Total functional capacity (TFC)[32], and 67 (51 manifest and 16 PreHD) from both the BIOHD (NCT01412125) and Repair-HD (NCT03119246) cohorts. PreHD participants were defined by a TFC score at 13 and a total motor score (TMS) of the UHDRS equal

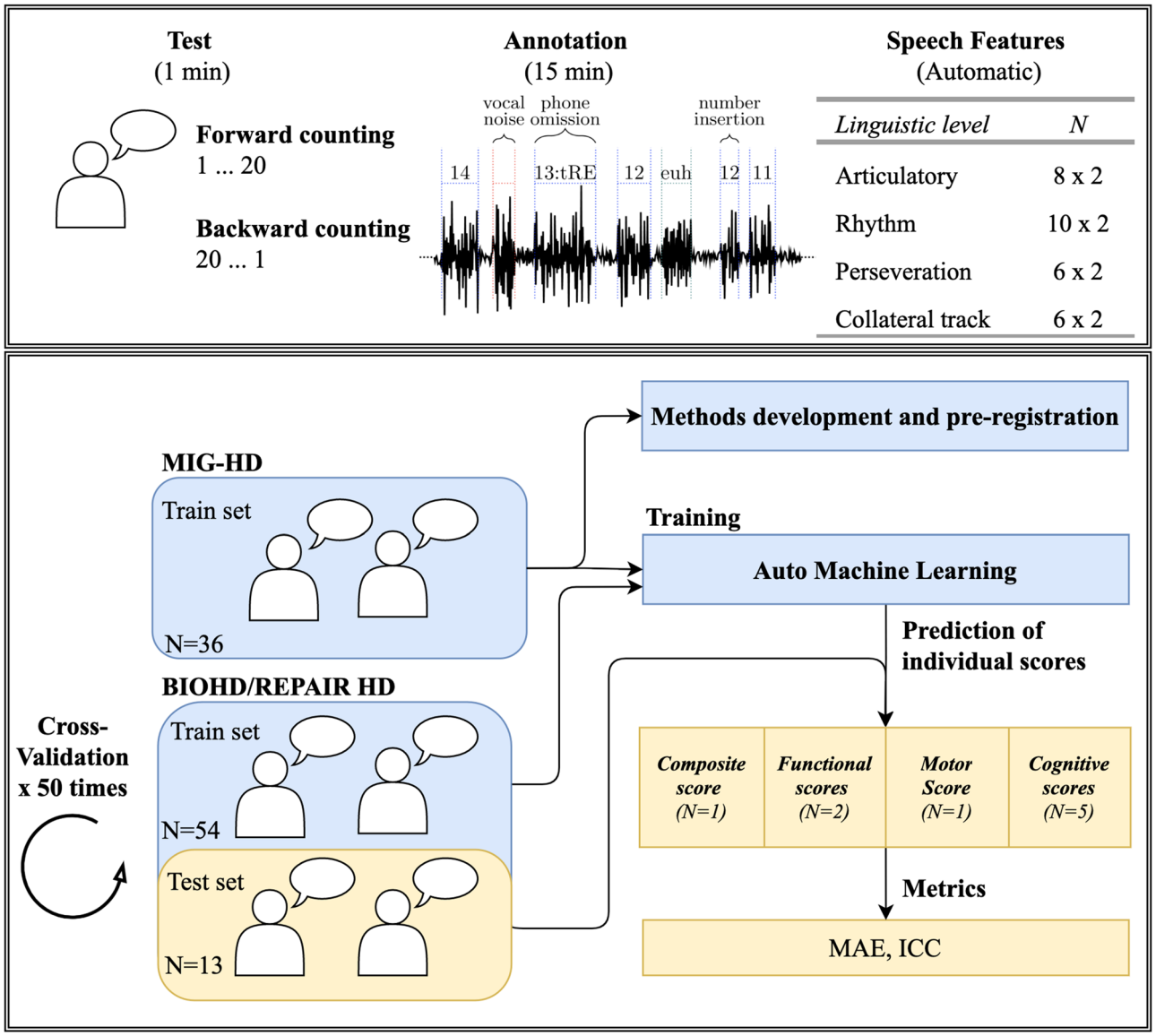


Fig. 1 Extraction of individual clinical scores from the speech samples. (Top panel) Examples of portions of the speech signal and various types of vocalizations and segmentation are provided. Similar speech features were extracted separately from the forward and backward counting tasks yielding to 60 features (30×2). (Bottom panel) Illustration of the methods developments, Machine learning training and evaluation of the predictions of the clinical scores. *N CAG* number of CAG repeats on the Huntingtin gene, *DBS* Disease Burden Score. *TFC* Total Functional capacity, *TMS* Total motor score, *SDMT* Symbol digit modality, *UHDRS IS* UHDRS Independence Scale, *MAE* Mean absolute error, *ICC* Intraclass correlation coefficient, *cUHDRS* composite UHDRS

Table 1 Demographics and clinical performance of the participants in the cohorts under study at baseline

	MIG-HD	BIOHD/REPAIRHD	Total
Number of participants	36	67	103
Premanifest/manifest	0/36	16/51	16/87
Number of visits per patient	1.4 (0.5) [1–2]	1.1 (0.3) [1–2]	1.2 (0.4) [1–2]
Gender	23F/13 M	40F/27 M	63F/40 M
Age at first visit	47.0 (9.1) [28–68]	52.7 (11.8) [27–88]	50.7 (11.2) [27–88]
Laterality	30R/5L/1A	59R/8L/0A	89R/13L/1A
Number of CAG repeats	45.3 (4.4) [37–60]	43.5 (3.1) [39–55]	44.0 (3.6) [37–60]
cUHDRS mean (SD) [range]	9.1 (2.5) [5.2–15.0]	11.1 (4.6) [2.5–18.8]	10.4 (4.0) [2.5–18.8]
Total motor score mean (SD) [range]	35.0 (13.6) [7–63]	26.7 (20.3) [0–60]	29.6 (18.6) [0–63]
TFC mean (SD) [range]	10.4 (1.7) [6–13]	11.0 (2.2) [5–13]	10.8 (2.0) [5–13]
UHDRS independence scale mean (SD) [range]	85.7 (8.5) [70–100]	88.9 (12.9) [60–100]	87.8 (11.8) [60–100]
Verbal fluency 1 min mean (SD) [range]	28.2 (8.5) [9–45]	27.6 (13.3) [9–62]	27.8 (11.8) [9–62]
Symbol digit modality test mean (SD) [range]	24.8 (7.6) [11–42]	31.9 (15.2) [3–67]	29.4 (13.4) [3–67]
Stroop word mean (SD) [range]	61.9 (15.0) [39–99]	70.7 (24.7) [23–117]	67.6 (22.1) [23–117]
Stroop color mean (SD) [range]	46.6 (11.9) [24–76]	52.3 (18.5) [16–89]	50.3 (16.7) [16–89]
Stroop interference mean (SD) [range]	26.7 (8.8) [11–45]	29.9 (12.8) [7–58]	28.8 (11.6) [7–58]

Mean, (Standard Deviations) [range]

F Female, *M* Male, *R* Right, *L* Left, *A* Ambidexter, *TFC* Total Functional Capacity

or below five [33]. The Disease Burden Score (DBS) was computed using the formulae: age × (CAG length − 35.5) [33]. All participants signed an informed consent. Ethical approval was given by the institutional review board from Henri Mondor Hospital (Créteil, France) for the French part of MIG-HD, Bio HD and Repair-HD, and the institutional review board from Erasme Hospital in Belgium. It complied with the Helsinki Declaration, current Good Clinical Practice guidelines, and local laws and regulations.

Clinical evaluation

Participants were assessed by certified examiners through nine measures classically used for both clinical practice and trial (Fig. 1): the UHDRS Total Motor Score (TMS), five cognitive assessments (the Symbol Digit Modalities Test (SDMT), the Verbal Fluency test 1-min (VF), and the three components of the Stroop test (word (SW); color (SC); interference (SI)), and two functional scales (the Total Functional Capacity (TFC) and the UHDRS Independence scale (UHDRS IS)). We also computed the composite cUHDRS

$$\left(\text{cUHDRS} = \left[\left(\frac{\text{TFC}-10.4}{1.9}\right) - \left(\frac{\text{TMS}-29.7}{14.9}\right) + \left(\frac{\text{SDMT}-28.4}{11.3}\right) + \left(\frac{\text{SW}-66.1}{20.1}\right)\right] + 10\right)$$

[5].

Standardised lightweight speech test

Speech samples were recorded through two brief controlled tasks by the examiner, who provided the instructions to the participants. Each participant was asked to (1) count aloud numbers from 1 to 20 (forward counting), then (2) to count the numbers backwards from 20 to 1 while holding his/her hands up and closing his/her eyes (backward counting). The rationales for these two subsequent tasks are: (1) we wanted to obtain a baseline performance for counting numbers with minimal cognitive load, (2) we wanted to measure performance of HD as cognitive load is higher, due to the inhibition of forward counting and dual tasking. Recording was performed either by video tape, microphone of the computer, or external microphone.

Speech features

Only samples without too much acoustic noise, as perceptually determined blindly by two speech therapists before data delivery were retained. Thirty five files were discarded in total (33 from MIG-HD, 2 from BIOHD/REPAIRHD). This yielded the analysis of 126 samples, from 103 patients. In the case of a second visit for a participant, this visit can be separated between 1 and 36 months after the first visit. Then, the two speech therapists blindly transcribed each sample at the word level; and when there was a mispronunciation, the word was transcribed at the phonetic level with the Speech Assessment Methods Phonetic Alphabet using the software Praat [34] and the Seshat platform [35]. This is based on the listening of the acoustic signal, and also visualisation of the acoustic signal along the spectrogram.

They identified paraphasias, phone perseverations, abnormal breathing, vocal noises, filled pauses ("euh", "um"), blocks, and prolongations (Table 2). Paraphasias, phone perseverations, blocks, and prolongations are pooled together to count as "pronunciation error". Abnormal breathing, vocal noises, and filled pauses are considered to play an important part in communication and are referred to as collateral track additions [36].

Time and categorizations of events differences between raters were systematically discussed until agreement between both annotators. Phones were then force-aligned using Hidden Markov models combined with Gaussian Mixture models based on the Kaldi toolkit [37]. An automatic pipeline algorithm was developed to extract the speech features previously selected on previous analyses of the MIG-HD, the exploration cohort. After exploration on MIG-HD, we preregistered all the methodologies before running the analyses on the BIOHD/REPAIRHD cohort.

Based on these annotations, the forced-alignment and the acoustic waveform, we extracted different speech deficits dimension already reported in HD: articulatory and phonatory deficiencies [16, 17, 27, 38], rhythm and temporal statistics [26, 39], filled pauses and vocalizations additions [26, 27, 29], sequence (the order of numbers), and perseveration errors (introduced here for measuring target sequence errors). In total, we examined 60 features that do not need to be adapted to a specific language (See Table 2 for the full detailed list of speech features).

Table 2 List of speech and language features extracted from the recitation of numbers

Dimension	Speech/language feature
Articulatory and phonatory deficiencies	Total number of pronunciations errors
	Ratio of pronunciation errors
	Pronunciation error per second
	Mean intelligibility based on non-intrusive normed speech-to-reverberation modulation energy ratio metric [40]
	SD of the fundamental frequency
	Range of the fundamental frequency
	SD of normalized intensity of vocalizations
	Normalized range of intensity of vocalizations
Rhythm and temporal statistics	Task duration
	Temporal rate of the pronounced numbers
	Mean duration of pronounced numbers
	Pronounced numbers per second
	SD of the duration of pronounced numbers
	Phones per second
	TR of the silences
	Mean duration of silences
	SD of the duration of silences
	Total number of silences
Sequence errors and perseverations	Levenshtein distance between the pronounced numbers and the target sequence (1, 2, …, 19, 20)
	Gestalt similarity between the pronounced numbers and the target sequence (1, 2, …, 19, 20)
	Levenshtein distance between the pronounced phones and the target sequence (phones of 1, phones of 2, …, phones of 19, phones of 20)
	Gestalt similarity between the pronounced phones and the target sequence (phones of 1, phones of 2, …, phones of 19, phones of 20)
	Total number of pronounced numbers
	Total number of pronounced phones
Collateral track additions	Total number of involuntary/abnormal vocalizations
	Involuntary/Abnormal vocalizations per second
	Temporal rate of the involuntary/abnormal vocalizations
	Total number of filled pauses
	Filled pauses per second
	Temporal rate of the filled pauses

SD stands for standard deviation, Temporal rate is defined as the ratio of the total time of a specific class on the total time to perform the task

Machine learning

We used the auto-machine-learning system, auto-sklearn [41] to predict the clinical variables from the speech features. Auto-sklearn uses Bayesian optimization algorithms to find the model with the best cross-validated performance on the training set. The model selection process is performed independently for each clinical score, yielding different predictors and models. We ran and compared three automatic machine learning pipelines by using different sets of inputs:

1) The speech features (Table 2) with the Demographic variables (Gender, Age, Number of CAG repeats, and Disease Burden Score). In machine learning experiments the relationship between the features and target variable is not always linear. Sometimes the relationship between dependent and independent variables is more complex such as polynomial transformation. That is why we used the combination of the Disease Burden Score alongside the Age and Number of CAG repeats.
2) Demographics variables alone, which allow predicting disease's onset and progression in HD (Gender, Age, Number of CAG repeats, and Disease Burden Score), and represent an important baseline to be compared to [42].
3) The mean baseline performance of each clinical score on the training set (called Cohort Mean Performance in the following sections), which represents the average performance of individuals in the training set. This Cohort Mean Performance is equivalent to what is usually performed with classic statistical methodologies when there is a will to replicate results across cohorts in medicine.

For the auto-machine learning approach, we followed the approach described in detail in the auto-sklearn article [41]. For the auto-machine learning approach, we set a 2-min time limit for each model training for each clinical score as defined by the auto-sklearn toolkit. Each training is limited to 30 s. We used 24 parallel processes for each clinical score and each model. Thus, the minimum number of models tested was therefore 96 models. Then all the best 50 models found on training data during this search are combined (through ensemble strategy).

To assess the respective importance of each speech feature to predict each clinical score, we used a linear regression model with an ElasticNet regularization (Fig. 5). We also ran an ablation study to evaluate the contributions of the backward and forward speech features. An ablation study is a term from the machine learning literature to refer to an experiment to evaluate contributions of specific features. This means that we run the same machine learning analyses based on the subset of features extracted of the forward counting, and on the subset of features extracted on the backward counting, to evaluate contributions of each.

Validation of models

We used both the Mean Absolute Error (MAE) and the intraclass correlation coefficient (ICC) between the predicted and the observed scores provided by the clinicians. The ICC measures how much the predicted clinical score outputted by the Machine Learning model resembles the observed score. ICC values were calculated using a two-way random model with absolute agreement. The use of ICC allows comparing the machine learning model to the interrater reliability of clinicians. The MAE quantifies the absolute errors between the observed clinical scores and the predicted scores.

Concerning the sample size of the current study, we wanted enough visits to train the models and enough visits to test the models. The problem of sample size and model validation for machine learning applied in Neurology and Psychiatry has been extensively studied with simulation in these studies [43, 44]. As underlined by the authors, "leave-one-out" strategy leads to unstable and biased estimates of the true performance of a model, and repeated random splits method should be preferred. 20% should be left out for the test set.

Thus, we splitted the data into two sets: "train set" (86% of the participants, i.e. 89 participants, including all participants of MIG-HD and 80% of the ones of RepairHD/BIOHD) for fitting and developing the various models and an independent "test set" (14% of the participants, i.e. 14 participants, consisting in the 20% remaining participants of RepairHD/BIOHD) for model evaluations. We conducted 50 repeated learning-tests to obtain reliable estimates of the performances. There was no overlap between participants of the training and of the test sets to ensure the generalisation of the results. Multiple visits of the same patients were assigned either to the training set, either to the test set to ensure independence.

In addition, the number of samples should be at least 100 to obtain less than 10% of variance on the test score based on the simulation [43, 44]. We had 103 participants and 126 visits in total in this study, which fulfilled all these requirements.

Identifying Significant Relationships with the Striatum.

The association between each of the 60 speech features and the striatal volumes was assessed in thirty-six participants from the BIOHD/REPAIRHD cohorts (23 females, mean age: 52.98 ± 12.56). High-resolution brain MRI scans were obtained on a Siemens Skyra including T1 3D anatomical MP-RAGE images (repetition time: 2300 ms; echo time: 2900 ms; inversion time: 900 ms; flip angle: 9°; acquisition matrix: 256×240; slice thickness: 1.2 mm, no inter-slice gap, 176 sagittal sections). We used the FreeSurfer software

(https://surfer.nmr.mgh.harvard.edu/) [45] for extracting subcortical volumes. Percentage of striatal volume relative to the estimated intracranial volume was obtained from the caudate nucleus, ventral striatum, and putamen volumes.

When number of associations to be tested is large with limited data, the assessment of significance of variables must consider that: (1) Measures of relationships need to yield a good probability of making a correct decision when assessing significance (power property), (2) the capability to measure the strength of any relationship (linear or not) at a given noise level (equitability property) and (3) the multi-comparison issue. We therefore used the mutual information-based estimators procedure, including the Total Information Coefficient estimator (TICe) and the Maximal Information Coefficient estimator (MICe) [46] to identify and measure the strengths of their relationships [47]. The TICe allows the screening of variables because of its high power, but low equitability and the MICe estimates the strengths of the relationships because of its high equitability but lower power. In addition, speech variables and clinical scores correlations were corrected for multiple comparisons with the Maximum Statistic correction to take into account the correlations between the variables [48].

Results

The duration for the forward (backward) recitation of numbers is 10.7 ± 3.6 (15.6 ± 5.6) seconds. The annotation lasted less than 15 min per file. Illustration of prediction performances of the cUHDRS and TMS are shown in Fig. 2; where each individual prediction error on one visit contributes to the MAE. Predicted clinical scores on the Test Set are displayed in Fig. 3 using the MAE metric and Fig. 4 using the ICC. Models based on the Speech features performed significantly better for the MAE, for all clinical variables, than the ones using the Demographics variables (Age, Gender, Numbers of CAG, and Disease Burden Score) or the Cohort Mean Performance (all P values < 0.0001 except for the Verbal Fluency P value $= 3.25 \times 10^{-3}$, Fig. 3). Models using the Demographics variables performed more accurately than the ones using the Cohort Mean Performance, (all P values < 0.0001 except for the Stroop Interference P value $= 1.32 \times 10^{-1}$, Fig. 3). Models based on the Speech variables performed significantly better for the ICC for all clinical variables, than the ones using the Demographics variables (all P values < 0.0001, Fig. 4). Among all variables the cUHDRS was the best predicted based on the ICC. This score is predicted with on average 2.3 points error using the combination of the speech features and demographics (MAE = 2.3 ± 0.5;*ICC*=0.72 ± 0.10). Speech and demographic features allowed 19.4% and 29.2% improvement over demographics alone for MAE and ICC respectively, and 40.1% over Cohort Mean Performance models for MAE.

An ablation study showed that the speech features from the backward counting obtain better results overall than the forward ones, and even better results than when combined with the forward ones. Forward speech features obtained for the different scores: cUHDRS MAE = 2.6 ± 0.5; TMS MAE = 11.7 ± 1.8; TFC MAE = 1.5 ± 0.2;

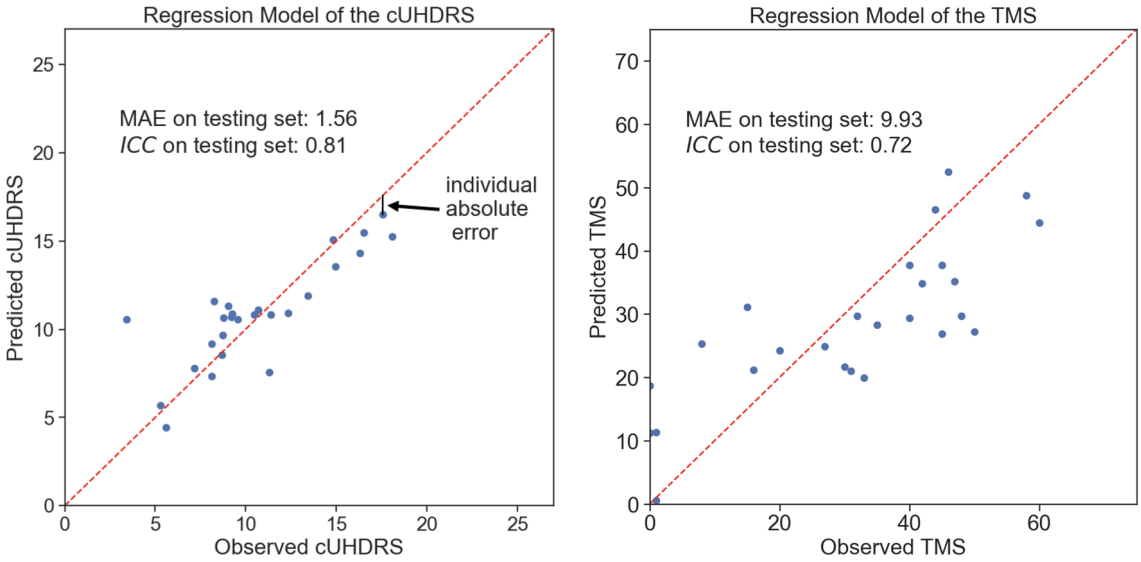


Fig. 2 Illustration of individual predictions of the cUHDRS (Left) and the TMS (Right) based on the speech features. Each individual blue dot is the difference between the predicted and the observed score for a particular assessment of an individual of the test set. The red dashed line is the line '$y = x$'. The black line is the individual contribution of a point (individual absolute error) to obtain the Mean Absolute Error (MAE)

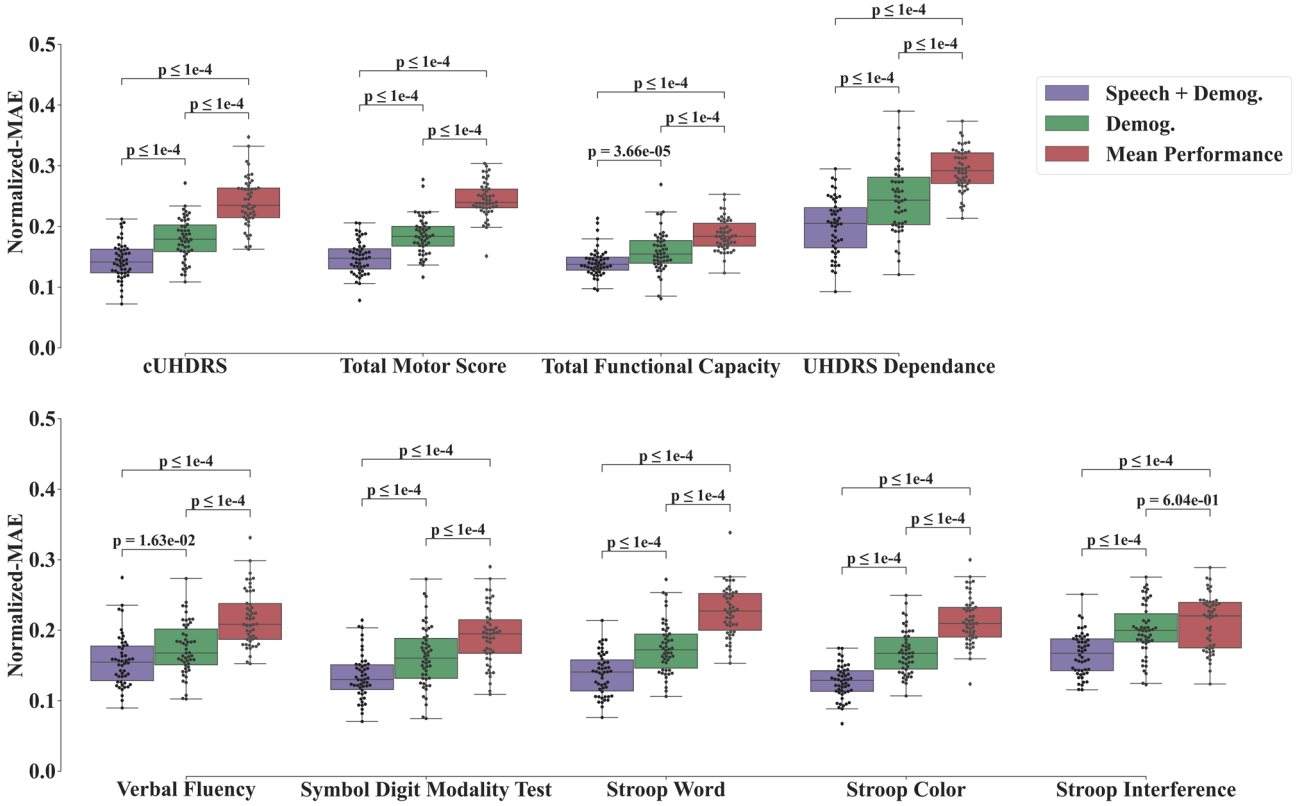


Fig. 3 Boxplots of mean-absolute-error (MAE) on the test set for the repeated-learning testing experiment. A MAE at zero means that the predicted value equals the observed one. Horizontal lines are the medians, boxes are upper and lower quartiles, and whiskers are 1.5 × IQR (Interquartile Range). First row displays the cUHDRS, functional, and motor predicted scores; whereas the second row displays the predicted Cognitive Scores. Statistical Significance was assessed with Wilcoxon-test and was Bonferroni-corrected

UHDRS IS MAE = 8.8 ± 1.2; VF MAE = 9.2 ± 1.3, SDMT MAE = 9.8 ± 1.8; SW MAE = 14.9 ± 3.1; SC MAE = 10.9 ± 1.8; SI MAE = 8.9 ± 1.7. The backward speech features obtained for the different scores: cUHDRS MAE = 2.4 ± 0.4; TMS MAE = 12.0 ± 1.8; TFC MAE = 1.3 ± 0.2; UHDRS IS MAE = 8.1 ± 1.2; VF MAE = 8.0 ± 1.0, SDMT MAE = 8.9 ± 1.8; SW MAE = 13.3 ± 2.2; SC MAE = 9.6 ± 1.7; SI MAE = 7.8 ± 1.5.

Some clinical variables (cUHDRS, TMS, SW, SDMT, and UHDRS IS) and speech features (both Mean duration and Standard Deviation of durations of Silences during backward recitation) correlated with the measure of the striatal atrophy (Table 3). Comparison correction was performed with the Maximum Statistic [48]. The Mean duration of Silences obtained the strongest strength of relationship based on the *MICe*, while the cUHDRS obtained the strongest linear relationship with the Pearson coefficient *R*.

The features that are the most used for predictions are the ones from backward counting (Fig. 5). Speech features extracted from the collateral track additions were less used overall than the other dimensions. Rhythm and temporal statistics were useful for both counting forward and backward.

Even if some coefficients have been set to 0, they may still be related to the clinical score outcome. The model chose to diminish their weights because they bring no additional information in comparison to the other speech features.

Discussion

Our multicentered prospective study aimed at predicting the clinical scores of different visits of 103 individuals carrying the mutant Htt gene leading to Huntington's disease, using machine learning analyses of speech productions. We used speech features extracted from forward and backward counting—a task that lasts less than 40 s, even in patients at an advanced stage. We showed that measures of speech production accurately predict the clinical measures in HD, within the 12% to 20% range for the functional, motor, and cognitive, and composite cUHDRS (The Mean Absolute Error is divided by the maximum observed range to obtain these values). Speech features improved predictions from demographics and genetics characteristics alone by around 17% in relative terms. In particular, the predicted cUHDRS had an equivalent inter-rater agreement score (ICC) in the

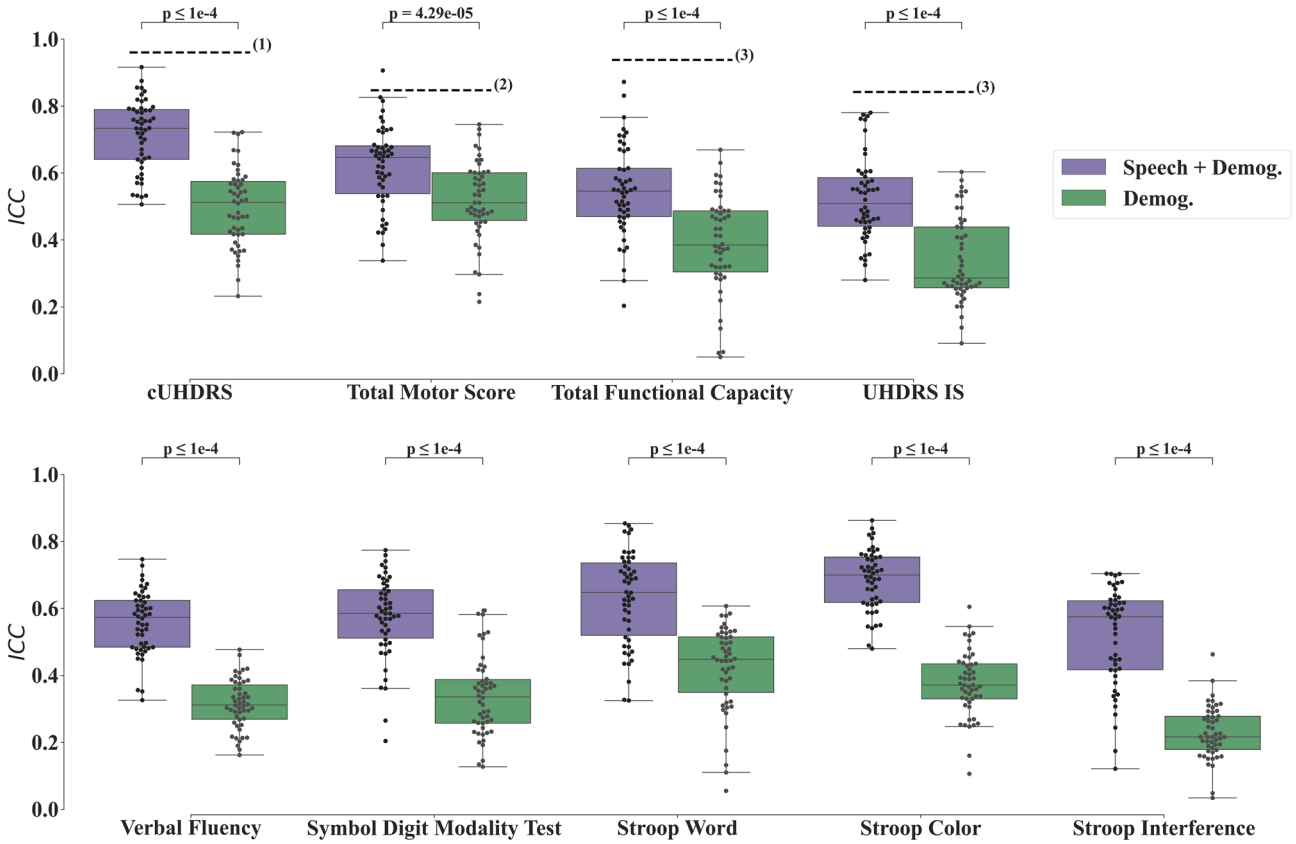


Fig. 4 Boxplots of intraclass correlation coefficients (ICC) on the test set for the repeated-learning testing experiment. An ICC at 1 means that the predicted value equals the observed one. Horizontal lines are the medians, boxes are upper and lower quartiles, and whiskers are 1.5 × IQR (Interquartile Range). First row displays the cUHDRS, functional, and motor predicted scores; whereas the second row displays the predicted Cognitive Scores. Statistical Significance was assessed with Wilcoxon-test and was Bonferroni-corrected. The dashed lines figure the ICCs obtained between Neurologists for the clinical scores namely: (1) ICC for cUHDRS ICC = 0.92 [49], (2) for TMS ICC = 0.847 [3], (3) for TFC ICC = 0.938, and for UHDRS IS ICC = 0.842 [4]. The ICC cannot be computed for the Mean Cohort Performance as its standard deviation is zero

Table 3 Summary of the speech and clinical variables with significant correlation with the Normalized Volume of the Striatum

	TICe P value	*MICe*	Pearson *R*	Spearman ρ
Speech				
Mean duration of the silences during backward recitation	0.0024	0.57	−0.35	−0.56
Standard deviation of the duration of the silences during backward recitation	0.026	0.49	−0.41	−0.60
Clinical variables				
cUHDRS	0.0050	0.40	0.65	0.68
UHDRS total motor score	0.0090	0.38	0.52	0.57
Stroop word	0.021	0.38	0.61	0.64
Symbol digit modality test	0.030	0.36	−0.63	−0.63
UHDRS independence scale	0.040	0.33	0.58	0.57

The comparison between the *TICe*'s *P* values [46], the measure of linear relationship with the Pearson *R* coefficient, the Spearman rank correlation coefficient ρ, the measure of strength of the relationship with the *MICe* shows that Mean duration of Silences and the Standard Deviation of the duration of Silences are as well correlated with the striatal volume than the regular clinical scores. Multiple Comparison correction is done with the Maximum Statistic [48]

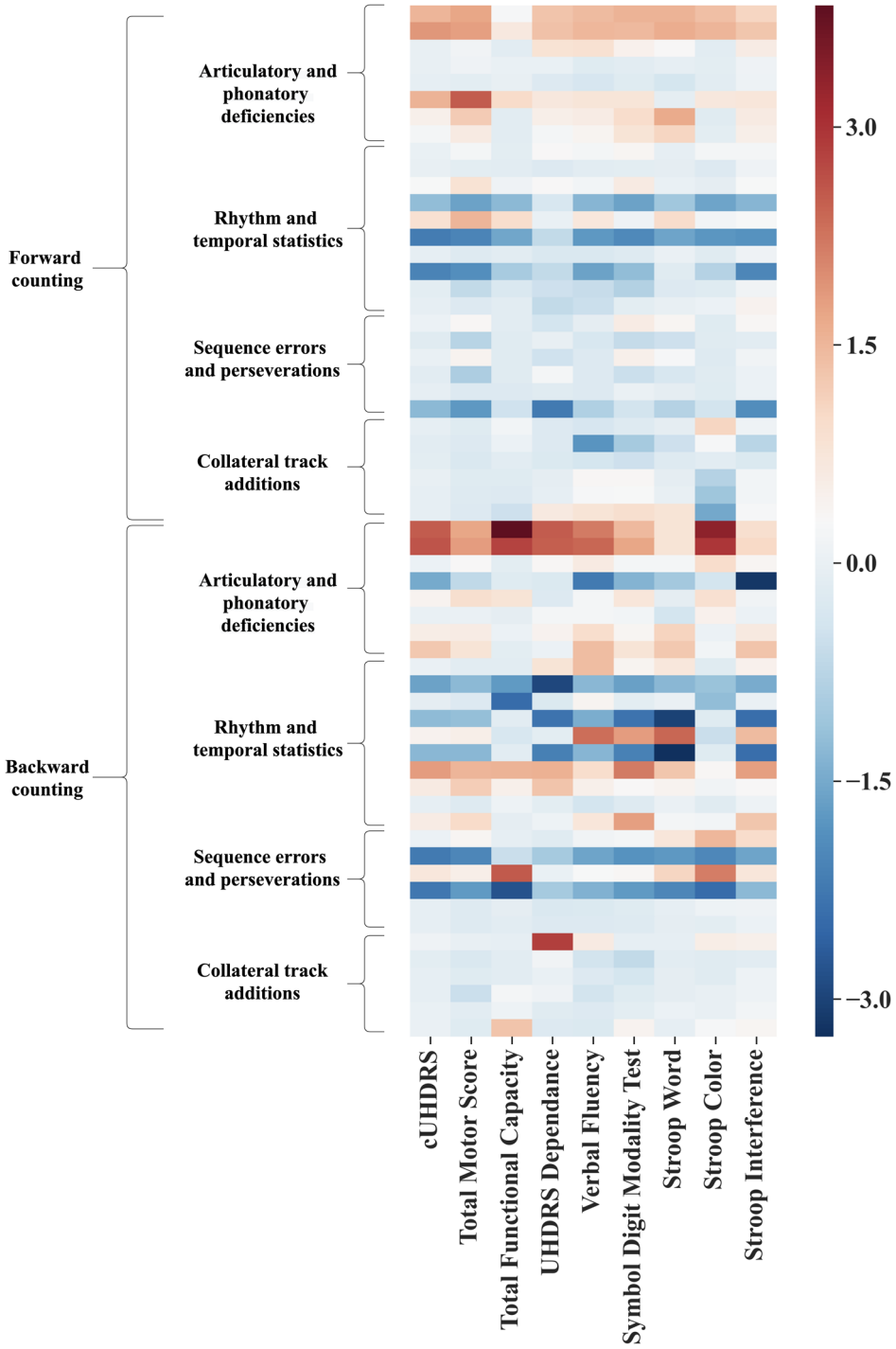


Fig. 5 Coefficient importance of the different speech features for the predictions of the clinical scores. Each line represents a feature of Table 2 and the rank is the order introduced in Table 2. These mean weights are obtained with a linear Elastic Net model for interpretability. The weights are z-scored per clinical score to be one the same scale. The weights for the clinical scores are reversed, so that a higher feature weight can be interpreted as a higher clinical impairment

"good" reliability range. Finally, the Mean Duration and Standard Deviation of Durations of Silences correlated significantly with the atrophy of the striatum.

These results may lead to the construction of reliable, discriminative and applicable diagnostic tools for the prediction of the progress of the symptoms. Our forward/backward counting task provides a good compromise between the different requirements for a usable language-based battery in a clinical setting: accuracy (to measure the evolution of the condition), ease of use, and multidimensionality (capability for one single marker to capture several dimensions of the disease [5]).

As for accuracy, for machine learning systems to be clinically valuable, assessing only the statistical significance of the group performance (here the Cohort Mean Performance) is insufficient [7]. The derived scores should be predictive

enough at the individual level to be used for clinical decision making. This is why, to assess their accuracy, we compared our predicted scores with standard tests performed by neurologists [12]. As expected, the ICCs from machine learning models did not match the ones of expert clinical raters [3, 4, 49]. However, their capacity to assess the patients frequently could reduce the cost to evaluate clinical therapies in HD by increasing the measures of an individual, thus permitting the reduction of the required number of participants in clinical trials [50].

As for ease of use, the forward and backward counting does not require the involvement of any expert nor training for patients' recording. This constitutes a major progress considering that despite its worldwide dissemination and its excellent acceptability, the interrater reliability of the UHDRS between neurologists decreases in absence of annual certification [3]. Audio data can be collected over the phone, allowing not only remote but also out of sync assessments between health professionals and patients [51]. The limited vocabulary and deterministic sequences expected from participants allows easier development of fully automated procedures potentially reducing further annotation time. In contrast, other batteries like the Cantab [52], and the HD-CAB [6], require longer assessments, are not easy to administer and cannot currently be performed remotely.

Finally, as regards multidimensionality, our simple speech test, allows measuring, on the top of language, the different components of the UHDRS (cognitive, motor and functional).

Our results are consistent with previous ones in HD concerning the different dimensions that are affected during spoken language production. Our 60 speech features coded articulatory and phonatory deficiencies, rhythm and temporal statistics, and added seldom studied collateral track additions, sequence and perseverations. We showed that rhythmic and articulatory features were particularly sensitive to the progress of the disease. Rhythmic features well reflected motor and cognitive disabilities (Fig. 5) and correlated the most with the striatal volume (Table 3). This latter result is consistent with Hinzen's findings on a storytelling task in which the composite quantitative score capturing the rhythm was the only one correlated with striatal atrophy. This confirms the involvement of the striatum in motor programs of phones and syllables, and their sequential structure and timing[14] Besides, we also found that articulatory features were linked to various HD deficits (global, motor, functional but not as much to the cognitive scores: SDMT, Stroop Word and Stroop Interference) like in previous reading tasks [18, 27, 29] and storytelling tasks [26].

We obtained robust estimations of clinical scores, even though using a relatively simple task. Yet, the strength of rhythmic and phonatory impairments in HD has been shown to depend on the cognitive load of the task used to elicit speech. Vogel and authors studied the speech disturbances of manifest and premanifest mutant Htt gene carriers while performing a spectrum of tasks from low to high cognitive load [27]. In their study, rhythmic deficits correlated with the TMS only when measured from a reading task (Percentage of silence $R = 0.4$) and a monologue task (Percentage of silence $R = 0.5$) but not from automated speech (recitation of the days of the week, Percentage of silence $R = 0.08$). Similarly, although HD participants have difficulties to sustain the vowel /a/ steadily for a few seconds compared to premanifest patients [30], speech features extracted from this simple task could not improve the clinical score extracted from demographics alone [38]. In our present study, we used both an automatic task (counting forward) and a more cognitive complex task (counting backward). A post hoc analysis shows that the forward counting task alone, which involves an automatic sequence yields lower predictions than the backward counting sequence. As described in the methods, when participants perform the backward counting, they need to inhibit the automatic number forward recitation and disengage from the overlearned forward sequence of numbers just previously performed. In addition, we used a dual task (of holding hands and closing eyes) [53] which is known to increase reaction times and errors [54]. As seen in Fig. 5, perseveration features are more salient in the backward compared to the forward test confirming the importance of cognitive load when estimating the symptoms of HD participants.

Here, we focused on the measurement of rather low-level acoustic features in a rather simple task for its potential for automation and applicability in different languages with minimal adaptation. Other studies have demonstrated that HD symptoms also include higher levels of language processing (conceptual, lexical, syntactic planning) [26]. Adding such high level features could improve the accuracy of a test battery over low level speech features. However, it was shown [55] that the extraction of high level features from 10 min of speech imposed two hours of annotation by experts including the identification of "Who speaks when?", "What is said?", and "How is it said?". Current Artificial Intelligence (AI) research is being done to replace the expert linguist by automatic systems in order to reduce the cost of analyzing such tests. an automatic speech recognition system that could recognize the words was built [29] ("What is said?") directly from audio recordings of the 'GrandFather Passage' story yielding to 85% accuracy when classifying HD from controls using speech features (speech rate, pauses, fillers, and goodness of pronunciation). However, humans were still required to segment manually the turns between doctor and patient, and the boundaries between sentences before feeding the automatic transcriber. Surprisingly, "Who speaks when" is still more challenging for algorithms than for humans when the audio comes from naturalistic and

clinical settings (see the low performance in engineering DIHARD challenges [56]). Even when using state-of-the-art models, the reliability of "Who speaks when" in a clinical context remains too low for clinical use [38]. More powerful models and larger datasets will eventually overcome these limitations. The combination of different objective sources is an opportunity to increase the predictive power of the clinical scores based on speech features. In future work, this would be of great interest to combine speech features to other objective measures such as the Q-motor [57]. Yet, this still represents a technical challenge as the number of dimensions to analyze increase.

Our study presents some limitations that could be overcome in future works. The number of participants limited to a hundred here might impact the generalization results. Focusing on French gene-carriers of the mutant Htt gene should not constitute a problem, the analysis of results from five languages in Parkinson's disease was found equivalent [58]. Our task was designed with as much as language-independent features, but it does not warrant the generalization of our results across languages and centers. Despite Huntington's disease combining the major features of NDDs—motor, psychiatric and cognitive disorders, the dissemination of our method requires validation in each individual disease of interest.

In conclusion, this is the first machine learning model combined with speech study that reliably estimated the scores of classical scales assessing several domains for pre-HD individuals and HD participants. One of its strengths is that the reliability of the predictive models closely match the observed data from neurologists and neuropsychologists for HD, without any ambiguity on the reliability of the data as methods were pre-registered before analyses. Being able to evaluate the severity of the different symptoms so quickly and potentially remotely has both clinical and experimental relevance in HD. This will likely reduce the human and financial burden for the follow-up of patients and help to reduce the cost of future disease modifying therapeutic trials.

Acknowledgements We are very thankful to the patients that participated in our study. We thank Marvin Lavechin, Alex Cristia for the multiple helpful discussions and we thank the speech pathologists for the annotations of the speech data.

Author contributions RR, ED, ACBL: research project: (A) conception, (B) organization, (C) execution, (D) supervision; statistical analysis: (A) design, (B) execution, (C) review and critique; manuscript preparation: (A) writing of the first draft, (B) review and critique. ML, HT, XNC, JHB, LL, JM, AS, KY, LCDL: statistical analysis: (A) design, (B) execution, (C) review and critique; manuscript preparation: (A) writing of the first draft, (B) review and critique.

Funding MIG-HD was funded from the "Direction de la Recherche Clinique" (Assistance Publique— Hôpitaux de Paris) by AOM00139 and AOM04021 grants, Repair-HD from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n°602245 (www.repair-hd-eu), Bio-HD by the Henri-Mondor Hospital National Reference Centre for Huntington's disease (Ministry of Health). The team is supported by—NeurATRIS ANR-11-INBS-0011/Agence Nationale de la Recherche (French National Research Agency) and ANR-17-EURE-0017. This work is also funded in part from the Agence Nationale pour la Recherche (ANR-17-EURE-0017 Frontcog, ANR-10-IDEX-0001-02 PSL*, ANR-19-P3IA-0001 PRAIRIE 3IA Institute). ED was funded through his EHESS role by Facebook AI Research (Research Gift) and CIFAR (Learning in Minds and Brains).

Declarations

Conflicts of interest Nothing to report.

References

1. Ross CA, Tabrizi SJ (2011) Huntington's disease: from molecular pathogenesis to clinical treatment. Lancet Neurol 10(1):83–98. https://doi.org/10.1016/S1474-4422(10)70245-3
2. (1996) Unified Huntington's disease rating scale: reliability and consistency. Huntington study group. Mov Disord Off J Mov Disord. Soc 11(2): 136–142. https://doi.org/10.1002/mds.870110204.
3. Winder JY, Roos RAC, Burgunder J, Marinus J, Reilmann R (2018) Interrater reliability of the unified huntington's disease rating scale-total motor score certification. Mov Disord Clin Pract 5(3):290–295. https://doi.org/10.1002/mdc3.12618
4. Winder JY, Achterberg WP, Marinus J, Gardiner SL, Roos RAC (2018) Assessment scales for patients with advanced Huntington's disease: comparison of the UHDRS and UHDRS-FAP. Mov Disord Clin Pract 5(5):527–533. https://doi.org/10.1002/mdc3.12646
5. Schobel SA et al (2017) Motor, cognitive, and functional declines contribute to a single progressive factor in early HD. Neurology 89(24):2495–2502. https://doi.org/10.1212/WNL.0000000000004743
6. Stout JC et al (2014) HD-CAB: a cognitive assessment battery for clinical trials in Huntington's disease 1,2,3. Mov Disord Off J Mov Disord Soc 29(10):1281–1288. https://doi.org/10.1002/mds.25964
7. Mason SL et al (2018) Predicting clinical diagnosis in Huntington's disease: an imaging polymarker. Ann Neurol 83(3):532–543. https://doi.org/10.1002/ana.25171
8. Scahill RI et al (2020) Biological and clinical characteristics of gene carriers far from predicted onset in the Huntington's disease young adult study (HD-YAS): a cross-sectional analysis. Lancet

Neurol 19(6):502–512. https://doi.org/10.1016/S1474-4422(20)30143-5
9. Zhan A et al (2018) Using smartphones and machine learning to quantify Parkinson disease severity: the mobile parkinson disease score. JAMA Neurol 75(7):876–880. https://doi.org/10.1001/jamaneurol.2018.0809
10. Bechtel N et al (2010) Tapping linked to function and structure in premanifest and symptomatic Huntington disease. Neurology 75(24):2150–2160. https://doi.org/10.1212/WNL.0b013e3182020123 (**e-Pub ahead of print**)
11. Gajos KZ et al (2020) Computer mouse use captures ataxia and parkinsonism, enabling accurate measurement and detection. Mov Disord 35(2):354–358
12. Wilkinson J et al (2020) Time to reality check the promises of machine learning-powered precision medicine. Lancet Digit Health 2(12):e677–e680. https://doi.org/10.1016/S2589-7500(20)30200-4
13. Fernandes BS, Williams LM, Steiner J, Leboyer M, Carvalho AF, Berk M (2017) The new field of 'precision psychiatry.' BMC Med 15(1):80. https://doi.org/10.1186/s12916-017-0849-x
14. Guenther FH (2016) Neural control of speech. MIT Press, Cambridge
15. Levelt WJM (1993) Speaking: from intention to articulation. MIT Press, Cambridge
16. Rusz J et al (2013) Objective acoustic quantification of phonatory dysfunction in Huntington's Disease. PLoS ONE 8(6):e65881. https://doi.org/10.1371/journal.pone.0065881
17. Rusz J, Saft C, Schlegel U, Hoffman R, Skodda S (2014) Phonatory dysfunction as a preclinical symptom of Huntington Disease. PLoS ONE 9(11):e113412. https://doi.org/10.1371/journal.pone.0113412
18. Rusz J et al (2014) Characteristics and occurrence of speech impairment in Huntington's disease: possible influence of antipsychotic medication. J Neural Transm 121(12):1529–1539. https://doi.org/10.1007/s00702-014-1229-8
19. Skodda S et al (2016) Two different phenomena in basic motor speech performance in premanifest Huntington disease. Neurology 86(14):1329–1335. https://doi.org/10.1212/WNL.0000000000002550
20. Skodda S, Schlegel U, Hoffmann R, Saft C (1996) Impaired motor speech performance in Huntington's disease. J Neural Transm 121(4):399–407. https://doi.org/10.1007/s00702-013-1115-9
21. Ramig LA (1986) Acoustic analyses of phonation in patients with Huntington's disease. Preliminary report. Ann Otol Rhinol Laryngol 95(3 Pt 1):288–293. https://doi.org/10.1177/000348948609500315
22. Velasco García MJ, Cobeta I, Martín G, Alonso-Navarro H, Jimenez-Jimenez FJ (2011) Acoustic analysis of voice in Huntington's disease patients. J Voice Found 25(2):208–217. https://doi.org/10.1016/j.jvoice.2009.08.007
23. Németh D et al (2012) Language deficits in Pre-Symptomatic Huntington's Disease: Evidence from Hungarian. Brain Lang 121(3):248–253. https://doi.org/10.1016/j.bandl.2012.04.001
24. Wallesch C-W, Fehrenbach RA (1988) On the neurolinguistic nature of language abnormalities in Huntington's disease. J Neurol Neurosurg Psychiatry 51(3):367–373
25. Chenery HJ, Copland DA, Murdoch BE (2002) Complex language functions and subcortical mechanisms: evidence from Huntington's disease and patients with non-thalamic subcortical lesions. Int J Lang Commun Disord 37(4):459–474. https://doi.org/10.1080/1368282021000007730
26. Hinzen W et al (2018) "A systematic linguistic profile of spontaneous narrative speech in pre-symptomatic and early stage Huntington's disease", Cortex. J Devoted Study Nerv Syst Behav 100:71–83. https://doi.org/10.1016/j.cortex.2017.07.022
27. Vogel AP, Shirbin C, Churchyard AJ, Stout JC (2012) Speech acoustic markers of early stage and prodromal Huntington's disease: a marker of disease onset? Neuropsychologia 50(14):3273–3278. https://doi.org/10.1016/j.neuropsychologia.2012.09.011
28. Hertrich I, Ackermann H (1994) Acoustic analysis of speech timing in Huntington's disease. Brain Lang 47(2):182–196. https://doi.org/10.1006/brln.1994.1048
29. Perez et al M (2018) Classification of huntington disease using acoustic and lexical features. In: Interspeech, ISCA, Hyderabad India, pp.1898–1902. https://www.isca-speech.org/archive_v0/Interspeech_2018/abstracts/2029.html
30. Romana A, Bandon J, Carlozzi N, Roberts A, Provost EM (2020) Classification of manifest Huntington disease using vowel distortion measures. Interspeech 2020:4966–4970. https://doi.org/10.21437/interspeech.2020-2724
31. Chan JCS, Stout JC, Vogel AP (2019) Speech in prodromal and symptomatic Huntington's disease as a model of measuring onset and progression in dominantly inherited neurodegenerative diseases. Neurosci Biobehav Rev 107:450–460. https://doi.org/10.1016/j.neubiorev.2019.08.009
32. Shoulson I (1981) Huntington disease: functional capacities in patients treated with neuroleptic and antidepressant drugs. Neurology 31(10):1333–1335
33. Tabrizi SJ et al (2009) Biological and clinical manifestations of Huntington's disease in the longitudinal TRACK-HD study: cross-sectional analysis of baseline data. Lancet Neurol 8(9):791–801. https://doi.org/10.1016/S1474-4422(09)70170-X
34. Boersma P (2006) Praat: doing phonetics by computer," https://www.Praat.Org.
35. Titeux et al H (2020) Seshat: a tool for managing and verifying annotation campaigns of audio data. In: Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, May 2020, pp. 6976–6982. Accessed: Nov. 09, 2020. [Online]. Available: https://www.aclweb.org/anthology/2020.lrec-1.861
36. Clark HH (1996) Using language. Cambridge University Press, Cambridge
37. Povey et al D (2014) The Kaldi speech recognition toolkit. In: Proc. ASRU, 2011, pp. 1–4. Accessed: Nov. 19, 2014. [Online]. Available: http://www.researchgate.net/publication/228828379_The_Kaldi_speech_recognition_toolkit/file/79e4150743dc6ce65c.pdf
38. Riad et al R (2020) Vocal markers from sustained phonation in Huntington's disease. Proc. Interspeech, 1893–1897, https://doi.org/10.21437/Interspeech.2020-1057
39. Ludlow CL, Connor NP, Bassich CJ (1987) Speech timing in Parkinson's and Huntington's disease. Brain Lang 32(2):195–214. https://doi.org/10.1016/0093-934x(87)90124-6
40. Santos JF, Falk TH (2014) Updating the SRMR-CI metric for improved intelligibility prediction for cochlear implant users. IEEEACM Trans Audio Speech Lang Process 22(12):2197–2206. https://doi.org/10.1109/TASLP.2014.2363788
41. Feurer M, Klein A, Eggensperger K, Springenberg J, Blum M, Hutter F (2015) Efficient and robust automated machine learning. Adv Neural Inf Process Syst 28:2962–2970
42. Rodrigues FB et al (2020) Mutant huntingtin and neurofilament light have distinct longitudinal dynamics in Huntington's disease. Sci Transl Med. https://doi.org/10.1126/scitranslmed.abc2888
43. Poldrack RA, Huckins G, Varoquaux G (2019) Establishment of best practices for evidence for prediction: a review. JAMA Psychiat. https://doi.org/10.1001/jamapsychiatry.2019.3671
44. Varoquaux G (2017) Cross-validation failure: small sample sizes lead to large error bars. Neuroimage. https://doi.org/10.1016/j.neuroimage.2017.06.061

45. Fischl B et al (2002) Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33(3):341–355. https://doi.org/10.1016/S0896-6273(02)00569-X
46. Reshef YA, Reshef DN, Finucane HK, Sabeti PC, Mitzenmacher M (2016) Measuring dependence powerfully and equitably. J Mach Learn Res 17(211):1–63
47. Albanese D, Riccadonna S, Donati C, Franceschi P (2018) A practical tool for maximal information coefficient analysis. GigaScience. https://doi.org/10.1093/gigascience/giy032
48. Nichols TE, Holmes AP (2002) Nonparametric permutation tests for functional neuroimaging: A primer with examples. Hum Brain Mapp 15(1):1–25. https://doi.org/10.1002/hbm.1058
49. Trundell D, Palermo G, Schobel S, Long JD, Leavitt BR, Tabrizi SJ (2018) F23 Validity, reliability, ability to detect change and meaningful within-patient change of the CUHDRS. BMJ Publishing Group Ltd, London
50. Yi Q, Panzarella T (2002) Estimating sample size for tests on trends across repeated measurements with missing data based on the interaction term in a mixed model. Control Clin Trials 23(5):481–496. https://doi.org/10.1016/S0197-2456(02)00223-4
51. Arias-Vergara T, Klumpp P, Vasquez J, Orozco JR, Noeth E (2017) Parkinson's disease progression assessment from speech using a mobile device-based application. Springer, Cham, pp 371–379. https://doi.org/10.1007/978-3-319-64206-2_42
52. Robbins TW, James M, Owen AM, Sahakian BJ, McInnes L, Rabbitt P (1994) Cambridge neuropsychological test automated battery (CANTAB): a factor analytic study of a large sample of normal elderly volunteers. Dement Geriatr Cogn Disord 5(5):266–281
53. Lo J et al (2020) Dual tasking impairments are associated with striatal pathology in Huntington's disease. Ann Clin Transl Neurol 7(9):1608–1619. https://doi.org/10.1002/acn3.51142
54. Mayr U, Keele SW (2000) Changing internal constraints on action: the role of backward inhibition. J Exp Psychol Gen 129(1):4
55. Rofes A et al (2018) Language in individuals with left hemisphere tumors: is spontaneous speech analysis comparable to formal testing? J Clin Exp Neuropsychol 40(7):722–732
56. Ryant N, Church K, Cieri C, Cristia A, Du J, Ganapathy S, Liberman M(2019) The Second DIHARD Diarization Challenge: Dataset, Task, and Baselines. Interspeech 978–982. https://doi.org/10.21437/Interspeech.2019-1268
57. Reilmann R, Schubert R (2017) Motor outcome measures in Huntington disease clinical trials. Handb Clin Neurol 144:209–225. https://doi.org/10.1016/B978-0-12-801893-4.00018-3
58. Rusz J et al (2021) Speech biomarkers in rapid eye movement sleep behavior disorder and parkinson disease. Ann Neurol. https://doi.org/10.1002/ana.26085

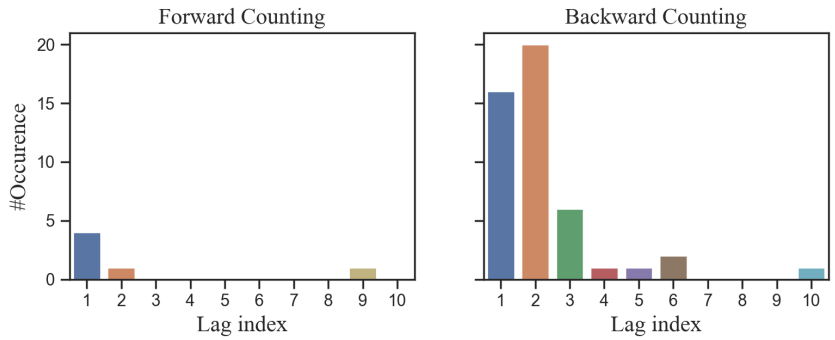


Fig. 4.3.: Lag distribution analyses during forward and backward counting of the first 20 numbers (1-20) and (20-1). These perseveration analyses are adapted from (Cohen & Dehaene, 1998). These are results for the BIO-HD/REPAIRHD cohort only for participants the Huntingtin's gene (preHD and HD). #Occurences refers to the number of occurences.

In addition to this study, we ran additional analyses only on the validation cohort, on the BasalVoice database. We wanted to better understand the perseveration patterns and not only assess the degree of severity. Indeed, edit-distance and gestalt-similarity provided us quantitative measures of the perseverations, they did not provide us any insights on *how* they occur.

That is why we followed the perseveration analysis introduced by Cohen and Dehaene (1998) to answer this question. In their study, the authors introduced a number of variables and methods to study the perseveration, especially in spontaneous speech productions. They introduce normalization of perseverations and also a way to correct for the chance level of phones. As we asked the participants to count numbers, we did not need these. That is why we used the simple 'Lag distribution analysis'. We looked at the series of responses and matched each error to a preceding response. We look backward in the counting series until we find a matching pattern. We then record the number that separates the two answers (the 'lag'). We can then visualise the frequency distribution of these lags, computed over all the errors.

We applied this lag distribution analysis to the forward and backward counting tasks. We gathered the errors for all HD and preHD participants in our study. Figure 4.3 shows the resulting lag distributions. As seen in the left figure (lag distribution for the counting forward task) there are almost no error, only a few hesitations, by just repeating the last number (ex: 4 perseverations of the previously pronounced number, i.e. lag=1). The results of the lag distribution analysis are very different for the counting backward. As we underlined in the discussions of the paper, the cognitive load is way more important in the backward setting. In this

case, individuals with HD exhibit way more difficulties to follow the good sequence of numbers, and just the automatic way to count number (lag=2, ex: 17 16 15 16). There seems to be an underlying distribution around lag=2, and a power law for bigger lag indexes. These perseveration results confirm the results from (Cohen & Dehaene, 1998). Further investigation of this phenomenon represents a great avenue to study internal decaying variable, and capacity to activate and deactivate previous symbols.

Tab. 4.4.: Results of the statistical analyses for spontaneous speech measures: Healthy Controls (HC), asymptomatic genetic carrier of Huntington's Disease (preHD), symptomatic genetic carrier of Huntington's Disease (HD). The p-values significativity of the tests (Kruskal-Wallis tests H-statistic (H stat) and post-hoc pairwise tests on the Cohen's d) are reported with *: $0.01 < p \leq 0.05$, **: $0.001 < p \leq 0.01$, ***: $p \leq 0.001$. The post-hoc statistics are corrected for multiple comparison feature wise. SD stands for standard deviation.

	Mean (SD)			H-stat	Effect size Cohen's d		
	HC (N=24)	preHD (N=27)	HD (N=74)		HD/PreHD	HD/HC	preHD/HC
Feature							
Ratio of Track 1	0.65 (0.07)	0.6 (0.15)	0.54 (0.16)	14.98**	-0.37*	-0.77**	-0.43
Ratio of Track 2	0.3 (0.06)	0.29 (0.09)	0.35 (0.11)	8.5	0.50*	0.47	-0.06
Ratio of incidental	0.17 (0.06)	0.17 (0.08)	0.24 (0.11)	17.103***	0.61**	0.70**	0.11
Ratio of ex-official	0.06 (0.03)	0.05 (0.04)	0.06 (0.04)	0.5	0.11	0.00	-0.12
Ratio of secondary	0.08 (0.03)	0.08 (0.03)	0.06 (0.03)	15.0**	-0.68**	-0.80**	-0.09

4.1.4 Markers from open-vocabulary speech tasks

In the previous sections of this chapter, we described our progresses concerning the validation of spoken language markers from vowel measures (sustained phonation and vowel distortion) and also from the combination of two closed-vocabulary tasks. Yet, it is important to evaluate the spoken language production of HD, in more open and spontaneous settings, to extend our knowledge and also aim for potential better measures with better real-world evidence.

Therefore, here we made the first attempts to examine if the tracks of communication and their timings are related to the progression of the disease. This analysis is made possible thanks to our parsing methodology introduced in the section 2.3. We normalized the duration of the different tracks of communications by the total duration of vocalizations of the interviewee, like the Speech Efficiency Score introduced in (Amir et al., 2018). We obtained ratio for each tracks and ran statistical comparisons like in the previous section 4.1.1 and the section 4.1.2. These group-level analyses on the BasalVoice database are still ongoing, and the results in this section are preliminary and exploratory. In this part we focused only on the spontaneous tasks.

We observed strong effects to differentiate HD from the other populations for Ratio of Track 1, Ratio of incidental and Ratio of secondary. This could be interpreted in a diminution of an overall efficiency of communication when the disease officially starts (Ratio of Track 1). This diminution of efficiency is large but there are competing effects that explain these variations. The Ratio of secondary track is reduced, this means that HD use for a shorter time filled pauses to regulate their conversations. A diminution of the Ratio of secondary track alone, would increase the Ratio of Track 1 in theory. Yet, the overall efficiency is diminished overall due to a large increase of Ratio of incidental. This could be interpreted as: Filled pauses are *normal* behaviors and signals during conversations, while incidental elements such as vocal noises, phonological fragments or unintelligble words are *symptoms* of the disease.

4.2 Vocal and linguistic expression impairments of emotions in Huntington's disease

In previous section, we investigated to what extent spoken language markers could be informative concerning HD's clinical symptoms. Yet, the emotional capabilities during spoken language production have not been studied so far. Yet, the reported disruption of social interactions and communication problems are the major cause of social withdrawal. It has been reported that HD patients are impaired in the perception of vocal and facial emotions and this hampers their abilities to navigate social situations (Kordsachia et al., 2017; Trinkler et al., 2013). Here, we developed the necessary methods to run the first experiments to investigate emotional speech production in HD. We did not rely on perceptual studies but we considered spoken language markers automatically extracted from speech and linguistics annotations. We quantified the identifiability of emotions in speech for each group HC, preHD and HD and compare accuracies. All the methods have been pre-registered before running any analyses. We include in the Appendix A.2 2 pre-registrations: 1 for the vocal modality and 1 for the linguistic modality.

This study has been carried out with the Master student Charlotte Gallezot, and the findings are in the journal Cortex (Gallezot* et al., 2022) and is included below.

Vocal and linguistic emotion expression deficits in Huntington's disease

Charlotte Gallezot[a,b,*], Rachid Riad[a,b,*], Hadrien Titeux[a], Laurie Lemoine[b,c], Justine Montillot[b,c], Agnes Sliwinski[b,c], Jennifer Hamet Bagnou[b,c], Xuan Nga Cao[a], Katia Youssov[b,c], Emmanuel Dupoux[a] and Anne-Catherine Bachoud Levi[b,c]

[a] CoML/ENS/CNRS/EHESS/INRIA/PSL Research University, Paris, France
[b] NPI/ENS/INSERM/UPEC/PSL Research University, Creteil, France
[c] Assistance Publique-Hôpitaux de Paris, National reference center for Huntington's Disease, Groupe Henri-Mondor Albert-Chenevier, Créteil, France

ARTICLE INFO

Keywords:
Huntington's disease
Speech
Machine learning
Emotions

ABSTRACT

Patients with Huntington's disease suffer from disturbances in the perception of emotions; they do not correctly read the body, vocal and facial expressions of others. With regard to the expression of emotions, it has been shown that they are impaired in expressing emotions through face but up until now, little research has been conducted about their ability to express emotions through spoken language. To better understand emotion production in both voice and language in Huntington's Disease (HD), we tested 115 individuals: 68 patients (HD), 22 participants carrying the mutant HD gene without any motor symptoms (pre-manifest HD), and 25 controls in a single-centre prospective observational follow-up study. Participants were recorded in interviews in which they were asked to recall sad, angry, happy, and neutral stories. Emotion expression through voice and language was investigated by comparing the identifiability of emotions expressed by controls, preHD and HD patients in these interviews. To assess separately vocal and linguistic expression of emotions in a blind design, we used machine learning models instead of a human jury performing a forced-choice recognition test. Results from this study showed that patients with HD had difficulty expressing emotions through both voice and language compared to preHD participants and controls, who behaved similarly and above chance. In addition, we did not find any differences in expression of emotions between preHD and healthy controls. We further validated our newly proposed methodology with a human jury on the speech produced by the controls. These results are consistent with the hypothesis that emotional deficits in HD are caused by impaired sensori-motor representations of emotions, in line with embodied cognition theories. This study also shows how machine learning models can be leveraged to assess emotion expression in a blind and reproducible way.

1. Introduction

Huntington's disease (HD) is a rare autosomal-dominant neurodegenerative disorder that primarily affects the striatum [58, 53]. It is characterised by motor, cognitive and psychiatric disorders, with a median progressive course leading to death within 35 years from symptom onset [59]. While motor disorders have been the subject of most studies, they are relatively well tolerated by the patients' entourage [22]. Conversely, the disruption of patients' social interactions and communication difficulties are one the major causes of patients' social withdrawal and family break-ups. Despite their major importance, they remain poorly understood and insufficiently quantified [16, 25, 17, 22]. Effective communication requires cognitive and linguistic skills, but also social abilities such as the perception and production of emotions. Deficits in emotion perception are a recognised symptom of patients with Huntington's disease. It impairs their ability to decipher and navigate social situations [7, 57]. However, less is known about these individuals' ability to express emotions, presumably because assessing emotion production is more technically challenging than emotion perception. In particular, the expression of emotions through spoken language has not been explored [26], despite its critical role for interpersonal communication. This information became even more crucial during the pandemic with the use of telephone communication as the only mean of interaction. The aim of our study is therefore to assess the ability of patients to produce emotions through spoken language.

There is ample evidence that patients with HD have impairments in the recognition of emotional faces [26, 7], body expressions [14, 62] and voices [50, 51, 26], and these impairments extend to negative and positive emotions [44]. These deficits can be detected even in the pre-manifest stage of the disease, before the onset of motor symptoms in carriers of the mutant huntingtin gene[7, 24]. Regarding emotion production, only the impairment of facial emotion production has been established [19, 57, 56, 26]. Given that there is a physiological congruence between body and facial expression [62], it can be assumed that body and facial emotion expression are concomitantly impaired in healthy participants. Even though speech plays a critical role in human interpersonal communication, to the best of our knowledge, no study has yet investigated the expression of emotions through spoken language in HD. See Figure 1 for schematic overview of what is known concerning emotional processes in HD. The communication of emotions through spoken language is reflected both in the voice and the linguistic content (called language here). These two media depend on different brain structures [15, 13] and can be altered separately. It has been shown that different dimensions of the voice, such as its fundamental frequency, energy, or speech rate, are affected by emotions [12, 47, 38]. Similarly, the individual's affective

*Authors contributed equally to this work.
ORCID(s):

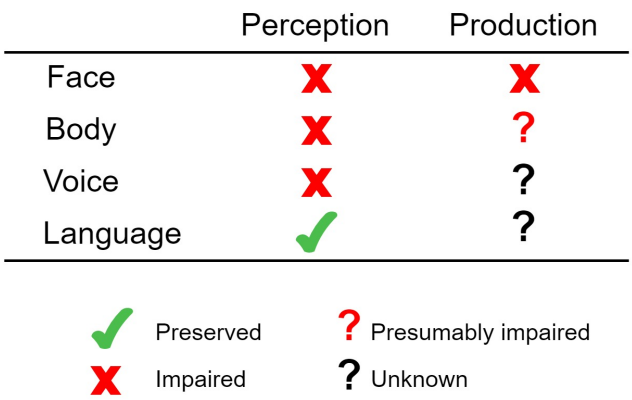


Figure 1: Summary of emotional processes in Huntington's disease.

state impacts on linguistic production, including word choice and syntactic structure [48]. Despite their language processing disorders such as syntax or morphology processing (See [23] for complete review), it has been shown that individuals with HD do not seem to be impaired in the perception of emotions conveyed by language. Indeed, as observed in a semantic task associating words and emotional content [18] and in the interpretation of a story [57], participants with HD obtain similar results to those of controls. In contrast, they are impaired in the perception of emotions through voice [18, 44, 26]. Assessing the production and the perception of emotions by voice or linguistic content is therefore essential to reach a global vision of emotion processing in HD patients.

Here, we sought to fill these gaps in the literature by testing independently the expression of emotions through voice and language in preHD and HD individuals. Yet, dissociating emotion expression carried by voice and language in speech represents a methodological and technological challenge. Studies exploring facial emotional expressions [19, 57] used external human scorers to subjectively classify emotions using a forced-choice recognition test. However this is not applicable to our study, since humans cannot perceptually separate the production of emotions by voice and language. Speech can be filtered to make language unintelligible, but at the cost of information loss. Moreover, HD individuals suffer from voice impairments unrelated to emotions that easily distinguishes them from controls [46, 40, 8, 43]. Indeed, patients exhibit speech disorders that are not emotion-specific, such as dysarthria or speech initiation disorders [28, 8, 46]. Mild speech impairments can even be detected in preHD participants [46, 8, 43]. Therefore, it is difficult to obtain a blind and deconfounded comparison of the emotions conveyed in voice and language with naive listeners. Apart from a human jury, studies on emotional prosody also used statistical analysis of the speech signal (ex: fundamental frequency, energy) [2, 33]. However, the link between emotional states and signal characteristics, such as fundamental frequency, intensity, formants are not straightforward [48]. This means that we do not know the exact formula that connects emotions, such as anger or joy, to the exact value of the speech signal, especially given the importance of variations across individuals. This limits the information from simple statistical analyses of these signal parameters to study vocal emotion production. The same notion applies to the linguistic information (syntax and semantic) that can be extracted from spoken sentences. To overcome these difficulties, we developed machine learning emotion recognition models that act as an expert panel to compare the expression of emotions through voice and language between controls, preHD participants and HD participants. This strategy allowed us to assess emotion expression from voice and language separately and without being biased by the disease severity or by spoken language problems unrelated to emotions.

We therefore collected recordings of neutral and emotional speech from healthy controls, preHD individuals and HD individuals. Taking as an assumption that recalling an emotional story elicits the corresponding emotion [34, 4], we obtained emotional discourses by recording participants while they narrate stories associated with fear, anger and joy. The participants' interviews were segmented and transcribed by speech therapy students. Then, in order to compare emotions identifiability between controls, preHD individuals and HD individuals, we compared how machine learning models classified emotions from the vocal signal or the annotated text for each group.

2. Material and methods

2.1. Overview

We collected emotional speech at the hospital by asking control, preHD and HD participants, to narrate emotional (anger, sadness, joy) and neutral stories. Interviews were split in stretches and labelled with the name of the elicited emotion. Stretches consist of semantically consistent chunks [49, 55]. Stretches from each group (HD, preHD, and controls) and each modality (voice and language) were combined into six sets (3 groups x 2 modalities). We then trained and tested one emotion classifier on each of our six datasets to assess emotion expression through both modalities and for all groups separately (methods are displayed in Figure 2A). Emotion classifiers are algorithms that can learn how to distinguish emotions on a dataset (either audio signal or text annotation). The classifiers play the role of an expert jury performing a forced-choice recognition test. We compared the labelling of emotion provided by the classifiers for each stretch to its actual label to measure the accuracy of each classifier. The rationale to use machine learning to examine emotion production is the following: all things being equal in terms of training size, machine learning model and all other parameters, the ability to classify accurately the emotion of a stretch should be equivalent for each group if there is no difference in emotion production. Any difference compared to control will indicate a difference in the group capacity to express emotions. For the sake of simplicity, we will use the term language for linguistic content in our article.

Table 1
Interviewee demographics and clinical scores. DBS: Disease Burden Score; cUHDRS: composite Unified Huntington's Disease Rating Scale; TFC: Total functional capacity; SW: Stroop word reading; SDMT Symbol Digit modality Test; TMS: Total Motor Score; CAG repeats: number of CAG repeats. The Disease Burden Score is computed as age $\times$ (CAG-35.5) (Langhben ref)

	Controls	Huntington's disease gene carriers	
Sub-groups		preHD	HD
N	25	22	68
Gender	12F/13M	11F/11M	42F/26M
Age (years)	53.6 (8.8)	49.0 (11.9)	54.3 (10.9)
CAG Triplets	-	41.7 (2.3)	43.8 (3.1)
DBS	-	42.9 (13.2)	46.1 (13.3)
cUHDRS	-	16.9 (1.4)	9.6 (3.5)
TFC	-	13 (0)	10.5 (2.1)
SW	-	98.8 (12.9)	64.6 (20.7)
SC	-	73.0 (12.1)	46.4 (14.9)
SI	-	43.9 (10.6)	26.0 (10.2)
SDMT	-	51.5 (11.5)	26.5 (9.8)
TMS	-	0.5 (1.2)	33.2 (15.8)

Methods for the voice and language experiments were preregistered before running any analyses to ensure the validity and avoid inflated results (links aspredicted). We further validated our methods with human judgement of the vocal and linguistic production as it is done in classical experiments [20, 57].

2.2. Participants

In this study, 115 participants were included from two observational cohorts (BIOHD NCT01412125 and Repair-HD NCT03119246) at the Hospital Henri-Mondor Créteil, France: 90 participants with at least 36 CAG repeats on the mutant Htt gene (including 22 gene carriers without the manifest disease (preHD), 39 patients at Stage 1, 27 Stage at 2, 2 at Stage 3 according to the total functional score (Shoulson, 1981)), and 25 healthy controls (Table 1). Participants carrying the mutant Htt gene were considered as preHD if both their Total Motor Score (TMS) is less than 5 [52] and their Total functional capacity (TFC) equals 13 using the Unified Huntington's Disease Rating Scale (UHDRS, Huntington Study Group, 1996). Participants were all French native speakers. They all signed an informed consent form. Ethical approval was given by the institutional review board from Henri Mondor Hospital (Créteil, France) for the Bio HD study and the CPP Saint Louis French part of the Repair-HD study. It complied with the Helsinki Declaration, current Good Clinical Practice guidelines, and local laws and regulations. None of the participants had any previous or current language, neurological or psychiatric history except HD.

Concerning the clinical evaluation, participants were assessed by certified examiners through the UHDRS. We reported nine measures classically used for both clinical practice and therapeutic trials (see Table 1): the UHDRS Total Motor Score (TMS), five cognitive assessments (the Symbol Digit Modalities Test (SDMT), the Verbal Fluency test 1-minute (VF), and the three components of the Stroop test (word (SW); colour (SC); interference (SI))), and two functional scales (the Total Functional Capacity (TFC) and the Independence scale (UHDRS IS)).

All the demographics and the summary of clinical scores are displayed in Table 1.

2.3. Data collection

Interviewers were all neuropsychologists. Participants completed a standardised battery of speech tasks at the hospital. They were asked to narrate both emotional and neutral stories. The stories were prompted by the interviewers asking standardised questions (ex:"May you tell me a story that you find sad, really unhappy, or really depressing". In most cases, participants spoke for less than 1 minute, therefore interviewers were instructed to prompt individuals with pre-defined instructions such as "For instance, you may tell me about an incident heard on the news, a movie, something you saw on television?"). It was clearly stated that they should trigger non personal stories as much as possible. Task 1 (neutral) consisted in describing the latest 24 hours and tasks 2, 3, and 4 were designed to elicit emotions by telling a story making the participants sad, angry, and happy, respectively. The speech tasks were separated by non-emotional speech (Automatic recitation of months of the year, the Cookie Theft description, and the storytelling of the Little Red Riding Hood) [31] not assessed here. Tasks' order was fixed. All sessions ended with the recall of the happy story, Task 4, to avoid ending the experiment with a negative emotion. The whole session lasted less than 15 minutes (14.5 ± 6.1 minutes on average). Patients could interrupt the session anytime. Participants were recorded in similar acoustic conditions, with a ZOOM H4n Pro recorder, sampled at 44.1 kHz with a 16-bit resolution.

2.4. Samples preprocessing

Speech therapists excluded speech samples with too high ambient noise precluding any analysis (two files discarded). Then, they transcribed the language content in text and split by *stretch* the stream of continuous speech from each task in stretches using the software Praat. Annotations were managed with the Seshat platform [55]. The annotation of a single interview lasted approximately 8 hours, therefore to ensure the quality of the annotation, we randomly selected five interviews and asked two speech pathologists to annotate them independently. Inter-annotator agreements computed with the Gamma Agreements , denoted γ, [30, 54] for stretch limits and turn-takings were high for the 4 tasks: Neutral ($\gamma = 78.0\% \pm 6.9$), Sad ($\gamma = 84.3\% \pm 7.9$), Angry ($\gamma = 77.6\% \pm 4.9$), Happy ($\gamma = 66.0\% \pm 19.2$). This allowed to annote the remaining 110 interviews by a single speech therapist. Each stretch was labeled with the emotion corresponding to the task it was uttered into (Task 1: neutral, Task 2: sad, Task 3: angry, and Task 4: happy). They were dispatched in data sets according to their group (HD, pre HD, and controls) and the modality (voice and language). We balanced the datasets classes to avoid confounding the performance of the machine learning models with the quantity of training data. To do so, stretches were removed randomly from the dataset to ensure the same number of stretches and same number of emotions for the voice and the text. This

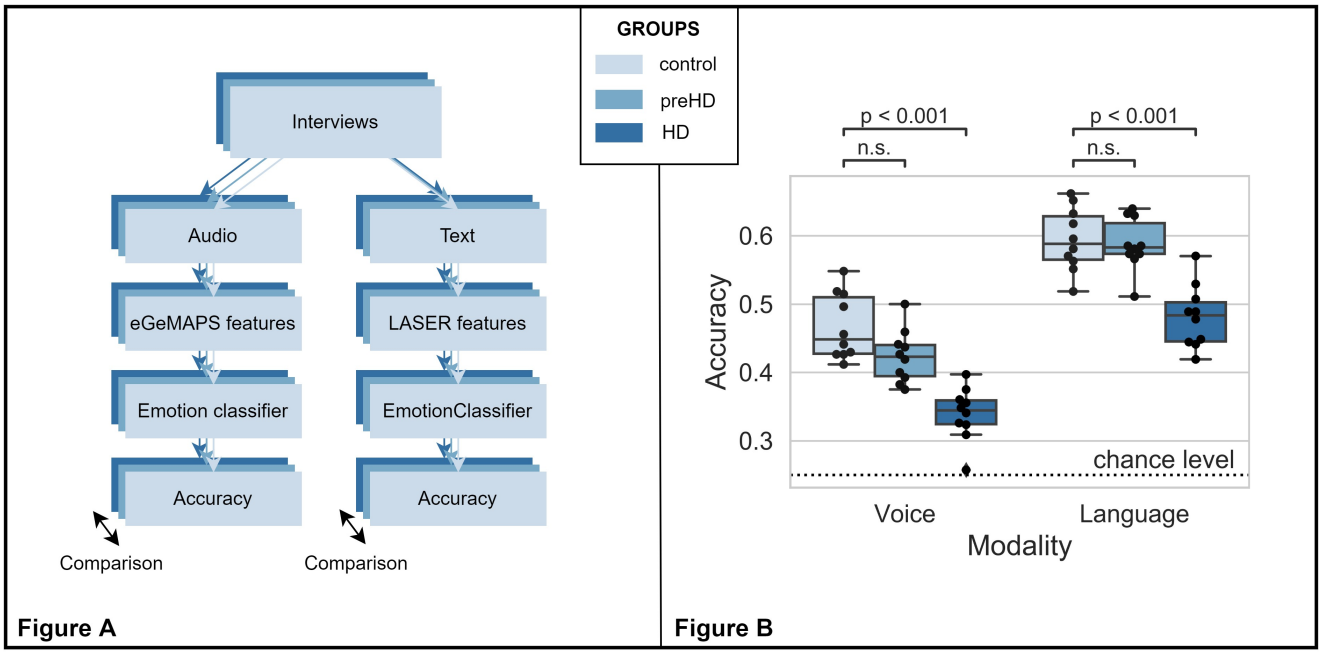


Figure 2: Voice and language experiments.
A. Method's flowchart. Voice and linguistic content (language) of the interviews are separated to create six datasets (3 groups x 2 modalities). Features are extracted from the data and one emotion classifier is trained and tested on each dataset. Emotion classifiers' accuracy is then compared across groups within each modality.
B. Results of accuracy comparison within each modality across the three groups. The HD group accuracy is significantly lower than controls, but this is not the case for the preHD group.

yielded 1356 stretches for each group, both for the voice and the language modalities.

We then extracted relevant affective information in a fixed-size vector from the audio stretches and linguistic content to be able to automatically classify emotion. We denoted this fixed-size vector the 'features'. For the audio stretches, we selected a minimal set of features, the Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [10] designed by interdisciplinary voice and speech scientists to provide a relevant set of features for affective computing. We chose these features due to their high performance to classify emotions in several voice dataset from different cultures [10]. The features are related to energy (e.g. harmonic to noise ratio), to pitch (e.g. fundamental frequency statistics), and articulation (e.g. formant statistics). We extracted these voice features with the openSMILE toolkit [11]. For the text modality, we used the features from the LASER (Language-Agnostic SEntence Representations) sentence embedding model [5]; LASER being a language-agnostic model which transforms a sentence of words of arbitrary length into a fixed-size vector. It was designed to perform well in a variety of natural language processing tasks in more than 90 languages, especially when the data were scarce to train the models. LASER obtained excellent performance in multi-lingual document classification or natural language inference (find relationships of entailments between sentences). Then, emotion classifiers were developed for each data set, trained and tested on the datasets' features. Emotion classifier algorithms were built with random forests, implemented in scikit-learn [39]. To train and test emotion classifiers on our datasets we repeated a 10-fold nested cross validation scheme proceeding as follows. We split our data within each dataset in 10 folds. Nine were used for training and then the remaining one as a test set. After 10 permutations between training sets and test sets, we obtained the assessment of the whole data set with a measure of accuracy for each test set, yielding 10 measures of accuracy for each of the six models. The accuracy of an emotion classifier on a set of stimuli (audio or text) labelled with emotions is computed as the percentage of correct predictions (predicted label equals actual label) on this set (see Figure 2 B). In order to assess whether the impairment in emotion production through voice in HD is restricted to a specific step during speech production, we ran our audio models with subsets of the features related to energy (Harmonic to noise ratio features, alpha ratio statistic, hammarberg index statistics), pitch (F0 statistics, jitter features, shimmer features), or articulation (MFCC features, formants statistics, spectral flux features). We selected these 3 different sets of features as they represent the main steps to produce fluent articulated speech: respiratory (energy), phonatory (pitch) and articulatory.

2.5. Statistical analysis

To assess the identifiability of emotion expression through voice and language, we compared the accuracy for the two modalities across the three participants 'groups (controls, preHD, and HD). We first used a Kruskal-Wallis test on the

three series of accuracies (controls, preHD, and HD), and when it led to significant difference between groups within each modality (rejection of the null hypothesis), we conducted post-hoc Wilcoxon Mann Whitney tests to compare accuracies group by group. The tests were one-tailed, and Bonferroni correction was applied. Because of the structure of the data, traditional tests could not be applied as such, and we resorted to resampling to perform the tests (see Appendix 1. for details).

2.6. Comparison between machine and human classification

To check whether our method was consistent, 12 healthy scorers with no hearing impairments (31.4 years by mean (10.5)), half females, half males) were asked to classify the voice and the text (language) modalities assessed by the machine learning algorithms. They were all French native speakers. For this purpose, annotations were used as the linguistic content and were provided as stretches of text to the scorer. The audio stretches were filtered (we used a Butterworth low pass filter of 4th order, with a cut-off frequency of 250Hz) in order to remove intelligibility of words and keep prosody. The 1356 text stretches and the 1356 filtered voice stretches were randomised and divided in six scoring sets for each modality, yielding twelve sets of scoring. Each scorer oversaw one audio scoring set and one text scoring set. Half of them scored first language and then voice (G1), the other half performed the scoring in a reverse order (G2). Text scoring lasted around 15-30 minutes whereas language scoring for a set lasted between 75 -90 minutes for each scorer. We measured agreement between human groups and our model on stretches that were correctly labelled by each individual group. We compared it with agreement between Group 1 (respectively Group 2) on stretches correctly labeled by Group 1 (respectively Group 2). Agreement was measured with the kappa score.

3. Results

Accuracies for each group and each modality are reported in Figure 2B. We found that our models' accuracies were significantly lower for HD patients than for controls and preHD for both modalities (corrected p-value <0.001 for each statistical test). This shows that HD patients are impaired in expression of emotions through voice and language compared to controls. All types of emotions are impaired across at least one modality (see Figure 3). Detailed results for each emotion are given in Appendix 2. Accuracies for the preHD group were similar to that of controls (voice: corrected p-value=0.14, language: corrected p-value >0.5). Hence, preHD participants show no sign of impairment in emotion expression through voice and language compared to controls.

The accuracy in classifying emotion was also significantly lower in HD than in controls when restraining the features to energy, pitch or articulation. In contrast, accuracy in classifying emotion for preHD did not differ from those of controls in any modality (Figure 4).

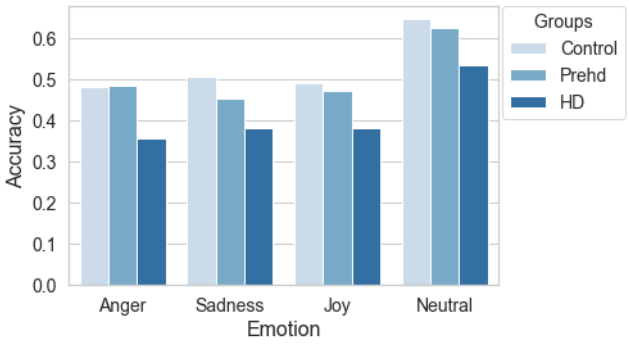


Figure 3: Mean of the voice and language accuracy for each emotion and each group. This figure provides a view per emotion, provided through the different splits and the different modalities (Voice and language). Yet, we took the accuracy averaged between the different splits and the different modalities thus precluding a direct statistical comparison but giving a hint on the lowering of all expressions through speech.

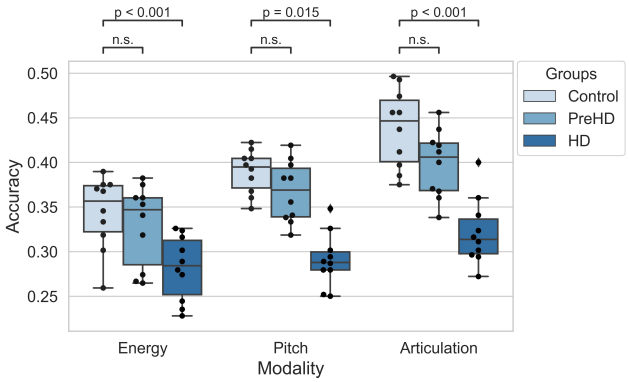


Figure 4: Accuracies obtained for each group when restraining the audio features to energy, pitch, or articulation features.

The comparison of our method to a human jury highlights its validity (See Figure 5). The accuracy of the human juries was lower than the machine ones for both modalities (audio: 31.5% on average for human groups vs 48% for our model; text: 48% on average for human groups vs 60% for our model). We also found that our model and the two human scorers groups correctly classify the same stretches for both modalities (audio: kappa score= 0.31 on average between human groups (G1 vs G2) and 0.39 on average between each human group and our model; text: 0.79 on average between the two human groups (G1 vs G2) and 0.7 on average between each human group and our model).

4. Discussion

Here, we assessed emotional expression through spoken language of individuals carrying the mutant Htt gene leading to HD in comparison with control participants. As speech combines both the voice (including prosody and vocal signals) and the linguistic content (language), we developed a machine learning method to disentangle emotional expression through voice and through language in a blind and ef-

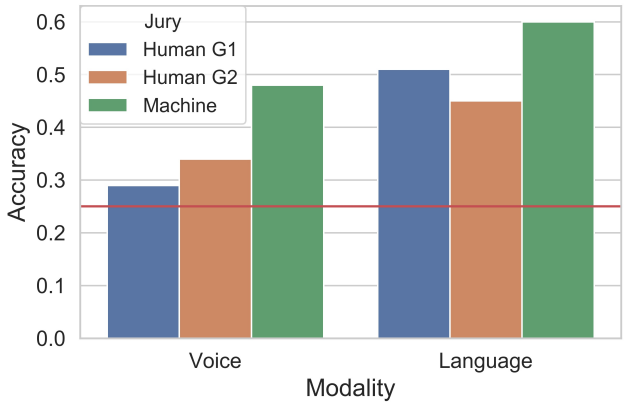


Figure 5: Accuracies Human/Machine on a subset of speech segments produced by Healthy Controls.

ficient manner. Elicited emotions were classified by emotion classifiers, acting as expert juries. Our models classified stretches of spoken language better than chance for preHD and HD patients, both for voice and language based-models. We found that HD participants have reduced expression of emotions both through voice and language compared to controls and preHD participants. Our machine learning models performed better than human scorers exposed to similar stretches. By studying voice and language separately, we added the missing evidence about the spoken language production of emotion. These results enrich the theories of emotional processing in HD. They also show how machine learning models can be leveraged to study emotion expression.

4.1. Automated evaluation of emotion expression

Speech algorithms are currently being developed to detect and classify emotions from spoken language [63, 1, 41] and more specifically for individuals affected with neurological disorders [46, 3]. However, the development of these models have been purely narrowed to an engineering point of view; 'only' trying to maximise the accuracy. Yet, these models offer a remarkable opportunity to compare the expressiveness of emotions in spoken language. To the best of our knowledge, machine learning models have never been used to quantify the identifiability of emotions in spoken language of individuals with neurological disorders.

Traditional methods seeking to compare the expression of emotions between patients and controls use either a statistical comparison of pre-defined spoken language markers [2] or a human jury [19, 57]. These methods preclude analysing variation of everyday emotions beyond stereotypical constrained intense emotions in static set-ups [61]. Our methods allow assessing naturally triggered emotions [34, 4]. They are closer to experiments using a human jury in such that they focus on comparing the identifiability of emotions across different populations. In addition, our methods present three advantages compared to a human jury. First, it is not sensitive to the impact of the disease on the motor aspects of speech, since one model is trained on each population. If the emotions are still expressed despite the motor problems, the models will be able to identify it. This is not the case with a human jury which can be biased by differences across participants unrelated to the expression of emotions. This was particularly sensitive in our study since Patients with Huntington's disease suffer from dysarthria, inappropriate pauses or breathing, and temporal irregularities of speech [46, 40, 43] independently of emotional expression. Additionally, machine Learning models allow selective testing of various dimensions that are intertwined in the speech signal, for instance through the choice of features. This enabled us not only to study language and voice separately, but also to compare emotion expression through different speech effectors (energy, pitch, and articulation; see Annex). Finally, using a standardised method makes it easier to reproduce the analysis, to compare it with new studies; it does not require organising a human experiment with a jury, which is challenging and time consuming. The use of human juries also brings other problems such as huge variability when the sample size is low. Thus, this method could even be adapted to study emotion expression through other modalities such as body and gestures by feeding the model pictures as input to machine learning algorithms.

This shows that machine learning models are great additions to classic approaches using human juries. This opens a room for more straightforward and reproducible research in the study of the production of emotions.

4.2. Embodied cognition & spoken language

Previous studies attempted to explain the reported deficits in expression and perception of emotions in HD with the theoretical framework of embodied cognition [56, 26]. In the embodied cognition theory, high-level cognitive processes use reactivation of sensory and motor systems. Applied to emotions, it suggests that perceiving others' emotional states involves sensory-motor reexperiencing as opposed to a mere mobilization of abstract conceptual representations of emotions [36]. Evidence from studies with neurotypical populations support this view. For instance, the dampening or amplification of facial feedback modulates the accuracy to perceive emotions [35]. The hypothesis made in previous studies is that in HD, the deterioration of motor processing components alters sensori-motor representation of emotions, causing the observed impairments in expression and recognition of emotion. Reported results provide credibility to this argument because they are consistent with three important predictions of this hypothesis.

First, this hypothesis states that expression and recognition mechanisms are caused by the same underlying mechanism. Thus, we shall observe joint deficits in perception and production of emotions, and these deficits should extend to all emotions. Results obtained on expression and recognition of facial emotion in HD are in line with this prediction. Hayes et al. [19] showed a coupled expression-recognition deficit for facial disgust in HD. Robotham et al. [44] by equalising the number of positive and negative stimuli, showed that both positive and negative emotions were

impaired similarly. Trinkler et al. [57] and Trinkler et al. [56] extended these results and found that facial emotion expression deficits were correlated with impairments in emotion recognition of facial emotions in HD and that these impairments extend to all emotions.

Second, this hypothesis predicts that emotional impairments are tied to motoric impairments rather than dysfunctions in internal experience of emotions or cognitive disorders. Results reported in the literature also support this second prediction. HD patients scored as controls on the alexithymia test [56] suggesting an intact internal experience of emotions. Additionally, evidence for an intact conceptual understanding of emotions in HD was reported in several studies [26]. Furthermore, Yitzhak et al. [61] showed that amongst 4 factors (cognitive screening, motor symptoms, depressive symptoms, and the estimated progression of HD pathology) only motor symptoms were correlated with HD patients' performances in a facial expression recognition task.

Third, following this hypothesis, patients suffering from other movement disorders should exhibit joint deficits in expression and recognition of emotions. Accordingly, it has been shown that in Parkinson disease [32] or in multiple sclerosis [42, 21] similar expression/recognition deficits are observed. Consistently, they extend to several emotions and modalities.

Our results on emotion expression through voice provide additional evidence to support the hypothesis of altered sensori-motor representations of emotions in HD. The hypothesis predicts that 1) expression of emotions through motor functions is impaired in HD, and thus expression of emotions through voice should be impaired 2) expression and recognition deficits are joint impairments, and as recognition of emotions through voice is impaired in HD, expression of emotions through voice should also be impaired. Consistently, we found that expression of emotions through voice is impaired in HD. Additionally, we found that this impairment does not seem to be effector (energy, pitch, and articulation) specific.

More difficult to integrate is the preserved perception of emotional language reported in previous studies [18, 57]. Following our finding that the expression of emotions through language is impaired, it derogates from the joint perception /production impairment predicted by embodied cognition theory. Nevertheless, as Winkielman et al. [60] points out, when the tasks do not require emotional implications, the conceptual channel could remain functional without simulating emotions. HD patients could be impaired in the perception of emotional language but could compensate for their deficit with intact conceptual abilities [26]. The two hypotheses are not exclusive: although embodied representations and conceptual representations rely on two different networks, their activation is not competitive but complementary. They depend on the context and the purpose of the task [60]. For example, when asked to list properties associated with emotional concepts (e.g., frustration), participants' facial muscles are activated more in an emotional context (they are asked to respond as they would to a good friend) than in a formal context (they are asked to respond as they would to a supervisor) [37, 60]. Coupled with the interaction between conceptual or embodied language and emotions, this could explain the differences with voice processing.

Another possibility is that the incongruence between perception and production of emotional language is not dependent on emotional processing. The tasks used in perception consisted of classifying either emotional words or emotional stories [18, 57]. Although in some cases emotional words may induce emotion, the context of the tasks may not activate emotions. Therefore it may not be comparable to our production tasks based on the elicitation of emotions with emotional stories. Alternatively, the emotional language impairment could be based on production deficits (decreased fluency, syntactic and semantic deficits in Huntington's disease patients), which are more marked in production than in perception [28]. The production deficit would therefore be independent of the emotional processing deficit.

The processing of action words has been the subject of a similar debate. Patients with Parkinson's and Huntington's disease have difficulties with syntax, verb perception and production, and the coupling between motor and action [6]. These results are consistent with both the theory of embodied cognition and what Mahon and Caramazza [29] have called the disembodied cognition hypothesis. Counterexamples such as the retention of nouns but the impossible use of tools in apraxic patients or abstract words such as "freedom" that do not induce either simulation or action would confine the embodiment theory to emotion or action words and could not be applied to language in general. This led to an intermediate theory of embodied cognition, the "grounding through interaction", in which conceptual representations can be maintained without motor activations, but sensory and motor systems complement and enrich abstract and symbolic representations [29]. The instantiation of a concept would involve the retrieval of specific sensory and motor information. The "suppression" of sensory and motor systems (as in the case of brain damage) would result in impoverished or "isolated" concepts. From this point of view, sensory and motor information contribute to the "complete" representation of a concept. The activation of sensory and motor processes during conceptual processing is not necessarily "incidental" or "irrelevant" to conceptual processing. The activation of specific sensory and motor representations complements the generality and flexibility of "abstract" and "symbolic" conceptual representations. However, this theory was dedicated to action words and not to emotion words and the entanglement between emotion theories and language processing remains to be clarified.

As our methodology tested voice and language in an equivalent way on the basis of elicited emotions, the impairment of emotion production by language and voice presumably reflects a deficit of emotion expression in HD patients. Future models of emotional processing may incorporate these results.

4.3. Clinical perspectives

HD patients have difficulties in expressing their emotions through both voice and language may partly account for the disrupted communication between patients and their caregivers [22]. This difficulty in sharing emotions might create misunderstanding and frustration. This contradicts our previous view that expression of emotions through spoken language was preserved [56] and could compensate impaired motor expression of emotions. Our results suggest that this might not be the case. Awareness of HD patients' emotion expression impairments should be raised amongst patients and caregivers to improve their interactions. Second, our results could also be of use for the design of a smart device to monitor Huntington disease at home. Currently, the disease's follow-up requires regular consultations at the hospital with several neurological, psychiatric, and cognitive tests, which implies both a financial and human cost. Even though HD patients are impaired in emotion expression through spoken language, it is possible to classify emotions in their speech above chance. Thus, an automatic emotion classifier for HD patients could help track their mood and alert caregivers when psychiatric syndromes such as depression, irritability or apathy start developing. They can be hard to spot on a day-to-day basis, and patients and caregivers would benefit from early medical care for these symptoms. In addition, methods are being developed to derive patients' clinical scores from speech[43, 40, 45]. Eventually, these methods could be used to monitor HD at home through the evolution of clinical scores. The identifiability of emotions for each HD individual may constitute an interesting clinical endpoint to monitor as well. It may also potentially be included in clinical trials in order to spare the capabilities of HD patients' social interactions of HD's individuals rather than on, and not only attenuate motor symptoms.

4.4. Limitations

Our study presents some limitations that could be overcome in future works. While being in range with the literature, the number of participants and the amount of data for each participant was limited in our study because of the complexity of retrieving speech data in controlled experiments. Hence, we could not conduct fine grained analysis such as correlations between the identifiability of emotions and motor and cognitive score for each individuals.

5. Conclusions

Machine learning allowed us to disentangle voice and language in overt speech in HD. The impact of HD on emotion perception can be potentially translated in models of striatal lesions and tested empirically. This represents an important line of subsequent work, for the specifications of the embodied cognition theory in a more quantitative manner, with direct implications for a complete understanding of emotions in HD. Models might apply to other conditions and emotions expressed by bodies or faces, for example.

6. Acknowledgements

We are very thankful to the participants that took part in our study. We also thank the speech pathologists for the annotations of the speech data. In addition, we thank Laurent Cléret de Langavant, Nicolas Fraisse, Tiffany Monnier, Amin Gharbi, Graça Morgado for their help to carry out this project and helpful discussions to improve this work. This work is funded in part by the Agence Nationale pour la Recherche (ANR-17-EURE-0017 Frontcog, ANR-10-IDEX-0001-02 PSL*, ANR-19-P3IA-0001 PRAIRIE 3IA Institute) and Grants from Neuratris, from Facebook AI Research (Research Gift), Google (Faculty Research Award), Microsoft Research (Azure Credits and Grant), and Amazon Web Service (AWS Research Credits).

7. Bibliography

References

[1] Akçay, M.B., Oğuz, K., 2020. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication 116, 56–76. URL: https://www.sciencedirect.com/science/article/pii/S0167639319302262, doi:10.1016/j.specom.2019.12.001.

[2] Alhinti, L., Christensen, H., Cunningham, S., 2021. Acoustic differences in emotional speech of people with dysarthria. Speech Communication 126, 44–60. URL: https://www.sciencedirect.com/science/article/pii/S016763932030296X, doi:10.1016/j.specom.2020.11.005.

[3] Alhinti, L., Cunningham, S., Christensen, H., 2020. Recognising Emotions in Dysarthric Speech Using Typical Speech Data, in: Interspeech 2020, ISCA. pp. 4821–4825. URL: https://www.isca-speech.org/archive/interspeech_2020/alhinti20_interspeech.html, doi:10.21437/Interspeech.2020-1825.

[4] Amodio, D.M., Zinner, L.R., Harmon-Jones, E., 2007. Social psychological methods of emotion elicitation. Handbook of emotion elicitation and assessment 91, 91–105.

[5] Artetxe, M., Schwenk, H., 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. arXiv:1812.10464 [cs] URL: http://arxiv.org/abs/1812.10464. arXiv: 1812.10464.

[6] Birba, A., García-Cordero, I., Kozono, G., Legaz, A., Ibáñez, A., Sedeño, L., García, A.M., 2017. Losing ground: Frontostriatal atrophy disrupts language embodiment in Parkinson's and Huntington's disease. Neuroscience & Biobehavioral Reviews 80, 673–687. URL: https://www.sciencedirect.com/science/article/pii/S0149763417300702, doi:10.1016/j.neubiorev.2017.07.011.

[7] Bora, E., Velakoulis, D., Walterfang, M., 2016. Social cognition in Huntington's disease: A meta-analysis. Behavioural Brain Research SreeTestContent1 297, 131–140. URL: https://www.sciencedirect.com/science/article/pii/S0166432815302151, doi:10.1016/j.bbr.2015.10.001.

[8] Chan, J.C., Stout, J.C., Vogel, A.P., 2019. Speech in prodromal and symptomatic huntington's disease as a model of measuring onset and progression in dominantly inherited neurodegenerative diseases. Neuroscience & Biobehavioral Reviews 107, 450–460.

[9] Efron, B., Hastie, T., 2016. Computer age statistical inference. volume 5. Cambridge University Press.

[10] Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., Andre, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S., Truong, K.P., 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Transactions on Affective Computing 7, 190–202. URL: http://ieeexplore.ieee.org/document/7160715/, doi:10.1109/TAFFC.2015.2457417.

[11] Eyben, F., Wöllmer, M., Schuller, B., 2010. openSMILE – The Mu-

nich Versatile and Fast Open-Source Audio Feature Extractor, pp. 1459–1462. doi:10.1145/1873951.1874246.

[12] Frick, R.W., 1985. Communicating emotion: The role of prosodic features. Psychological Bulletin 97, 412–429. doi:10.1037/0033-2909.97.3.412. place: US Publisher: American Psychological Association.

[13] Friederici, A.D., 2017. Language in Our Brain: The Origins of a Uniquely Human Capacity. MIT Press.

[14] de Gelder, B., Van den Stock, J., Balaguer, R.d.D., Bachoud-Lévi, A.C., 2008. Huntington's disease impairs recognition of angry and instrumental body language. Neuropsychologia 46, 369–373. doi:10.1016/j.neuropsychologia.2007.10.015.

[15] Guenther, F.H., 2016. Neural Control of Speech. MIT Press. Google-Books-ID: aRSvDAAAQBAJ.

[16] Hamilton, J.M., 2003. Behavioural abnormalities contribute to functional decline in Huntington's disease. Journal of Neurology, Neurosurgery & Psychiatry 74, 120–122. URL: https://jnnp.bmj.com/lookup/doi/10.1136/jnnp.74.1.120, doi:10.1136/jnnp.74.1.120.

[17] Hartelius, L., Jonsson, M., Rickeberg, A., Laakso, K., 2010. Communication and Huntington's disease: qualitative interviews and focus groups with persons with Huntington's disease, family members, and carers. International Journal of Language & Communication Disorders 45, 381–393. doi:10.3109/13682820903105145.

[18] Hayes, C.J., Stevenson, R.J., Coltheart, M., 2007. Disgust and Huntington's disease. Neuropsychologia 45, 1135–1151. URL: https://www.sciencedirect.com/science/article/pii/S002839320600412X, doi:10.1016/j.neuropsychologia.2006.10.015.

[19] Hayes, C.J., Stevenson, R.J., Coltheart, M., 2009a. Production of spontaneous and posed facial expressions in patients with Huntington's disease: Impaired communication of disgust. Cognition and Emotion 23, 118–134. URL: https://doi.org/10.1080/02699930801949090, doi:10.1080/02699930801949090. publisher: Routledge _eprint: https://doi.org/10.1080/02699930801949090.

[20] Hayes, C.J., Stevenson, R.J., Coltheart, M., 2009b. Production of spontaneous and posed facial expressions in patients with huntington's disease: Impaired communication of disgust. Cognition and Emotion 23, 118–134.

[21] Henry, J.D., Phillips, L.H., Beatty, W.W., Mcdonald, S., Longley, W.A., Joscelyne, A., Rendell, P.G., 2009. Evidence for deficits in facial affect recognition and theory of mind in multiple sclerosis. Journal of the International Neuropsychological Society 15, 277–285. doi:10.1017/S1355617709090195. publisher: Cambridge University Press.

[22] Ho, A.K., Hocaoglu, M., of Life Working Group, E.H.D.N.Q., 2011. Impact of huntington's across the entire disease spectrum: the phases and stages of disease from the patient perspective. Clinical Genetics 80, 235–239.

[23] Jacquemot, C., Bachoud-Lévi, A.C., 2021. Striatum and language processing: Where do we stand? Cognition , 104785.

[24] Johnson, S.A., Stout, J.C., Solomon, A.C., Langbehn, D.R., Aylward, E.H., Cruce, C.B., Ross, C.A., Nance, M., Kayson, E., Julian-Baros, E., Hayden, M.R., Kieburtz, K., Guttman, M., Oakes, D., Shoulson, I., Beglinger, L., Duff, K., Penziner, E., Paulsen, J.S., the Predict-HD Investigators of the Huntington Study Group, 2007. Beyond disgust: impaired recognition of negative emotions prior to diagnosis in Huntington's disease. Brain 130, 1732–1744. URL: https://doi.org/10.1093/brain/awm107, doi:10.1093/brain/awm107.

[25] Jona, C.M.H., Labuschagne, I., Mercieca, E.C., Fisher, F., Gluyas, C., Stout, J.C., Andrews, S.C., 2017. Families Affected by Huntington's Disease Report Difficulties in Communication, Emotional Involvement, and Problem Solving. Journal of Huntington's Disease 6, 169–177. URL: https://content.iospress.com/articles/journal-of-huntingtons-disease/jhd170250, doi:10.3233/JHD-170250. publisher: IOS Press.

[26] Kordsachia, C., Labuschagne, I., Stout, J., 2017. Beyond emotion recognition deficits: A theory guided analysis of emotion processing in Huntington's disease. undefined URL: /paper/Beyond-emotion-recognition-deficits%3A-A-theory-of-in-Kordsachia-Labuschagne/860bd5605e7abce1788e27150f264307b6b997bb.

[27] Kordsachia, C.C., Labuschagne, I., Andrews, S.C., Stout, J.C., 2018. Diminished facial EMG responses to disgusting scenes and happy and fearful faces in Huntington's disease. Cortex 106, 185–199. URL: https://www.sciencedirect.com/science/article/pii/S0010945218301758, doi:10.1016/j.cortex.2018.05.019.

[28] Ludlow, C.L., Connor, N.P., Bassich, C.J., 1987. Speech timing in parkinson's and huntington's disease. Brain and language 32, 195–214.

[29] Mahon, B.Z., Caramazza, A., 2008. A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. Journal of Physiology-Paris 102, 59–70. URL: https://linkinghub.elsevier.com/retrieve/pii/S0928425708000193, doi:10.1016/j.jphysparis.2008.03.004.

[30] Mathet, Y., Widlöcher, A., Métivier, J.P., 2015. The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. Computational Linguistics 41, 437–479.

[31] McNally, R.J., Litz, B.T., Prassas, A., Shin, L.M., Weathers, F.W., 1994. Emotional priming of autobiographical memory in post-traumatic stress disorder. Cognition & Emotion 8, 351–367.

[32] Mermillod, M., Vermeulen, N., Droit-Volet, S., Jalenques, I., Durif, F., Niedenthal, P., 2011. Embodying Emotional Disorders: New Hypotheses about Possible Emotional Consequences of Motor Disorders in Parkinson's Disease and Tourette's Syndrome. ISRN Neurology 2011, 1–6. URL: https://www.hindawi.com/journals/isrn/2011/306918/, doi:10.5402/2011/306918.

[33] Möbes, J., Joppich, G., Stiebritz, F., Dengler, R., Schröder, C., 2008. Emotional speech in Parkinson's disease. Movement Disorders 23, 824–829. doi:https://doi.org/10.1002/mds.21940.

[34] Nazareth, D.S., Tournier, E., Leimkötter, S., Janse, E., Heylen, D., Westerhof, G.J., Truong, K.P., 2019. An Acoustic and Lexical Analysis of Emotional Valence in Spontaneous Speech: Autobiographical Memory Recall in Older Adults, in: Proc. Interspeech 2019, pp. 3287–3291. URL: http://dx.doi.org/10.21437/Interspeech.2019-1823, doi:10.21437/Interspeech.2019-1823.

[35] Neal, D.T., Chartrand, T.L., 2011. Embodied Emotion Perception: Amplifying and Dampening Facial Feedback Modulates Emotion Perception Accuracy. Social Psychological and Personality Science 2, 673–678. URL: http://journals.sagepub.com/doi/10.1177/1948550611406138, doi:10.1177/1948550611406138.

[36] Niedenthal, P., 2007. Embodying Emotion. Science doi:10.1126/science.1136930.

[37] Niedenthal, P.M., Winkielman, P., Mondillon, L., Vermeulen, N., 2009. Embodiment of emotion concepts. Journal of Personality and Social Psychology 96, 1120–1136. doi:10.1037/a0015574. place: US Publisher: American Psychological Association.

[38] Paeschke, A., Kienast, M., Sendlmeier, W.F., . F0-CONTOURS IN EMOTIONAL SPEECH , 4.

[39] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., 2011. Scikit-learn: Machine learning in python. the Journal of machine Learning research 12, 2825–2830.

[40] Perez, M., Jin, W., Le, D., Carlozzi, N., Dayalu, P., Roberts, A., Provost, E.M., 2018. Classification of huntington disease using acoustic and lexical features, in: Interspeech, NIH Public Access. p. 1898.

[41] Picard, R.W., 2000. Affective computing. MIT press.

[42] Pöttgen, J., Dziobek, I., Reh, S., Heesen, C., Gold, S.M., 2013. Impaired social cognition in multiple sclerosis. Journal of Neurology, Neurosurgery & Psychiatry 84, 523–528. URL: https://jnnp.bmj.com/content/84/5/523, doi:10.1136/jnnp-2012-304157. publisher: BMJ Publishing Group Ltd Section: Multiple sclerosis.

[43] Riad, R., Titeux, H., Lemoine, L., Montillot, J., Bagnou, J.H., Cao, X.N., Dupoux, E., Bachoud-Lévi, A.C., 2020. Vocal markers from sustained phonation in Huntington's Disease. arXiv:2006.05365 [cs, eess] URL: http://arxiv.org/abs/2006.05365. arXiv: 2006.05365.

[44] Robotham, L., Sauter, D.A., Bachoud-Lévi, A.C., Trinkler, I., 2011. The impairment of emotion recognition in Huntington's disease extends to positive emotions. Cortex 47, 880–

5 General discussion

> *The history of science teaches us that, to be fruitful, a discipline must interfere with others both at the level of concepts and at the level of techniques. Science cannot be reduced to a single discourse and must be interdisciplinary. How then to understand an integrated system like that of the dream starting from the only molecular biology.*
>
> **— Michel Jouvet**
> French neuroscientist
> who discovered the paradoxical sleep.

This manuscript investigated speech and linguistic markers as potential cheap and non-invasive biomarkers in HD. In this part, we first summarize the main contributions of this thesis. Then, we discuss the effect of speech tasks and the types of spoken language markers that enables us to obtain good predictive performance. Finally, we overview the different limitations of our studies, and discuss potential future lines of research.

5.1 Spoken language as a window into HD

In this thesis, we developed a speech protocol that can be easily conducted by neuropsychologists to probe different spoken language capabilities of individuals with HD. We developed and open-sourced software technologies that allowed us to obtain a rich annotated resource, the BasalVoice database. To the best of our knowledge, it represents the largest available resource on the spoken language production in HD covering all stages of the disease (See Chapter 2). We developed novel ways to annotate spontaneous speech, taking into account the specifics of the production of HD. We made sure that our protocol fulfill the requirements for automatic analyses by computer programs. In addition, we proposed a novel formalization

and implementation of the psycholinguistic theory of tracks of communication of Herbert Clark (Clark, 1996). These tracks specify the roles of different vocalizations within an utterance. We provided specific application of this formalization on the BasalVoice database.

Thanks to this new database of interviews at the hospital, we investigated the possibility to automatize some aspects of annotations, identifications of turn-takings and tracks of communication directly from the waveform (See Chapter 3). We applied and adapted deep learning methods for our applications. We discovered that a system that identify the role of the speakers (interviewer/interviewee) worked better than speaker enrollment methods. We also found out that the size of the training set was a huge contributing factor for the performance of the speaker role recognition system, underlying the need to scale the database effort. We also started to extract spoken language markers from these automatic pipelines. We showed that we need to take into account these language markers during the development of automatic systems. In addition, we confirmed that it was harder to identify tracks of communication (disfluencies) from speech waveform in stuttered speech than from textual transcription. We extended this framework to our database. We tackled, for the first time, both turn-taking identification and identification of tracks of communication tasks at the same time, based only on the waveform.

Another contribution of this thesis concerns the replication, validation and extension of spoken language markers as biomarkers in HD. We assessed the predictive potential of different speech features from different speech tasks (See Chapter 4). We found out that it was possible to distinguish HD from HC and preHD participants based on simple sustained phonation of the vowel /a/. However, preHD and HC participants were harder to separate at the group and individual level. These markers from the vowel /a/ have some potential to reflect the main clinical scores in preHD and HD populations. Preliminary analyses on vowel distortion measures extracted from sustained phonation showed that we could not distinguish HC, preHD and HD participants at the group level. Yet, we found that these vowel distortion measures from open-vocabulary tasks present significant difference between preHD and HD participants for the VAI measure. Then, we combined features from two brief closed-vocabulary tasks, to obtain speech from minimal cognitive load (counting aloud numbers) and with higher cognitive load (backward counting aloud numbers while holding hands and closing eyes). When we combined speech features from these tasks with demographics variables, we obtained the current best performance to predict the main clinical measures in HD, for the functional, motor, and cognitive, and composite cUHDRS. Our methods obtained the best performance on the cUHDRS, the current clinical endpoint in clinical trials. This approach was even

further validated thanks to the separation of exploration and validation cohorts, with pre-registration methodologies and correlations with striatal volumes. Thanks to our new parsing analysis of spontaneous and dysarthric speech we made the first exploratory analysis concerning the speech phenomena in HD's speech and the tracks of communication. We observed that the ratio of the primary track decreases as the disease is declared, the ratio of incidental elements is greatly increased, and the ratio of secondary presentation is greatly decreased.

We also developed new machine-learning based methods to assess the emotional speech production deficits in HD. This allowed us to extend the current description of emotional processes in HD, and we observed that both vocal (energy, pitch and articulation) and linguistic modalities were affected. To quantify these differences, we used machine learning models and the same amount of training data for the three groups under study. It was harder to distinguish emotions in spoken language produced by HD than to distinguish emotions spoken language produced by HC. This effect was not found when the comparison was made between preHD and HC.

Our different contributions underline the importance of (1) cognitive load to obtain informative markers, (2) and timing and articulatory information to obtain strong predictive performance, underlying even more the need for robust speech algorithms. These different aspects can be examined from the different perspectives we mentioned during the construction of the BasalVoice database in Section 2: Clinical Neurology, Neurolinguistics, and Speech technology.

Concerning the cognitive load, Vogel et al. (2012) found out that spoken language markers in preHD and HD correlated with TMS only when reading or during in a monologue. In our study (Riad, Lunven, et al., 2022, p. 15), we also observed that we obtain better predictive performance when using spoken language markers extracted from counting backward than from counting forward. Besides, we also found that vowel distortion measures separated HD only in open-vocabulary tasks. This effect of the cognitive load on the task has also been found in another sub-cortical disease, PD, where spontaneous speaking tasks were the most informative on the disease (Rusz, Cmejla, et al., 2013; Kempler & Van Lancker, 2002). This underlines the critical role of sub-cortical parts of the brain to handle and support cognitive load. Furthermore, the need for cognitive load also has implications for speech technology. Getting spoken language markers is a computational challenge. Indeed, cognitive load provokes the rise of disfluent phenomena (Bortfeld et al., 2001; Lindström et al., 2008) and more demanding tasks such as monologue have potentially an open vocabulary.

The timing/rhythmic information of speech units (utterances, tracks of communications, words, phones) and silences during conversations seems to be one of the most reliable and recurrent spoken language markers in HD. In our counting study (Riad, Lunven, et al., 2022), we found that rhythm and temporal features were used to predict motor, cognitive, functional and global scores. The timing/rhythmic information of units have also been found to reflect HD in other studies (Perez et al., 2018; Vogel et al., 2012; Rusz, Klempir, Tykalova, et al., 2014; Hertrich & Ackermann, 1994; J. Chan et al., 2021). Silence statistics were the only features in our study that resisted multi-comparison correction for correlation with the striatal volumes. This reinforces the role of the striatum in timing coordination during spoken language production as proposed in the GODIVA model from (Guenther, 2016). Articulatory and phonatory deficits were also found to reflect the HD progression, as found in our different studies (Riad, Titeux, et al., 2020; Riad, Lunven, et al., 2022). These aspects are better characterized than timings, as they have been researched extensively / as indicated by the large number of studies the dysarthria in HD (Romana et al., 2020; Perez et al., 2021; Darley et al., 1969; Rusz, Cmejla, et al., 2013). The striatum is highly connected to motor areas in the cortex, and its role in speech articulation is central in the articulatory neural DIVA model (Guenther, 2016). Extraction of timing information and articulatory deficits pose also a challenge for speech technology. Classic approaches try to get rid of these speaker variability, and timing extraction is only targeted in a few studies (Lu et al., 2016).

In sum, we found that machine learning models with spoken language markers have the potential to reliably estimate the scores of classical motor, cognitive, functional and global scales for preHD and HD individuals. The capability to estimate the severity of this panel of symptoms in a short amount of time, and its potential to be carried remotely, is of great interest from a practical clinical perspective. This has the potential to diminish the human and financial cost for the regular follow-up of HD, and to reduce the cost of future clinical trials.

5.2 Limitations and future work

Let us now discuss the main limitations of this thesis. The first limitations concern the current setup of the BasalVoice database. Collection of the database has been done in a single clinical centre and only in French. Even though the content of this database represents the biggest resource in HD, it remains small by today's machine learning standards. Biomarker studies with small data samples (N<50) have large error bars and this can potentially harm the validity of the conclusions (Varoquaux,

2018). It is necessary in future work, if we want to impact clinical practice, to develop new strategies to scale the collection and annotation of speech data in HD. As we underlined in Section 2.1, data security is crucial to protect patients but complicates data-sharing. There are two potential strategies to overcome these difficulties. First, there are some progresses in *Federated Learning*, especially with clinical applications in mind (Wei et al., 2020; Kairouz et al., 2019). Federated learning is the machine learning setup where many data collectors (ex: mobile devices, different hospitals) collaborate to train a single machine learning model, while keeping the training data decentralized and private. The ultimate goal is to prevent leakage of the information of individuals while allowing the training of a global model. Another avenue to scale the training and test data to validate biomarkers is the use of active learning. With this setup, a model tries to indicate which annotations will be the most useful to make, reducing its total amount (Sun et al., 2021).

Besides, the setup of the BasalVoice database is currently only made of interviews at the hospital with Neuropsychologists. It is unknown how the algorithms and spoken language markers will generalize to more difficult conditions, such as home-based recordings. In addition, the order of the tasks was fixed to allow proper comparison across individuals, but this limits potentially the scope of our conclusions, especially for the study of emotional speech production.

Another limitation of the thesis resides in the construction of markers. We focused on the development of hand-crafted spoken language markers to quantify HD's progression. Their design and hyper-parameters might not be optimal for our ultimate goal, the prediction of clinical scores. In the future, it would be of great of interest to learn these markers directly from the speech signal, such as in (Riad et al., 2021; Millet & Zeghidour, 2019). Besides, we also want to automatically extract more spoken language markers directly based on algorithms developed in Section 3. We would like to examine their predictive power for the different symptoms in HD. In addition, in this thesis, we did not investigate the capabilities of spoken language markers to predict the *psychiatric* aspects of the disease, which represent an important line of future work. Subsequent work, already started now based on the current work, attempts to derive psychiatric markers obtained from the interviews and predict from them the psychiatric sub-scores of the PBA. Besides, we are starting to investigate the link between spoken language and visuo-spatial deficits using analyses of visual description images and of the story in the BasalVoice database.

Finally, it is of prime of importance to go one step further in the validation of HD markers, especially in the *prognostic task* that we mentioned in Section 4, and as

underlined in (Henley et al., 2005). As the database keeps growing, we will be able to test if spoken language markers can be used to follow-up longitudinally preHD and HD and if we can also predict the evolution of the symptoms at future timpe steps.

If spoken language markers prove to be useful in other NDDs, they have the potential to become a *fundamental tool* for clinical practitioners to follow-up NDDs. Indeed, the BasalVoice database is now being augmented with interviews with PD patients, using the same protocol.

The future of Clinical Neurology

As presented in the introduction, current clinical test batteries or candidate biomarkers present a too heavy burden for efficient clinical follow-up in NDDs. In this thesis, we worked towards the construction of algorithms to extract robust spoken language markers from interviews at the hospital in HD, that have the potential to lift this burden.

Fig. 5.1.: Illustration of the future of biomarkers in NDDs.

Spoken language represents one potential modality to capture NDDs symptoms, but the combination of these information with body movements or biological samples represent a promising to obtain a more complete overview of NDDs' symptoms in real life. We list in Figure 5.1, potential novel sensors and modalities that already show promising results to follow-up individuals with NDDs. Vocal assistants can potentially follow conversation of NDDs directly at home and limits visits at the hospital to obtain spoken language markers. In NDDs, motor symptoms are often inescapable, and smart watches appear to be one of the best answer to capture these symptoms (Powers et al., 2021). Finally, new sensors are being developed, using for instance Radio Frequency signals (Vahia et al., 2020) or Skin wearables (Xu et al., 2020), which can be used to obtain body movements and bio signals in a more passive way

Obviously, these passive monitoring sensors pose great *engineering* and *ethical* challenges that need to be answered before really distributing these tools to individuals with NDDs.

Bibliography

World-Health-Organization. (2007). Neurological disorders affect millions of people worldwide, new who report shows (cit. on p. 1).

DiLuca, M., & Olesen, J. (2014). The cost of brain diseases: A burden or a challenge? *Neuron*, *82*(6), 1205–1208 (cit. on p. 1).

Feigin, V. L., Nichols, E., Alam, T., Bannick, M. S., Beghi, E., Blake, N., Culpepper, W. J., Dorsey, E. R., Elbaz, A., Ellenbogen, R. G., et al. (2019). Global, regional, and national burden of neurological disorders, 1990–2016: A systematic analysis for the global burden of disease study 2016. *The Lancet Neurology*, *18*(5), 459–480 (cit. on p. 1).

Gusella, J. F., Wexler, N. S., Conneally, P. M., Naylor, S. L., Anderson, M. A., Tanzi, R. E., Watkins, P. C., Ottina, K., Wallace, M. R., Sakaguchi, A. Y., et al. (1983). A polymorphic dna marker genetically linked to huntington's disease. *Nature*, *306*(5940), 234–238 (cit. on pp. 2, 19).

Marder, K., Zhao, H., Myers, R., Cudkowicz, M., Kayson, E., Kieburtz, K., Orme, C., Paulsen, J., Penney, J., Siemers, E., et al. (2000). Rate of functional decline in huntington's disease. *Neurology*, *54*(2), 452–452 (cit. on pp. 2, 43).

Bachoud-Lévi, A.-C., Rémy, P., Nguyen, J.-P., Brugières, P., Lefaucheur, J.-P., Bourdet, C., Baudic, S., Gaura, V., Maison, P., Haddad, B., et al. (2000). Motor and cognitive improvements in patients with huntington's disease after neural transplantation. *The lancet*, *356*(9246), 1975–1979 (cit. on p. 2).

Tabrizi, S. J., Leavitt, B. R., Landwehrmeyer, G. B., Wild, E. J., Saft, C., Barker, R. A., Blair, N. F., Craufurd, D., Priller, J., Rickards, H., et al. (2019). Targeting huntingtin expression in patients with huntington's disease. *New England Journal of Medicine*, *380*(24), 2307–2316 (cit. on p. 2).

Scahill, R. I., Zeun, P., Osborne-Crowley, K., Johnson, E. B., Gregory, S., Parker, C., Lowe, J., Nair, A., O'Callaghan, C., Langley, C., et al. (2020). Biological and clinical characteristics of gene carriers far from predicted onset in the huntington's disease young adult study (hd-yas): A cross-sectional analysis. *The Lancet Neurology*, *19*(6), 502–512 (cit. on pp. 2, 24, 25, 30).

Tabrizi, S. J., Scahill, R. I., Owen, G., Durr, A., Leavitt, B. R., Roos, R. A., Borowsky, B., Landwehrmeyer, B., Frost, C., Johnson, H., et al. (2013). Predictors of phenotypic progression and disease onset in premanifest and early-stage huntington's disease in the track-hd study: Analysis of 36-month observational data. *The Lancet Neurology*, *12*(7), 637–649 (cit. on pp. 2, 44).

Stout, J. C., Queller, S., Baker, K. N., Cowlishaw, S., Sampaio, C., Fitzer-Attas, C., Borowsky, B., & Investigators, H.-C. (2014). Hd-cab: A cognitive assessment battery for clinical trials in huntington's disease1, 2, 3. *Movement Disorders*, *29*(10), 1281–1288 (cit. on pp. 2, 30, 40, 42, 51).

Lunven, M., Bagnou, J. H., Youssov, K., Gabadinho, A., Fliss, R., Montillot, J., Audureau, E., Bapst, B., Morgado, G., Reilmann, R., et al. (2021). Cognitive decline in huntington's disease in the digitalized arithmetic task (dat). *PLoS ONE* (cit. on pp. 2, 11, 30, 42).

Rusz, J., Saft, C., Schlegel, U., Hoffman, R., & Skodda, S. (2014). Phonatory dysfunction as a preclinical symptom of huntington disease. *PloS one*, *9*(11), e113412 (cit. on pp. 3, 35, 47, 109).

Rusz, J., Klempir, J., Baborova, E., Tykalova, T., Majerova, V., Cmejla, R., Ruzicka, E., & Roth, J. (2013). Objective acoustic quantification of phonatory dysfunction in huntington's disease. *PloS one*, *8*(6), e65881 (cit. on pp. 3, 35, 109).

Romana, A., Bandon, J., Carlozzi, N., Roberts, A., & Provost, E. M. (2020). Classification of Manifest Huntington Disease Using Vowel Distortion Measures. *Proc. Interspeech 2020*, 4966–4970 (cit. on pp. 3, 35, 47, 109, 154).

Perez, M., Jin, W., Le, D., Carlozzi, N., Dayalu, P., Roberts, A., & Provost, E. M. (2018). Classification of huntington disease using acoustic and lexical features. *Interspeech*, *2018*, 1898 (cit. on pp. 3, 12, 35, 154).

Vogel, A. P., Shirbin, C., Churchyard, A. J., & Stout, J. C. (2012). Speech acoustic markers of early stage and prodromal huntington's disease: A marker of disease onset? *Neuropsychologia*, *50*(14), 3273–3278 (cit. on pp. 3, 34, 47, 109, 119, 153, 154).

Ullman, M. T., Corkin, S., Coppola, M., Hickok, G., Growdon, J. H., Koroshetz, W. J., & Pinker, S. (1997). A neural dissociation within language: Evidence that the mental dictionary is part of declarative memory, and that grammatical rules are processed by the procedural system. *Journal of cognitive neuroscience*, *9*(2), 266–276 (cit. on pp. 3, 31, 34).

Teichmann, M., Dupoux, E., Kouider, S., Brugières, P., Boissé, M.-F., Baudic, S., Cesaro, P., Peschanski, M., & Bachoud-Lévi, A.-C. (2005). The role of the striatum in rule application: The model of huntington's disease at early stage. *Brain*, *128*(5), 1155–1167 (cit. on pp. 3, 31).

Hinzen, W., Rosselló, J., Morey, C., Camara, E., Garcia-Gorro, C., Salvador, R., & de Diego-Balaguer, R. (2017). A systematic linguistic profile of spontaneous narrative speech in pre-symptomatic and early stage huntington's disease. *Cortex; a journal devoted to the study of the nervous system and behavior* (cit. on pp. 3, 31, 34).

Friederici, A. D. (2017). *Language in our brain: The origins of a uniquely human capacity*. MIT Press. (Cit. on pp. 3, 6).

Guenther, F. H. (2016). *Neural control of speech*. MIT Press. (Cit. on pp. 3, 6–10, 32, 154).

Jacquemot, C., & Bachoud-Lévi, A.-C. (2021). Striatum and language processing: Where do we stand? *Cognition*, 104785 (cit. on pp. 3, 32).

Lahiri, R., Kumar, M., Bishop, S., & Narayanan, S. (2020). Learning domain invariant representations for child-adult classification from speech. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6749–6753 (cit. on p. 3).

Titeux*, H., Riad*, R., Cao, X.-N., Hamilakis, N., Madden, K., Cristia, A., Bachoud-Lévi, A.-C., & Dupoux, E. (2020). Seshat: A tool for managing and verifying annotation campaigns of audio data [* Equal contribution]. *Twelfth International Conference on Language Resources and Evaluation (LREC 2020)* (cit. on pp. 5, 72).

Riad, R., Bachoud-Lévi, A.-C., Rudzicz, F., & Dupoux, E. (2020a). Identification of primary and collateral tracks in stuttered speech. *Twelfth International Conference on Language Resources and Evaluation (LREC 2020)* (cit. on p. 5).

Riad, R., Titeux, H., Lemoine, L., Montillot, J., Bagnou, J. H., Cao, X. N., Dupoux, E., & Bachoud-Lévi, A.-C. (2020). Vocal markers from sustained phonation in huntington's disease. *INTERSPEECH* (cit. on pp. 5, 109, 116, 119, 123, 154).

Titeux, H., & Riad, R. (2021). Pygamma-agreement: Gamma measure for inter/intra-annotator agreement in python. *Journal of Open Source Software*, *6*(62), 2989 (cit. on pp. 5, 63, 71).

Riad*, R., Titeux*, H., Lemoine, L., Montillot, J., Sliwinski, A., Bagnou, J. H., Cao, X. N., Bachoud-Lévi, A.-C., & Dupoux, E. (2021). A comparison study on patient-psychologist voice diarization [Under review, * Equal contribution]. (Cit. on p. 5).

Riad, R., Lunven, M., Titeux, H., Cao, X. N., Bagnou, J. H., Lemoine, L., Montillot, J., Sliwinski, A., Youssov, K., de Langavant, L. C., Dupoux, E., & Bachoud-Lévi, A.-C. (2022). Predicting clinical scores in huntington's disease: A lightweight speech test. *Journal of Neurology* (cit. on pp. 5, 14, 153, 154).

Gallezot*, C., Riad*, R., Titeux, H., Cao, X. N., Bagnou, J. H., Lemoine, L., Montillot, J., Sliwinski, A., Youssov, K., de Langavant, L. C., Dupoux, E., & Bachoud-Lévi, A.-C. (2022). Vocal and linguistic expression impairments of emotions in huntington's disease [Submitted to Cortex, * Equal contribution]. (Cit. on pp. 6, 14, 140).

Riad, R., Karadayi, J., Bachoud-Lévi, A.-C., & Dupoux, E. (2021). Learning spectro-temporal representations of complex sounds with parameterized neural networks. *The Journal of the Acoustical Society of America*, *150*(1), 353–366 (cit. on pp. 6, 155).

Riad, R., Teboul, O., Grangier, D., & Zeghidour, N. (2022). Learning strides in convolutional neural networks. *ICLR* (cit. on p. 6).

Levelt, W. J. (1993). *Speaking: From intention to articulation* (Vol. 1). MIT press. (Cit. on pp. 6–8, 33, 80).

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *science*, *298*(5598), 1569–1579 (cit. on p. 6).

Tomasello, M. (2010). *Origins of human communication*. MIT press. (Cit. on p. 6).

Clark, H. H. (1996). *Using language*. Cambridge university press. (Cit. on pp. 6, 9–11, 81, 95, 152).

Stokoe, E. (2018). *Talk: The science of conversation*. Hachette UK. (Cit. on pp. 6, 80).

Scherer, K. R. (1995). Expression of emotion in voice and music. *Journal of voice*, *9*(3), 235–248 (cit. on p. 6).

Mehler, J., & Dupoux, E. (2002). *Naitre humain*. Odile Jacob. (Cit. on p. 6).

Johnson, K. (2004). Acoustic and auditory phonetics. *Phonetica*, *61*(1), 56–58 (cit. on p. 6).

Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge University Press. (Cit. on p. 6).

Jurafsky, D., & Martin, J. H. (2000). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. (Cit. on p. 6).

Shriberg, E. E. (1994). *Preliminaries to a theory of speech disfluencies* (Doctoral dissertation). Citeseer. (Cit. on pp. 10, 11, 55, 81).

Picard, R. W. (2000). *Affective computing*. MIT press. (Cit. on p. 11).

Pinker, S. (2007). *The stuff of thought: Language as a window into human nature*. Penguin. (Cit. on p. 11).

Cavanna, A. E., Schrag, A., Morley, D., Orth, M., Robertson, M., Joyce, E., Critchley, H., & Selai, C. (2008). The gilles de la tourette syndrome–quality of life scale (gts-qol): Development and validation. *Neurology*, *71*(18), 1410–1416 (cit. on p. 11).

Gramunt, N., Sánchez-Benavides, G., Buschke, H., Lipton, R. B., Masramon, X., Gispert, J. D., Peña-Casanova, J., Fauria, K., & Molinuevo, J. L. (2016). Psychometric properties of the memory binding test: Test-retest reliability and convergent validity. *Journal of Alzheimer's Disease*, *50*(4), 999–1010 (cit. on p. 11).

Brockwell, S. E., & Gordon, I. R. (2001). A comparison of statistical methods for meta-analysis. *Statistics in medicine*, *20*(6), 825–840 (cit. on p. 12).

Reshef, Y. A., Reshef, D. N., Finucane, H. K., Sabeti, P. C., & Mitzenmacher, M. (2016). Measuring dependence powerfully and equitably. *The Journal of Machine Learning Research*, *17*(1), 7406–7468 (cit. on p. 12).

Bzdok, D., & Ioannidis, J. P. (2019). Exploration, inference, and prediction in neuroscience and biomedicine. *Trends in neurosciences*, *42*(4), 251–262 (cit. on p. 12).

Myszczynska, M. A., Ojamies, P. N., Lacoste, A. M., Neil, D., Saffari, A., Mead, R., Hautbergue, G. M., Holbrook, J. D., & Ferraiuolo, L. (2020). Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nature Reviews Neurology*, *16*(8), 440–456 (cit. on p. 12).

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual review of clinical psychology*, *14*, 91–118 (cit. on p. 12).

Wilkinson, J., Arnold, K. F., Murray, E. J., van Smeden, M., Carr, K., Sippy, R., de Kamps, M., Beam, A., Konigorski, S., Lippert, C., et al. (2020). Time to reality check the promises of machine learning-powered precision medicine. *The Lancet Digital Health* (cit. on p. 12).

Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage*, *180*, 68–77 (cit. on pp. 12, 154).

Koval, I., Bône, A., Louis, M., Lartigue, T., Bottani, S., Marcoux, A., Samper-Gonzalez, J., Burgos, N., Charlier, B., Bertrand, A., et al. (2021). Ad course map charts alzheimer's disease progression. *Scientific Reports*, *11*(1), 1–16 (cit. on p. 12).

Gajos, K. Z., Reinecke, K., Donovan, M., Stephen, C. D., Hung, A. Y., Schmahmann, J. D., & Gupta, A. S. (2020). Computer mouse use captures ataxia and parkinsonism, enabling accurate measurement and detection. *Movement Disorders*, *35*(2), 354–358 (cit. on p. 12).

Boyd, S., Boyd, S. P., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press. (Cit. on p. 13).

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606 (cit. on p. 14).

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van't Veer, A. E., & Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in cognitive sciences*, *23*(10), 815–818 (cit. on p. 14).

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The journal of the acoustical society of america*, *8*(3), 185–190 (cit. on p. 15).

Elliott, T. M., & Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *PLoS computational biology*, *5*(3) (cit. on pp. 15, 16).

Elhilali, M., Chi, T., & Shamma, S. A. (2003). A spectro-temporal modulation index (stmi) for assessment of speech intelligibility. *Speech communication*, *41*(2-3), 331–348 (cit. on pp. 15, 109).

Edraki, A., Chan, W.-Y., Jensen, J., & Fogerty, D. (2019). Improvement and assessment of spectro-temporal modulation analysis for speech intelligibility estimation. *Interspeech 2019Annual Conference of the International Speech Communication Association*, 1378–1382 (cit. on p. 15).

Arnal, L. H., Flinker, A., Kleinschmidt, A., Giraud, A.-L., & Poeppel, D. (2015). Human screams occupy a privileged niche in the communication soundscape. *Current Biology*, *25*(15), 2051–2056 (cit. on pp. 15, 16).

Mesgarani, N., Slaney, M., & Shamma, S. A. (2006). Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(3), 920–930 (cit. on p. 15).

Chang, S.-Y., & Morgan, N. (2014). Robust cnn-based speech recognition with gabor filter kernels. *Fifteenth annual conference of the international speech communication association* (cit. on p. 15).

Vuong, T., Xia, Y., & Stern, R. M. (2020). Learnable Spectro-Temporal Receptive Fields for Robust Voice Type Discrimination. *Proc. Interspeech 2020*, 1957–1961 (cit. on p. 15).

Talkin, D., & Kleijn, W. B. (1995). A robust algorithm for pitch tracking (rapt). *Speech coding and synthesis*, *495*, 518 (cit. on p. 16).

Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages*. Blackwell Publishers. (Cit. on p. 16).

Mporas, I., & Ganchev, T. (2009). Estimation of unknown speaker's height from speech. *International Journal of Speech Technology*, *12*(4), 149–160 (cit. on p. 17).

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., & Narayanan, S. (2010). The interspeech 2010 paralinguistic challenge. *Proc. INTERSPEECH 2010, Makuhari, Japan*, 2794–2797 (cit. on p. 17).

Weiss, B., & Burkhardt, F. (2010). Voice attributes affecting likability perception. *Eleventh Annual Conference of the International Speech Communication Association* (cit. on p. 17).

Krajewski, J., Batliner, A., & Golz, M. (2009). Acoustic sleepiness detection: Framework and validation of a speech-adapted pattern recognition approach. *Behavior research methods*, *41*(3), 795–804 (cit. on p. 17).

Gideon, J., Provost, E. M., & McInnis, M. (2016). Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder. *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, 2359–2363 (cit. on p. 17).

Schuller, B., Steidl, S., Batliner, A., Schiel, F., & Krajewski, J. (2011). The interspeech 2011 speaker state challenge (cit. on p. 17).

Laskowski, K., Ostendorf, M., & Schultz, T. (2008). Modeling vocal interaction for text-independent participant characterization in multi-party conversation. *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, 148–155 (cit. on p. 17).

Schuller, B. W. (2018). Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, *61*(5), 90–99 (cit. on p. 17).

Hansen, J. H., Bou-Ghazale, S. E., Sarikaya, R., & Pellom, B. (1997). Getting started with susas: A speech under simulated and actual stress database. *Eurospeech*, *97*(4), 1743–46 (cit. on p. 17).

Simon, D., Craig, K. D., Gosselin, F., Belin, P., & Rainville, P. (2008). Recognition and discrimination of prototypical dynamic expressions of pain and emotions. *PAIN®*, *135*(1-2), 55–64 (cit. on p. 17).

Schuller, B., & Batliner, A. (2013). *Computational paralinguistics: Emotion, affect and personality in speech and language processing*. John Wiley & Sons. (Cit. on p. 17).

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., MuLler, C., & Narayanan, S. (2013). Paralinguistics in speech and language—state-of-the-art and the challenge. *Computer Speech & Language*, *27*(1), 4–39 (cit. on p. 17).

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: The munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM international conference on Multimedia*, 1459–1462 (cit. on p. 17).

Williamson, J. R., Quatieri, T. F., Helfer, B. S., Ciccarelli, G., & Mehta, D. D. (2014). Vocal and facial biomarkers of depression based on motor incoordination and timing. *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 65–72 (cit. on p. 18).

Smith, K. M., Williamson, J. R., & Quatieri, T. F. (2017). Vocal markers of motor, cognitive, and depressive symptoms in parkinson's disease. *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 71–78 (cit. on p. 18).

Quatieri, T. F., Williamson, J. R., Smalt, C. J., Patel, T., Perricone, J., Mehta, D. D., Helfer, B. S., Ciccarelli, G., Ricke, D., Malyska, N., et al. (2015). Vocal biomarkers to discriminate cognitive load in a working memory task. *Sixteenth Annual Conference of the International Speech Communication Association* (cit. on p. 18).

Perez, M., Romana, A., Roberts, A., Carlozzi, N., Miner, J. A., Dayalu, P., & Provost, E. M. (2021). Articulatory Coordination for Speech Motor Tracking in Huntington Disease. *Proc. Interspeech 2021*, 1409–1413 (cit. on pp. 18, 35, 47, 154).

Moro-Velazquez, L., Gomez-Garcia, J. A., Godino-Llorente, J. I., & Andrade-Miranda, G. (2015). Modulation spectra morphological parameters: A new method to assess voice pathologies according to the grbas scale. *BioMed research international*, *2015* (cit. on pp. 18, 109).

Kodrasi, I., & Bourlard, H. (2020). Spectro-temporal sparsity characterization for dysarthric speech detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 1210–1222 (cit. on pp. 18, 109).

Janbakhshi, P., Kodrasi, I., & Bourlard, H. (2019). Pathological speech intelligibility assessment based on the short-time objective intelligibility measure. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6405–6409 (cit. on pp. 18, 109).

Bozkurt, E., Toledo-Ronen, O., Sorin, A., & Hoory, R. (2014). Exploring modulation spectrum features for speech-based depression level classification. *Fifteenth Annual Conference of the International Speech Communication Association* (cit. on p. 18).

Thoret, E., Andrillon, T., Gauriau, C., Leger, D., & Pressnitzer, D. (2020). Sleep deprivation impacts speech spectro-temporal modulations. *e-FA2020 (e-Forum Acusticum 2020)* (cit. on p. 18).

Ghio, A., Lalain, M., Giusti, L., Pouchoulin, G., Robert, D., Rebourg, M., Fredouille, C., Laaridh, I., & Woisard, V. (2018). Une mesure d'intelligibilité par décodage acoustico-phonétique de pseudo-mots dans le cas de parole atypique. *XXXIIe Journées d'Etudes sur la Parole*, 285–293 (cit. on p. 18).

Fredouille, C., Ghio, A., Laaridh, I., Lalain, M., & Woisard, V. (2019). Acoustic-phonetic decoding for speech intelligibility evaluation in the context of head and neck cancers. *International Congress of Phonetic Sciences (ICPhS)*, 3051–3055 (cit. on p. 18).

Romana, A., Bandon, J., Perez, M., Gutierrez, S., Richter, R., Roberts, A., & Provost, E. M. (2021). Automatically detecting errors and disfluencies in read speech to predict cognitive impairment in people with parkinson's disease. *Proc. Interspeech 2021*, 1907–1911 (cit. on p. 18).

Huntington, G. et al. (1872). On chorea (cit. on p. 19).

Rubinsztein, D. C. (2003). Molecular biology of huntington's disease (hd) and hd-like disorders. *Genetics of movement disorders* (pp. 365–383). Elsevier. (Cit. on p. 19).

Mangiarini, L., Sathasivam, K., Seller, M., Cozens, B., Harper, A., Hetherington, C., Lawton, M., Trottier, Y., Lehrach, H., Davies, S. W., et al. (1996). Exon 1 of the hd gene with an expanded cag repeat is sufficient to cause a progressive neurological phenotype in transgenic mice. *Cell*, *87*(3), 493–506 (cit. on p. 19).

Walker, F. O. (2007). Huntington's disease. *The Lancet*, *369*(9557), 218–228 (cit. on pp. 19–24, 26, 27).

Bates, G. P. (2005). The molecular genetics of huntington disease—a history. *Nature Reviews Genetics*, *6*(10), 766–773 (cit. on p. 20).

Shaw, M., & Caro, A. (1982). The mutation rate to huntington's chorea. *Journal of medical genetics*, *19*(3), 161–167 (cit. on p. 20).

Bates, G., Harper, P., & Jones, L. (2002). Huntington's disease. ed. *Press OU, editor2002* (cit. on pp. 20, 21, 27).

Takano, H., Cancel, G., Ikeuchi, T., Lorenzetti, D., Mawad, R., Stevanin, G., Didierjean, O., Dürr, A., Oyake, M., Shimohata, T., et al. (1998). Close associations between prevalences of dominantly inherited spinocerebellar ataxias with cag-repeat expansions and frequencies of large normal cag alleles in japanese and caucasian populations. *The American Journal of Human Genetics*, *63*(4), 1060–1066 (cit. on p. 20).

Wright, H. H., Still, C. N., & Abramson, R. K. (1981). Huntington's disease in black kindreds in south carolina. *Archives of neurology*, *38*(7), 412–414 (cit. on p. 20).

McColgan, P., & Tabrizi, S. J. (2018). Huntington's disease: A clinical review. *European journal of neurology*, *25*(1), 24–34 (cit. on p. 20).

Crowell, V., Houghton, R., Tomar, A., Fernandes, T., & Squitieri, F. (2021). Modeling manifest huntington's disease prevalence using diagnosed incidence and survival time. *Neuroepidemiology*, *55*(5), 361–368 (cit. on p. 20).

Tabrizi, S. J., Flower, M. D., Ross, C. A., & Wild, E. J. (2020). Huntington disease: New insights into molecular pathogenesis and therapeutic opportunities. *Nature Reviews Neurology*, *16*(10), 529–546 (cit. on pp. 21, 24).

Quarrell, O., O'Donovan, K. L., Bandmann, O., & Strong, M. (2012). The prevalence of juvenile huntington's disease: A review of the literature and meta-analysis. *PLoS currents*, *4* (cit. on p. 21).

Chaganti, S. S., McCusker, E. A., & Loy, C. T. (2017). What do we know about late onset huntington's disease? *Journal of Huntington's disease*, *6*(2), 95–103 (cit. on p. 21).

Cox, S. M., & McKellin, W. (1999). "there's this thing in our family": Predictive testing and the construction of risk for huntington disease. *Sociology of Health and Illness*, *21*(5), 622–646 (cit. on p. 22).

Quaid, K. A., & Morris, M. (1993). Reluctance to undergo predictive testing: The case of huntington disease. *American journal of medical genetics*, *45*(1), 41–45 (cit. on p. 22).

Kessler, S., Field, T., Worth, L., Mosbarger, H., Opitz, J. M., & Reynolds, J. F. (1987). Attitudes of persons at risk for huntington disease toward predictive testing. *American journal of medical genetics*, *26*(2), 259–270 (cit. on p. 22).

Langbehn, D. R., Brinkman, R. R., Falush, D., Paulsen, J. S., Hayden, M., & an International Huntington's Disease Collaborative Group. (2004). A new model for prediction of the age of onset and penetrance for huntington's disease based on cag length. *Clinical genetics*, *65*(4), 267–277 (cit. on p. 22).

Schulte, J., & Littleton, J. T. (2011). The biological function of the huntingtin protein and its relevance to huntington's disease pathology. *Current trends in neurology*, *5*, 65 (cit. on p. 22).

Braude, P. R., De Wert, G. M., Evers-Kiebooms, G., Pettigrew, R. A., & Geraedts, J. P. (1998). Non-disclosure preimplantation genetic diagnosis for huntington's disease: Practical and ethical dilemmas. *Prenatal Diagnosis: Published in Affiliation With the International Society for Prenatal Diagnosis*, *18*(13), 1422–1426 (cit. on p. 22).

Van Rij, M. C., De Rademaeker, M., Moutou, C., Dreesen, J. C., De Rycke, M., Liebaers, I., Geraedts, J. P., De Die-Smulders, C. E., & Viville, S. (2012). Preimplantation genetic diagnosis (pgd) for huntington's disease: The experience of three european centres. *European journal of human genetics*, *20*(4), 368–375 (cit. on p. 22).

Ross, C. A., Aylward, E. H., Wild, E. J., Langbehn, D. R., Long, J. D., Warner, J. H., Scahill, R. I., Leavitt, B. R., Stout, J. C., Paulsen, J. S., et al. (2014). Huntington disease: Natural history, biomarkers and prospects for therapeutics. *Nature Reviews Neurology*, *10*(4), 204–216 (cit. on p. 22).

Ready, R. E., Mathews, M., Leserman, A., & Paulsen, J. S. (2008). Patient and caregiver quality of life in huntington's disease. *Movement Disorders*, *23*(5), 721–726 (cit. on p. 23).

Rodrigues, F. B., Abreu, D., Damásio, J., Goncalves, N., Correia-Guedes, L., Coelho, M., Ferreira, J. J., & of the European Huntington's Disease Network, R. I. (2017). Survival, mortality, causes and places of death in a european huntington's disease prospective cohort. *Movement disorders clinical practice*, *4*(5), 737–742 (cit. on p. 23).

Rangone, H., Humbert, S., & Saudou, F. (2004). Huntington's disease: How does huntingtin, an anti-apoptotic protein, become toxic? *Pathologie Biologie*, *52*(6), 338–342 (cit. on p. 24).

De Diego Balaguer, R., & Bachoud, A. (2006). Aspects cliniques et neuropsychologiques de la maladie de huntington. *Actualités sur les démences: aspects cliniques et neuropsychologiques*, 381–411 (cit. on p. 24).

Douaud, G., Gaura, V., Ribeiro, M.-J., Lethimonnier, F., Maroy, R., Verny, C., Krystkowiak, P., Damier, P., Bachoud-Levi, A.-C., Hantraye, P., et al. (2006). Distribution of grey matter atrophy in huntington's disease patients: A combined roi-based and voxel-based morphometric study. *Neuroimage*, *32*(4), 1562–1575 (cit. on pp. 24, 26).

Stoffers, D., Sheldon, S., Kuperman, J., Goldstein, J., Corey-Bloom, J., & Aron, A. (2010). Contrasting gray and white matter changes in preclinical huntington disease: An mri study. *Neurology*, *74*(15), 1208–1216 (cit. on p. 24).

Phillips, O., Squitieri, F., Sanchez-Castaneda, C., Elifani, F., Caltagirone, C., Sabatini, U., & Di Paola, M. (2014). Deep white matter in huntington's disease. *PloS one*, *9*(10), e109676 (cit. on p. 24).

Matsui, J. T., Vaidya, J. G., Wassermann, D., Kim, R. E., Magnotta, V. A., Johnson, H. J., Paulsen, J. S., Investigators, P.-H., & of the Huntington Study Group, C. (2015). Prefrontal cortex white matter tracts in prodromal h untington disease. *Human brain mapping*, *36*(10), 3717–3732 (cit. on p. 24).

Graff-Radford, J., Williams, L., Jones, D. T., & Benarroch, E. E. (2017). Caudate nucleus as a component of networks controlling behavior. *Neurology*, *89*(21), 2192–2197 (cit. on pp. 24, 25).

Alexander, G. E., DeLong, M. R., & Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual review of neuroscience*, *9*(1), 357–381 (cit. on pp. 24, 25).

Turner, R. S., & Desmurget, M. (2010). Basal ganglia contributions to motor control: A vigorous tutor. *Current opinion in neurobiology*, *20*(6), 704–716 (cit. on p. 25).

Lieberman, P. (2002). *Human language and our reptilian brain: The subcortical bases of speech, syntax, and thought*. Harvard University Press. (Cit. on p. 25).

Poldrack, R. A., Prabhakaran, V., Seger, C. A., & Gabrieli, J. D. (1999). Striatal activation during acquisition of a cognitive skill. *Neuropsychology*, *13*(4), 564 (cit. on p. 25).

van Schouwenburg, M. R., O'Shea, J., Mars, R. B., Rushworth, M. F., & Cools, R. (2012). Controlling human striatal cognitive function via the frontal cortex. *Journal of Neuroscience*, *32*(16), 5631–5637 (cit. on p. 25).

Báez-Mendoza, R., & Schultz, W. (2013). The role of the striatum in social behavior. *Frontiers in Neuroscience*, *7*, 233 (cit. on p. 25).

Daniel, R., & Pollmann, S. (2014). A universal role of the ventral striatum in reward-based learning: Evidence from human studies. *Neurobiology of learning and memory*, *114*, 90–100 (cit. on p. 25).

Frank, M. J., Seeberger, L. C., & O'reilly, R. C. (2004). By carrot or by stick: Cognitive reinforcement learning in parkinsonism. *Science*, *306*(5703), 1940–1943 (cit. on p. 25).

Vonsattel, J.-P., Myers, R. H., Stevens, T. J., Ferrante, R. J., Bird, E. D., & Richardson, E. P. (1985). Neuropathological classification of huntington's disease. *Journal of Neuropathology & Experimental Neurology*, *44*(6), 559–577 (cit. on p. 24).

Roos, R., Pruyt, J., De Vries, J., & Bots, G. (1985). Neuronal distribution in the putamen in huntington's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, *48*(5), 422–425 (cit. on p. 24).

Vonsattel, J. P. G. (2008). Huntington disease models and human neuropathology: Similarities and differences. *Acta neuropathologica*, *115*(1), 55–69 (cit. on p. 24).

Halliday, G., McRitchie, D., Macdonald, V., Double, K., Trent, R., & McCusker, E. (1998). Regional specificity of brain atrophy in huntington's disease. *Experimental neurology*, *154*(2), 663–672 (cit. on p. 26).

Novak, M. J., & Tabrizi, S. J. (2010). Huntington's disease. *Bmj*, *340* (cit. on pp. 26, 27).

Kremer, H., Group, H. S. et al. (1996). Unified huntington's disease rating scale: Reliability and consistency. *Movement disorders*, *11*, 136–142 (cit. on pp. 26, 41, 42).

Duff, K., Paulsen, J., Mills, J., Beglinger, L., Moser, D., Smith, M., Langbehn, D., Stout, J., Queller, S., Harrington, D., et al. (2010). Mild cognitive impairment in prediagnosed huntington disease. *Neurology*, *75*(6), 500–507 (cit. on pp. 26, 29).

Paulsen, J. S., Nehl, C., Hoth, K. F., Kanz, J. E., Benjamin, M., Conybeare, R., McDowell, B., & Turner, B. (2005). Depression and stages of huntington's disease. *The Journal of neuropsychiatry and clinical neurosciences*, *17*(4), 496–502 (cit. on p. 26).

Schobel, S. A., Palermo, G., Auinger, P., Long, J. D., Ma, S., Khwaja, O. S., Trundell, D., Cudkowicz, M., Hersch, S., Sampaio, C., et al. (2017). Motor, cognitive, and functional declines contribute to a single progressive factor in early hd. *Neurology*, *89*(24), 2495–2502 (cit. on pp. 26, 43).

Bates, G. P., Dorsey, R., Gusella, J. F., Hayden, M. R., Kay, C., Leavitt, B. R., Nance, M., Ross, C. A., Scahill, R. I., Wetzel, R., et al. (2015). Huntington disease. *Nature reviews Disease primers*, *1*(1), 1–21 (cit. on pp. 27, 29).

Youssov, K., & Bachoud-Levi, A. (2008). Maladie de huntington: Aspects diagnostiques actuels et applications pratiques. *EMC-Neurol*, *5*, 1–13 (cit. on pp. 27, 29).

Kirkwood, S. C., Su, J. L., Conneally, P. M., & Foroud, T. (2001). Progression of symptoms in the early and middle stages of huntington disease. *Archives of neurology*, *58*(2), 273–278 (cit. on pp. 27, 28).

Louis, E. D., Lee, P., Quinn, L., & Marder, K. (1999). Dystonia in huntington's disease: Prevalence and clinical characteristics. *Movement disorders: official journal of the Movement Disorder Society*, *14*(1), 95–101 (cit. on p. 28).

Unti, E., Mazzucchi, S., Palermo, G., Bonuccelli, U., & Ceravolo, R. (2017). Antipsychotic drugs in huntington's disease. *Expert review of neurotherapeutics*, *17*(3), 227–237 (cit. on p. 28).

Ross, C. A., Pantelyat, A., Kogan, J., & Brandt, J. (2014). Determinants of functional disability in huntington's disease: Role of cognitive and motor dysfunction. *Movement disorders*, *29*(11), 1351–1358 (cit. on p. 28).

Hamilton, J., Salmon, D., Corey-Bloom, J., Gamst, A., Paulsen, J., Jerkins, S., Jacobson, M., & Peavy, G. (2003). Behavioural abnormalities contribute to functional decline in huntington's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, *74*(1), 120–122 (cit. on p. 28).

Duff, K., Paulsen, J. S., Beglinger, L. J., Langbehn, D. R., Wang, C., Stout, J. C., Ross, C. A., Aylward, E., Carlozzi, N. E., Queller, S., et al. (2010). 'frontal' behaviors before the diagnosis of huntington's disease and their relationship to markers of disease progression: Evidence of early lack of awareness. *The Journal of neuropsychiatry and clinical neurosciences*, *22*(2), 196–207 (cit. on p. 29).

De Diego-Balaguer, R., Couette, M., Dolbeau, G., Durr, A., Youssov, K., & Bachoud-Levi, A.-C. (2008). Striatal degeneration impairs language learning: Evidence from huntington's disease. *Brain*, *131*(11), 2870–2881 (cit. on p. 29).

Aretouli, E., & Brandt, J. (2010). Episodic memory in dementia: Characteristics of new learning that differentiate alzheimer's, huntington's, and parkinson's diseases. *Archives of clinical neuropsychology*, *25*(5), 396–409 (cit. on p. 29).

Glikmann-Johnston, Y., Fink, K. D., Deng, P., Torrest, A., & Stout, J. C. (2019). Spatial memory in huntington's disease: A comparative review of human and animal data. *Neuroscience & Biobehavioral Reviews*, *98*, 194–207 (cit. on p. 29).

Glikmann-Johnston, Y., Mercieca, E.-C., Carmichael, A. M., Alexander, B., Harding, I. H., & Stout, J. C. (2021). Hippocampal and striatal volumes correlate with spatial memory impairment in huntington's disease. *Journal of neuroscience research*, *99*(11), 2948–2963 (cit. on p. 29).

Thompson, J. C., Harris, J., Sollom, A. C., Stopford, C. L., Howard, E., Snowden, J. S., & Craufurd, D. (2012). Longitudinal evaluation of neuropsychiatric symptoms in huntington's disease. *The Journal of neuropsychiatry and clinical neurosciences*, *24*(1), 53–60 (cit. on p. 29).

Paulsen, J. S., Ready, R., Hamilton, J., Mega, M., & Cummings, J. (2001). Neuropsychiatric aspects of huntington's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, *71*(3), 310–314 (cit. on p. 29).

Solberg, O. K., Filkuková, P., Frich, J. C., & Feragen, K. J. B. (2018). Age at death and causes of death in patients with huntington disease in norway in 1986–2015. *Journal of Huntington's disease*, *7*(1), 77–86 (cit. on p. 29).

Killoran, A., & Biglan, K. (2016). Biomarkers for huntington's disease: A brief overview. *Journal of Rare Diseases Research & Treatment*, *1*(2) (cit. on p. 29).

Henley, S. M., Bates, G. P., & Tabrizi, S. J. (2005). Biomarkers for neurodegenerative diseases. *Current opinion in neurology*, *18*(6), 698–705 (cit. on pp. 30, 156).

Mason, S. L., Daws, R. E., Soreq, E., Johnson, E. B., Scahill, R. I., Tabrizi, S. J., Barker, R. A., & Hampshire, A. (2018). Predicting clinical diagnosis in huntington's disease: An imaging polymarker. *Annals of neurology*, *83*(3), 532–543 (cit. on p. 30).

Zeun, P., Scahill, R. I., Tabrizi, S. J., & Wild, E. J. (2019). Fluid and imaging biomarkers for huntington's disease. *Molecular and Cellular Neuroscience*, *97*, 67–80 (cit. on p. 30).

Rodrigues, F. B., Byrne, L. M., Tortelli, R., Johnson, E. B., Wijeratne, P. A., Arridge, M., De Vita, E., Ghazaleh, N., Houghton, R., Furby, H., et al. (2020). Mutant huntingtin and neurofilament light have distinct longitudinal dynamics in huntington's disease. *Science Translational Medicine*, *12*(574) (cit. on p. 30).

Rusz, J., Klempir, J., Tykalova, T., Baborova, E., Cmejla, R., Ruzicka, E., & Roth, J. (2014). Characteristics and occurrence of speech impairment in huntington's disease: Possible influence of antipsychotic medication. *Journal of Neural Transmission*, *121*(12), 1529–1539 (cit. on pp. 30, 35, 109, 154).

Darley, F. L., Aronson, A. E., & Brown, J. R. (1969). Differential diagnostic patterns of dysarthria. *Journal of speech and hearing research*, *12*(2), 246–269 (cit. on pp. 30, 119, 154).

Diehl, S. K., Mefferd, A. S., Lin, Y.-C., Sellers, J., McDonell, K. E., de Riesthal, M., & Claassen, D. O. (2019). Motor speech patterns in huntington disease. *Neurology*, *93*(22), e2042–e2052 (cit. on pp. 31, 119).

Chan, J. C., Stout, J. C., & Vogel, A. P. (2019). Speech in prodromal and symptomatic huntington's disease as a model of measuring onset and progression in dominantly inherited neurodegenerative diseases. *Neuroscience & Biobehavioral Reviews*, *107*, 450–460 (cit. on pp. 31, 119).

Speedie, L. J., Brake, N., Folstein, S. E., Bowers, D., & Heilman, K. M. (1990). Comprehension of prosody in huntington's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, *53*(7), 607–610 (cit. on p. 31).

Profant, O., Roth, J., Bures, Z., Balogova, Z., Liskova, I., Betka, J., & Syka, J. (2017). Auditory dysfunction in patients with huntington's disease. *Clinical Neurophysiology*, *128*(10), 1946–1953 (cit. on p. 31).

Sambin, S., Teichmann, M., de Diego Balaguer, R., Giavazzi, M., Sportiche, D., Schlenker, P., & Bachoud-Levi, A.-C. (2012). The role of the striatum in sentence processing: Disentangling syntax from working memory in huntington's disease. *Neuropsychologia*, *50*(11), 2625–2635 (cit. on p. 31).

Giavazzi, M., Daland, R., Palminteri, S., Peperkamp, S., Brugieres, P., Jacquemot, C., Schramm, C., de Langavant, L. C., & Bachoud-Lévi, A.-C. (2018). The role of the striatum in linguistic selection: Evidence from huntington's disease and computational modeling. *cortex*, *109*, 189–204 (cit. on p. 31).

Teichmann, M., Dupoux, E., Cesaro, P., & Bachoud-Levi, A.-C. (2008). The role of the striatum in sentence processing: Evidence from a priming study in early stages of huntington's disease. *Neuropsychologia*, *46*(1), 174–185 (cit. on p. 31).

Teichmann, M., Dupoux, E., Kouider, S., & Bachoud-Lévi, A.-C. (2006). The role of the striatum in processing language rules: Evidence from word perception in huntington's disease. *Journal of Cognitive Neuroscience*, *18*(9), 1555–1569 (cit. on p. 31).

PODOLL, K., CASPARY, P., LANGE, H. W., & NOTH, J. (1988). Language functions in huntington's disease. *Brain*, *111*(6), 1475–1503 (cit. on p. 31).

Wallesch, C.-W., & Fehrenbach, R. A. (1988). On the neurolinguistic nature of language abnormalities in huntington's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, *51*(3), 367–373 (cit. on p. 31).

Saldert, C., Fors, A., Stroberg, S., & Hartelius, L. (2010). Comprehension of complex discourse in different stages of huntington's disease. *International journal of language & communication disorders*, *45*(6), 656–669 (cit. on p. 31).

Chenery, H. J., Copland, D. A., & Murdoch, B. E. (2002). Complex language functions and subcortical mechanisms: Evidence from huntington's disease and patients with non-thalamic subcortical lesions. *International journal of language & communication disorders*, *37*(4), 459–474 (cit. on pp. 31, 34).

Gordon, W. P., & Illes, J. (1987). Neurolinguistic characteristics of language production in huntington's disease: A preliminary report. *Brain and Language*, *31*(1), 1–10 (cit. on pp. 31, 34).

Gagnon, M., Barrette, J., & Macoir, J. (2018). Language disorders in huntington disease: A systematic literature review. *Cognitive and Behavioral Neurology*, *31*(4), 179–192 (cit. on p. 31).

Rusz, J., & Tykalova, T. (2021). Reader response: Motor speech patterns in huntington disease (cit. on p. 31).

Longworth, C., Keenan, S., Barker, R., Marslen-Wilson, W., & Tyler, L. (2005). The basal ganglia and rule-governed language use: Evidence from vascular and degenerative conditions. *Brain*, *128*(3), 584–596 (cit. on p. 31).

Hertrich, I., & Ackermann, H. (1994). Acoustic analysis of speech timing in huntington' s disease. *Brain and language*, *47*(2), 182–196 (cit. on pp. 34, 154).

Skodda, S., Grönheit, W., Lukas, C., Bellenberg, B., von Hein, S. M., Hoffmann, R., & Saft, C. (2016). Two different phenomena in basic motor speech performance in premanifest huntington disease. *Neurology*, *86*(14), 1329–1335 (cit. on p. 34).

Rusz, J., Tykalova, T., Ramig, L. O., & Tripoliti, E. (2021). Guidelines for speech recording and acoustic analyses in dysarthrias of movement disorders. *Movement Disorders*, *36*(4), 803–814 (cit. on pp. 34, 119).

Novotny, M., Rusz, J., Cmejla, R., Ruzickova, H., & Klempir, E., Jiriand Ruzicka. (2016). Hypernasality associated with basal ganglia dysfunction: Evidence from parkinson's disease and huntington's disease. *PeerJ*, *4*, e2530 (cit. on p. 34).

Chan, J., Stout, J. C., Shirbin, C. A., & Vogel, A. P. (2021). Listener detection of objectively validated acoustic features of speech in huntington's disease. *Journal of Huntington's Disease*, (Preprint), 1–9 (cit. on pp. 34, 154).

Ghio, A., Pouchoulin, G., Teston, B., Pinto, S., Fredouille, C., De Looze, C., Robert, D., Viallet, F., & Giovanni, A. (2012). How to manage sound, physiological and clinical data of 2500 dysphonic and dysarthric speakers? *Speech Communication*, *54*(5), 664–679 (cit. on p. 37).

Strauss, E., Sherman, E. M., Spreen, O., et al. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary*. American Chemical Society. (Cit. on p. 40).

Messick, S. (1995). Standards of validity and the validity of standards in performance asessment. *Educational measurement: Issues and practice*, *14*(4), 5–8 (cit. on p. 40).

Bechtel, N., Scahill, R., Rosas, H. D., Acharya, T., van den Bogaard, S. J., Jauffret, C., Say, M. J., Sturrock, A., Johnson, H., Onorato, C. E., et al. (2010). Tapping linked to function and structure in premanifest and symptomatic huntington disease. *Neurology*, *75*(24), 2150–2160 (cit. on p. 41).

Barker, R. A., Mason, S. L., Harrower, T. P., Swain, R. A., Ho, A. K., Sahakian, B. J., Mathur, R., Elneil, S., Thornton, S., Hurrelbrink, C., et al. (2013). The long-term safety and efficacy of bilateral transplantation of human fetal striatal tissue in patients with mild to moderate huntington's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, *84*(6), 657–665 (cit. on p. 42).

Kingma, E. M., van Duijn, E., Timman, R., van der Mast, R. C., & Roos, R. A. (2008). Behavioural problems in huntington's disease using the problem behaviours assessment. *General hospital psychiatry*, *30*(2), 155–161 (cit. on p. 42).

Craufurd, D., Thompson, J. C., & Snowden, J. S. (2001). Behavioral changes in huntington disease. *Cognitive and Behavioral Neurology*, *14*(4), 219–226 (cit. on p. 42).

Dorsey, E. R., Glidden, A. M., Holloway, M. R., Birbeck, G. L., & Schwamm, L. H. (2018). Teleneurology and mobile technologies: The future of neurological care. *Nature Reviews Neurology*, *14*(5), 285–297 (cit. on p. 47).

Rusz, J., Svihlik, J., Kryze, P., Novotny, M., & Tykalova, T. (2021). Reproducibility of voice analysis with machine learning. *Movement Disorders*, *36*(5), 1282–1283 (cit. on p. 47).

Zhan, A., Mohan, S., Tarolli, C., Schneider, R. B., Adams, J. L., Sharma, S., Elson, M. J., Spear, K. L., Glidden, A. M., Little, M. A., et al. (2018). Using smartphones and machine learning to quantify parkinson disease severity: The mobile parkinson disease score. *JAMA neurology*, *75*(7), 876–880 (cit. on p. 47).

Rusz, J., Cmejla, R., Tykalova, T., Ruzickova, H., Klempir, J., Majerova, V., Picmausova, J., Roth, J., & Ruzicka, E. (2013). Imprecise vowel articulation as a potential early marker of parkinson's disease: Effect of speaking task. *The Journal of the Acoustical Society of America*, *134*(3), 2171–2181 (cit. on pp. 47, 119, 120, 153, 154).

Tsanas, A., Little, M. A., McSharry, P. E., Spielman, J., & Ramig, L. O. (2012). Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease. *IEEE transactions on biomedical engineering*, *59*(5), 1264–1271 (cit. on p. 47).

Bot, B. M., Suver, C., Neto, E. C., Kellen, M., Klein, A., Bare, C., Doerr, M., Pratap, A., Wilbanks, J., Dorsey, E. R., et al. (2016). The mpower study, parkinson disease mobile data collected using researchkit. *Scientific data*, *3*(1), 1–9 (cit. on p. 47).

Vogel, A. P., Saft, C., Churchyard, A., & Stout, J. (2016). Two different phenomena in basic motor speech performance in premanifest huntington diseaseauthor response. *Neurology*, *87*(21), 2283–2283 (cit. on p. 47).

Lo, J., Reyes, A., Pulverenti, T. S., Rankin, T. J., Bartlett, D. M., Zaenker, P., Rowe, G., Feindel, K., Poudel, G., Georgiou-Karistianis, N., et al. (2020). Dual tasking impairments are associated with striatal pathology in huntington's disease. *Annals of clinical and translational neurology*, *7*(9), 1608–1619 (cit. on p. 48).

Mayr, U., & Keele, S. W. (2000). Changing internal constraints on action: The role of backward inhibition. *Journal of Experimental Psychology: General*, *129*(1), 4 (cit. on p. 48).

Sacks, O. (1998). *The man who mistook his wife for a hat: And other clinical tales*. Simon & Schuster. (Cit. on p. 48).

Kaplan, E., Goodglass, H., Weintraub, S., et al. (2001). Boston naming test (cit. on pp. 48, 50).

Boschi, V., Catricala, E., Consonni, M., Chesi, C., Moro, A., & Cappa, S. F. (2017). Connected speech in neurodegenerative language disorders: A review. *Frontiers in psychology*, *8*, 269 (cit. on pp. 48, 49).

Wilson, S. M., Henry, M. L., Besbris, M., Ogar, J. M., Dronkers, N. F., Jarrold, W., Miller, B. L., & Gorno-Tempini, M. L. (2010). Connected speech production in three variants of primary progressive aphasia. *Brain*, *133*(7), 2069–2088 (cit. on p. 48).

Sajjadi, S. A., Patterson, K., Tomek, M., & Nestor, P. J. (2012). Abnormalities of connected speech in semantic dementia vs alzheimer's disease. *Aphasiology*, *26*(6), 847–866 (cit. on p. 49).

Alexander, K. J., Miller, P. J., & Hengst, J. A. (2001). Young children's emotional attachments to stories. *Social Development*, *10*(3), 374–398 (cit. on p. 49).

Symons, D. K., Peterson, C. C., Slaughter, V., Roche, J., & Doyle, E. (2005). Theory of mind and mental state discourse during book reading and story-telling tasks. *British journal of developmental psychology*, *23*(1), 81–102 (cit. on p. 49).

Kenan, N., Zachor, D. A., Watson, L. R., & Ben-Itzchak, E. (2019). Semantic-pragmatic impairment in the narratives of children with autism spectrum disorders. *Frontiers in psychology*, *10*, 2756 (cit. on p. 49).

MacWhinney, B. (2019). Understanding spoken language through talkbank. *Behavior research methods*, *51*(4), 1919–1927 (cit. on pp. 49, 55, 56).

Ash, S., Menaged, A., Olm, C., McMillan, C. T., Boller, A., Irwin, D. J., McCluskey, L., Elman, L., & Grossman, M. (2014). Narrative discourse deficits in amyotrophic lateral sclerosis. *Neurology*, *83*(6), 520–528 (cit. on p. 49).

Ekman, P. (1999). Basic emotions. *Handbook of cognition and emotion*, *98*(45-60), 16 (cit. on p. 49).

Guez, A., Peyre, H., Williams, C., Labouret, G., & Ramus, F. (2021). The epidemiology of cognitive development. *Cognition*, 104690 (cit. on p. 49).

Eriksen, H.-L. F., Kesmodel, U. S., Underbjerg, M., Kilburn, T. R., Bertrand, J., & Mortensen, E. L. (2013). Predictors of intelligence at the age of 5: Family, pregnancy and birth characteristics, postnatal influences, and postnatal growth. *PloS one*, *8*(11), e79200 (cit. on p. 49).

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., et al. (2019). Do no harm: A roadmap for responsible machine learning for health care. *Nature medicine*, *25*(9), 1337–1340 (cit. on pp. 49, 50).

Culnane, C., Rubinstein, B. I., & Teague, V. (2017). Health data in an open world. *arXiv preprint arXiv:1712.05627* (cit. on p. 52).

Goldwater, S., Jurafsky, D., & Manning, C. D. (2010). Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, *52*(3), 181–200 (cit. on pp. 54, 88).

Riviere, M., Copet, J., & Synnaeve, G. (2021). Asr4real: An extended benchmark for speech models. *arXiv preprint arXiv:2110.08583* (cit. on p. 54).

Rudzicz, F., Namasivayam, A. K., & Wolff, T. (2012). The torgo database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, *46*(4), 523–541 (cit. on p. 54).

Christensen, H., Cunningham, S., Fox, C., Green, P., & Hain, T. (2012). A comparative study of adaptive, automatic recognition of disordered speech. *Thirteenth Annual Conference of the International Speech Communication Association* (cit. on p. 54).

van Gompel, M., & Reynaert, M. (2014). Folia: A practical xml format for linguistic annotation - a descriptive and comparative study. *CLIN 2014* (cit. on p. 55).

Carletta, J., Evert, S., Heid, U., & Kilgour, J. (2005). The nite xml toolkit: Data model and query language. *Language Resources and Evaluation*, *39*, 313–334 (cit. on p. 55).

Wang, I., Pelletier, A., Antoine, J.-Y., & Halftermeyer, A. (2020). Odil_syntax: A free spontaneous spoken french treebank annotated with constituent trees. *Proceedings of the 12th Language Resources and Evaluation Conference*, 5301–5307 (cit. on p. 55).

Blanche-Benveniste, C., Bilger, M., Rouget, C., Van Den Eynde, K., Mertens, P., & Willems, D. (1990). Le français parlé(études grammaticales). *Sciences du langage* (cit. on p. 55).

Gibbon, D., Moore, R., & Winski, R. (1997). *Handbook of standards and resources for spoken language systems*. Walter de Gruyter. (Cit. on p. 59).

Bernard, M., & Titeux, H. (2021). Phonemizer: Text to phones transcription for multiple languages in python. *Journal of Open Source Software*, *6*(68), 3958 (cit. on p. 61).

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications. (Cit. on p. 62).

Mathet, Y., Widlocher, A., & Metivier, J.-P. (2015). The unified and holistic method gamma for inter-annotator agreement measure and alignment. *Computational Linguistics*, *41*(3), 437–479 (cit. on pp. 63, 71).

Tanno, R., Saeedi, A., Sankaranarayanan, S., Alexander, D. C., & Silberman, N. (2019). Learning from noisy labels by regularized estimation of annotator confusion. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11244–11253 (cit. on p. 71).

Zue, V., Seneff, S., & Glass, J. (1990). Speech database development at mit: Timit and beyond. *Speech communication*, *9*(4), 351–356 (cit. on p. 71).

Boersma, P. et al. (2002). Praat, a system for doing phonetics by computer. *Glot international*, *5* (cit. on p. 72).

Clark, H. H. (2002). Speaking in time. *Speech communication*, *36*(1-2), 5–13 (cit. on p. 80).

Levelt, W. J. (1983). Monitoring and self-repair in speech. *Cognition*, *14*(1), 41–104 (cit. on p. 80).

Clark, H. H., & Tree, J. E. F. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73–111 (cit. on pp. 81, 82).

Dingemanse, M., Torreira, F., & Enfield, N. J. (2013). Is "huh?" a universal word? conversational infrastructure and the convergent evolution of linguistic items. *PloS one*, *8*(11), e78273 (cit. on p. 81).

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291 (cit. on p. 81).

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, *20*(11), 818–829 (cit. on p. 81).

Grice, P. (2020). *Meaning/bedeutung (englisch/deutsch): Great papers philosophie*. Reclam Verlag. (Cit. on p. 81).

Amir, O., Shapira, Y., Mick, L., & Yaruss, J. S. (2018). The speech efficiency score (ses): A time-domain measure of speech fluency. *Journal of fluency disorders*, *58*, 61–69 (cit. on pp. 82, 83, 139).

Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., & Vinyals, O. (2012). Speaker diarization: A review of recent research. *IEEE Transactions on audio, speech, and language processing*, *20*(2), 356–370 (cit. on p. 88).

Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., & Narayanan, S. (2022). A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, *72*, 101317 (cit. on p. 88).

Barker, J., Watanabe, S., Vincent, E., & Trmal, J. (2018). The fifth'chime'speech separation and recognition challenge: Dataset, task and baselines. *Interspeech 2018-19th Annual Conference of the International Speech Communication Association* (cit. on p. 88).

Ryant, N., Singh, P., Krishnamohan, V., Varma, R., Church, K., Cieri, C., Du, J., Ganapathy, S., & Liberman, M. (2020). The third dihard diarization challenge. *arXiv preprint arXiv:2012.01477* (cit. on p. 88).

Bernstein, R. N., & MacWhinney, B. (2018). Fluency bank: A new resource for fluency research and practice. *Journal of fluency disorders*, *56*, 69 (cit. on pp. 88, 96).

Ferguson, J., Durrett, G., & Klein, D. (2015). Disfluency detection with a semi-markov model and prosodic features. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 257–262 (cit. on p. 95).

Zayats, V., Ostendorf, M., & Hajishirzi, H. (2016). Disfluency detection using a bidirectional lstm. *arXiv preprint arXiv:1604.03209* (cit. on p. 95).

Riad, R., Bachoud-Lévi, A.-C., Rudzicz, F., & Dupoux, E. (2020b). Identification of primary and collateral tracks in stuttered speech. *Proceedings of the 12th Language Resources and Evaluation Conference*, 1681–1688 (cit. on p. 96).

Bredin, H. (2017). pyannote.metrics: a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems. *Interspeech* (cit. on p. 105).

Ravanelli, M., & Bengio, Y. (2018). Speaker recognition from raw waveform with sincnet. *Spoken Language Technology Workshop (SLT)*, 1021–1028 (cit. on p. 106).

Oue, S., Marxer, R., & Rudzicz, F. (2015). Automatic dysfluency detection in dysarthric speech using deep belief networks. *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, 60–64 (cit. on p. 106).

Obozinski, G., Taskar, B., & Jordan, M. (2006). Multi-task feature selection (cit. on p. 116).

Audibert, N., & Fougeron, C. (2012). Distorsions de l'espace vocalique: Quelles mesures? application à la dysarthrie (distortions of vocalic space: Which measurements? an application to dysarthria.)[in french]. *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 1: JEP*, 217–224 (cit. on pp. 119, 120).

Huet, K., & Harmegnies, B. (2000). Contribution à la quantification du degré d'organisation des systèmes vocaliques. *Actes des JEP'2000*, 225–228 (cit. on p. 120).

Roy, N., Nissen, S. L., Dromey, C., & Sapir, S. (2009). Articulatory changes in muscle tension dysphonia: Evidence of vowel space expansion following manual circumlaryngeal therapy. *Journal of communication disorders*, *42*(2), 124–135 (cit. on p. 120).

Oosterloo, M., Craufurd, D., Nijsten, H., & van Duijn, E. (2019). Obsessive-compulsive and perseverative behaviors in huntington's disease. *Journal of Huntington's disease*, *8*(1), 1–7 (cit. on p. 122).

Cohen, L., & Dehaene, S. (1998). Competition between past and present. assessment and interpretation of verbal perseverations. *Brain: a journal of neurology*, *121*(9), 1641–1659 (cit. on pp. 122, 138, 139).

Kordsachia, C., Labuschagne, I., & Stout, J. (2017). Beyond emotion recognition deficits: A theory guided analysis of emotion processing in Huntington's disease. *undefined* (cit. on p. 140).

Trinkler, I., de Langavant, L. C., & Bachoud-Lévi, A.-C. (2013). Joint recognition–expression impairment of facial emotions in huntington's disease despite intact understanding of feelings. *Cortex*, *49*(2), 549–558 (cit. on p. 140).

Kempler, D., & Van Lancker, D. (2002). Effect of speech task on intelligibility in dysarthria: A case study of parkinson's disease. *Brain and language*, *80*(3), 449–464 (cit. on p. 153).

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and speech*, *44*(2), 123–147 (cit. on p. 153).

Lindström, A., Villing, J., Larsson, S., Seward, A., Åberg, N., & Holtelius, C. (2008). The effect of cognitive load on disfluencies during in-vehicle spoken dialogue. *Ninth Annual Conference of the International Speech Communication Association* (cit. on p. 153).

Lu, L., Kong, L., Dyer, C., Smith, N. A., & Renals, S. (2016). Segmental recurrent neural networks for end-to-end speech recognition. *Interspeech 2016*, 385–389 (cit. on p. 154).

Wei, K., Li, J., Ding, M., Ma, C., Yang, H. H., Farokhi, F., Jin, S., Quek, T. Q., & Poor, H. V. (2020). Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, *15*, 3454–3469 (cit. on p. 155).

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2019). Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977* (cit. on p. 155).

Sun, J. J., Kennedy, A., Zhan, E., Anderson, D. J., Yue, Y., & Perona, P. (2021). Task programming: Learning data efficient behavior representations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2876–2885 (cit. on p. 155).

Millet, J., & Zeghidour, N. (2019). Learning to detect dysarthria from raw speech. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5831–5835 (cit. on p. 155).

Powers, R., Etezadi-Amoli, M., Arnold, E. M., Kianian, S., Mance, I., Gibiansky, M., Trietsch, D., Alvarado, A. S., Kretlow, J. D., Herrington, T. M., et al. (2021). Smartwatch inertial sensors continuously monitor real-world motor fluctuations in parkinson's disease. *Science translational medicine*, *13*(579) (cit. on p. 156).

Vahia, I. V., Kabelac, Z., Hsu, C.-Y., Forester, B. P., Monette, P., May, R., Hobbs, K., Munir, U., Hoti, K., & Katabi, D. (2020). Radio signal sensing and signal processing to monitor behavioral symptoms in dementia: A case study. *The American Journal of Geriatric Psychiatry*, *28*(8), 820–825 (cit. on p. 156).

Xu, C., Yang, Y., & Gao, W. (2020). Skin-interfaced sensors in digital medicine: From materials to applications. *Matter*, *2*(6), 1414–1445 (cit. on p. 156).

Appendix

A

> *"You need both. You need people who focus very very hard on one very narrow problem working for years, become a very deep expert. But them you need the people who can connect things in fragmented fields. I make my living [...] by understanding one field X and taking some ideas from that and applying it to field Y. But I couldn't do that if there weren't people who are very deeply working in field X, and so forth.".*
>
> **— Terence Tao**
> Professor of Mathematics at UCLA

A.1 Additional information concerning the BasalVoice database

A.2 Pre-registration for vocal and linguistic markers studies

Here, we include the pre-registration for different studies introduced in the section 4. We used the pre-registration methodology from aspredicted.org, which have a closed set of questions to frame pre-registration.

CONFIDENTIAL - FOR PEER-REVIEW ONLY

PREDICT CLINICAL MARKERS FROM SPEECH AND LANGUAGE -- Paris, Aug. 2020 (#46794)

Created: 08/27/2020 07:06 AM (PT)

1) Have any data been collected for this study already?
It's complicated. We have already collected some data but explain in Question 8 why readers may consider this a valid pre-registration nevertheless.

2) What's the main question being asked or hypothesis being tested in this study?
Are speech features able to provide and predict sensitive clinical measures for other functions (cognitive, motor, and functional) in Huntington's Disease?

3) Describe the key dependent variable(s) specifying how they will be measured.
We collect 60 variables from speech recordings of the recitation onward and backward of numbers. We measure the: timing of the speech units and vocalisations, the fundamental frequency, speech intelligibility, perseverations, the paraphasias, vocal noises, and the counting errors statistics.
We collect 8 clinical variables to measure the motor, cognitive symptoms and the functional capacity of the patient: The UHDRS Motor Score, Symbol Digit Modalities Test (SDMT), Fluency test 1-minute, the 3 components of the Stroop test (word, colour, and interference), Total Functional Capacity (TFC), UHDRS Independence scale. We also calculate and predict the composite score cUHDRS.

4) How many and which conditions will participants be assigned to?
There is no condition for the included individuals. We included data points where audio recordings were available.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
We will split the data into two data sets: training (80%) for fitting and the development of the various models, and test (20%) for model testing. The splitting of data will take into account the Centres. The number of CAG triplets, Age, Sex, Education level, and the Disease Burden Score will be included as input predictive values to the models. The test set will only contain individuals from the Confirmation cohort, not from the Exploration cohort. The train set will contain individuals from both cohorts. We will conduct 50 repeated learning-testing to obtain good estimates of the performances.
We will use the auto-machine-learning toolkit, auto-sklearn to predict the clinical outcomes based on the speech and language markers. We will use a nested K-fold on the training set to select the best model and the best set of variables for predictions. The final model will be evaluated on the test set with the Coefficient of determination R2, Mean Absolute Error and range-based Normalized Mean Absolute Error. Specific performances on pre-symptomatic and symptomatic patients will be analysed.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We exclude observations when ambient noises preclude analyses of speech. Level of noise is determined perceptually by the speech therapists, before delivering data for analysis.

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
Data are collected for individuals with Huntington's disease, carrying the genetic mutation (pre-symptomatic and symptomatic carriers). Data collection was completed in March 2020 and linguistic annotations will be delivered in September 2020 for analysis.
The first cohort, the Exploration cohort, MIG-HD (NCT00190450) of patients (N=44) have done the protocol in 6 different centres in France and Belgium. The first cohort helped us design the speech and language features. The second cohort, Validation cohort, pooled together gene carriers from BioHD (NCT01412125) and RepairHD (NCT03119246) who have performed the speech protocol (N=68) in 1 Centre in France. The number of samples to estimate predictive performance needed to be at least 100.

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
Even though the audio samples have been collected, the annotations of audio of the second cohort, validation cohort, are not finished and delivered. We will measure correlations between speech markers and the striatum.

CONFIDENTIAL - FOR PEER-REVIEW ONLY

Expression of vocal emotions impairment in Huntington's Disease. (#59319)

Created: 02/25/2021 07:00 AM (PT)

1) Have any data been collected for this study already?
It's complicated. We have already collected some data but explain in Question 8 why readers may consider this a valid pre-registration nevertheless.

2) What's the main question being asked or hypothesis being tested in this study?
Is vocal expression of emotion impaired in Huntington disease (HD)? We hypothesize that emotions are harder to identify in HD patients' speech than in healthy individuals' speech.

3) Describe the key dependent variable(s) specifying how they will be measured.
The dataset will be split into 4 groups, based on the disease status of the speaker: healthy, pre-symptomatic, symptomatic early stage (total functional capacity score >=11), symptomatic late stage (total functional capacity score <11). We will train four speech emotion recognition (SER) algorithms, each on one of these four groups.
We will use random forests as the SER algorithm. To perform this algorithm we will use scikit learn. The algorithm will take as input features extracted from the data, as well as the sex of the speaker as a metadata. The features used will be the extended GeMAPS [1], and they will be extracted with the python openSMILE package [2]. We will also add the duration of each audio extract in seconds.
We will compute the accuracies of the models with 10-fold nested cross validation. The hyperparameter space used will be: n_estimators={1,10,100,200,500}, min_samples_split={2,5,10}, min_samples_leaf={1,2,4}. All the other parameters will be fixed to their default values.

[1] F. Eyben et al., "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," in IEEE Transactions on Affective Computing, April-June 2016,
[2] F. Eyben et al., 2010c Eyben, F., Wöllmer, M., Schuller, B., 2010c. openSMILE – the munich versatile and fast open-source audio feature Extractor. In: Proc. ACM Multimedia, Florence, Italy

4) How many and which conditions will participants be assigned to?
There is no condition for the included gene-carrier Huntington's Disease individuals. We included data points where audio recordings were available.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
To assess the difference between the four groups' accuracies' we will perform a Kruskal-Wallis test by resampling. If the null hypothesis is rejected, we will conduct post hoc Wilcoxon-Mann-Whitney tests (with Bonferroni corrections) by resampling.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We exclude observations when ambient noises preclude proper analyses of speech. Level of noise is determined perceptually by the speech pathologists, before delivering the data for analysis. We will exclude every audio segment that does not contain linguistic content (i.e. single filled pauses and single vocal noises are not used).

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
Data are collected for individuals with Huntington's disease, carrying the genetic mutation (pre-symptomatic and symptomatic carriers). Data collection was completed in March 2020 and linguistic annotations will be delivered in March 2021 for analysis. We will use data from at least 15 different speakers for each emotion considered (happiness, fear, anger, neutral).

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
The dataset we will use has already been collected, but the annotations of emotions content is not delivered and was not analyzed so far.
We will run an additional analysis with the same method but with 3 data groups, by merging the two symptomatic groups.
We will perform an ablation study of the features, to examine the role of each dimension (articulation, phonation, prosody, respiration) of the speech production of emotions. We will also perform an ablation study on the size of the training set. We will repeat the experiment removing the first 10% of each patients' interview to cope with a possible delay in emotion elicitation. To control this additional experiment we will also replicate the analysis removing randomly 5% to 100% of each interview.
We will replicate the statistical analysis done in the study [3], using the speech features extracted in our experiment.

[3] Alhinti L, Christensen H, Cunningham S. Acoustic differences in emotional speech of people with dysarthria. Speech Communication,. 2021

CONFIDENTIAL - FOR PEER-REVIEW ONLY

Linguistic expression of emotions in Huntington's Disease. (#63189)

Created: 04/12/2021 04:24 AM (PT)

1) Have any data been collected for this study already?
It's complicated. We have already collected some data but explain in Question 8 why readers may consider this a valid pre-registration nevertheless.

2) What's the main question being asked or hypothesis being tested in this study?
Is linguistic expression of emotions impaired in Huntington disease (HD)?
We hypothesize that emotions are less identifiable in HD patients' speech linguistic content than in controls' speech linguistic content.

3) Describe the key dependent variable(s) specifying how they will be measured.
The dataset is composed of speech annotations with emotional labels as well as information regarding the speaker's demographics and clinical conditions. The dataset will be split into 4 groups, based on the disease status of the speaker: healthy, pre-symptomatic, symptomatic early stage (total functional capacity score >=11), symptomatic late stage (total functional capacity score <11). Subgroups will be balanced such that they all have the same size and same composition emotion-wise. We will train four emotion recognition models based on natural language processing pipelines. This process will be repeated for each sub-group.
The machine learning models will take as inputs sentences embeddings obtained with the LASER library. LASER is a library that embeds sentences in fixed-size representation. We will use random forests to classify each emotion (implemented with scikit learn). We will compute the accuracies of the models with a 10-fold nested cross validation schema. The hyperparameter space will be: number of decision trees (1,10,100,200,500), the minimum number of samples to split an internal node (2,5,10), the minimum number of samples at a leaf node (1,2,4). All the other parameters will be fixed to their default value.

4) How many and which conditions will participants be assigned to?
There is no condition for the inclusion of gene-carrier Huntington's Disease individuals in the cohort (pre-symptomatic and symptomatics will be included). We will include data points where textual annotations were available.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
To assess the difference between the four groups' accuracies' we will perform a Kruskal-Wallis test based on resampling. If the null hypothesis (equality of the medians) is rejected, we will conduct post hoc analyses with Wilcoxon-Mann-Whitney tests (with Bonferroni corrections) based on resampling.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We exclude observations when ambient noises preclude proper analyses of speech. Level of noise is determined perceptually by the speech pathologists, before delivering the data for analysis. We will normalize textual annotation such that it contains only linguistic content (i.e. single filled pauses and vocal noises are removed), and remove phonetic distortions, stuttering, blocks and prolongations

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
Data are collected for individuals with Huntington's disease, carrying the genetic mutation (pre-symptomatic and symptomatic carriers). Data collection was completed in March 2020 and linguistic annotations will be delivered in March 2021 for analysis. We will use data from at least 15 different speakers for each emotion considered (happiness, fear, anger, neutral).

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
The dataset we will use has already been collected, but the annotations of emotions content is not delivered and was not analyzed so far.
We will run the following additional analysis:
-main experiment with 3 data groups (early and late symptomatics merged)
-main experiment with alternative classifiers: linear discriminant analysis, gradient boosting
-main experiment with the set of features used in ([1]) except for the acoustic features
-main experiment with first 10% of each interview removed
-main experiment with 10% of each interview randomly removed
-ablation study on the features
-ablation study on the size of the training set

[1] Kathleen C Fraser, Jed A Meltzer et Frank Rudzicz. "Linguistic features identify Alzheimer's disease in narrative speech". (2016)

A.3 Learning spectro-temporal representations of complex sounds with parameterized neural networks

JASA ARTICLE

Learning spectro-temporal representations of complex sounds with parameterized neural networks[a)]

Rachid Riad,[1,b)] Julien Karadayi,[1] Anne-Catherine Bachoud-Lévi,[2] and Emmanuel Dupoux[1,c)]

[1]*Ecole des Hautes Etudes en Sciences Sociales, CNRS, Institut National de Recherche informatique et Automatique, Département d'Études Cognitives, Ecole Normale Supérieure-Paris Sciences et Lettres University, 29 Rue d'Ulm, 75005 Paris, France*

[2]*NeuroPsychologie Interventionnelle, Département d'Études Cognitives, Ecole Normale Supérieure, Institut National de la Santé et de la Recherche Médicale, Institut Mondor de Recherche Biomédicale, Neuratris, Université Paris-Est Créteil, Paris Sciences et Lettres University, 29 Rue d'Ulm, 75005 Paris, France*

ABSTRACT:

Deep learning models have become potential candidates for auditory neuroscience research, thanks to their recent successes in a variety of auditory tasks, yet these models often lack interpretability to fully understand the exact computations that have been performed. Here, we proposed a parametrized neural network layer, which computes specific spectro-temporal modulations based on Gabor filters [learnable spectro-temporal filters (STRFs)] and is fully interpretable. We evaluated this layer on speech activity detection, speaker verification, urban sound classification, and zebra finch call type classification. We found that models based on learnable STRFs are on par for all tasks with state-of-the-art and obtain the best performance for speech activity detection. As this layer remains a Gabor filter, it is fully interpretable. Thus, we used quantitative measures to describe distribution of the learned spectro-temporal modulations. Filters adapted to each task and focused mostly on low temporal and spectral modulations. The analyses show that the filters learned on human speech have similar spectro-temporal parameters as the ones measured directly in the human auditory cortex. Finally, we observed that the tasks organized in a meaningful way: the human vocalization tasks closer to each other and bird vocalizations far away from human vocalizations and urban sounds tasks. https://doi.org/10.1121/10.0005482

(Received 18 February 2021; revised 4 June 2021; accepted 8 June 2021; published online 14 July 2021)

[Editor: Marie Roch] Pages: 353–366

I. INTRODUCTION

The main objective of auditory neuroscience is to build models that can predict both the brain neural responses to relevant sounds and the behaviors associated with these responses (Kell and McDermott, 2019; Pillow and Sahani, 2019). While most of the auditory neuroscience research has focused on the neural side, there is growing recognition of the importance of also matching the performance of living organisms on a variety of behavioral tasks (Yarkoni and Westfall, 2017). In recent years, major progress has been achieved with deep neural networks (DNNs), which, after training with supervised classification objectives on large datasets, proved able to perform near human performance on a variety of audio tasks, such as automatic speech recognition (Amodei *et al.*, 2016), speaker verification (Snyder *et al.*, 2018), or audio scene classification (Salamon and Bello, 2017). These trained systems therefore become potential candidate models for auditory neuroscience (Koumura *et al.*, 2019) and have already started to be used to account for perceptual results (Saddler *et al.*, 2020) and brain data (Kell *et al.*, 2018) in humans.

DNN models typically take as input a spectral representation [although some new trends consist in sidestepping this representation and work directly from the raw waveform (Ravanelli and Bengio, 2018; Zeghidour *et al.*, 2018)]. Working from a spectral representation has biological plausibility, since it matches approximately what we know about the first stage of auditory processing (Stevens *et al.*, 1937). However, DNN models are less biologically motivated regarding the next steps. Most of them use rather generic connectivity patterns (fully connected, convolutional, or recurrent networks), which, while being very powerful in learning task-specific representations from an engineering point of view, lack both interpretability and support in the auditory neuroscience. To further the understanding of both the artificial and real neural networks, there have been some attempts to decode the representation extracted from biological measurements or computed by deep learning models (Ondel *et al.*, 2019; Thoret *et al.*, 2020). Even though these methods allow uncovering the important aspects of the stimuli, they rely on simplifying hypotheses [linearity of the responses, independence across neurons (Meyer *et al.*, 2017; Shamma, 1996)], and they do not provide an in depth explanation of how the DNNs made their decisions.

Fortunately, the stages beyond the extraction of the acoustic spectrum have been studied over the past few years with novel understanding of the representations and

[a)]This paper is part of a special issue on Machine Learning in Acoustics.

[b)]Also at: NeuroPsychologie Interventionnelle, Ecole Normale Supérieure, 75005 Paris, France. Electronic mail: riadrachid3@gmail.com, ORCID: 0000-0002-7753-1219.

[c)]Also at: Facebook AI Research, Paris, France.

processing involved (McDermott, 2018). Slow spectral and temporal modulation built on top of the spectrum have been shown in psychophysical tests to be useful for several audio tasks solved by mammals: they contribute to speech intelligibility (Edraki *et al.*, 2019; Elhilali *et al.*, 2003; Elliott and Theunissen, 2009), and they help to boost performance for speech processing in noisy environments (Chang and Morgan, 2014; Mesgarani *et al.*, 2006; Vuong *et al.*, 2020). In addition, the responses to such spectral and temporal modulations of natural sounds can be decoded from human functional magnetic resonance imaging (fMRI) (Santoro *et al.*, 2017) and have been measured directly with invasive techniques in ferrets (Depireux *et al.*, 2001), in birds (Woolley *et al.*, 2005), and also in the human brain (Hullett *et al.*, 2016).

Analytic models (time-frequency analysis) of these modulations in the spectrogram have been proposed (Chang and Morgan, 2014; Chi *et al.*, 2005; Ezzat *et al.*, 2007; Schädler *et al.*, 2012), for instance, with a 2D discrete wavelet decomposition of the spectrogram. The idea is that on top of the spectrum, spectro-temporal wavelets (such as Morlet/Gabor wavelets) can be defined that drive both behavioral responses and brain signals. The problem of such analytic models is that they only propose a potentially very large representation space and provide no method to select which Gabor patch is relevant for which task. But analyses of brain signals show that the responses from the auditory cortex are not fixed, but vary depending on the task at hand (Francis *et al.*, 2018; Fritz *et al.*, 2003; Jääskeläinen *et al.*, 2007). Therefore, what is needed is a model that can learn the characteristics of the spectro-temporal representations that are relevant to the task.

This is the goal of this work. We introduce a parametrized neural network that explicitly represents spectro-temporal filters (STRFs), but whose parameters are differentiable and can therefore be tuned to each particular task. There are two advantages of this approach, as illustrated in Fig. 1. First, as analytic models, and contrary to standard DNN models, this model is fully interpretable. The parameters of each filter can be directly read off the model and compared to physiological or neural data (Fig. 2). Second, as DNNs, but contrary to analytic models, this model can be tuned to different tasks, accounting both for behavioral results and for the task-specificity of the brain representations. As a side issue, since the model is constrained and has few parameters, it has the potential to explain the perceptual learning aspect of plasticity with a lot less training data than typically used in generic DNNs. Therefore, the model makes direct and testable predictions about the auditory representation as a function of the task.

The paper is organized as follows: Sec. II presents the methods with our parametrized neural network model and the different ways to analyze the distribution of the learned spectro-temporal modulations; in Sec. III, we describe the experimental setup with the different computational tasks, data, state-of-the-art systems, and evaluations; Sec. IV presents the performance results, the analysis of the learned distribution of spectro-temporal modulation for each setup, and the discussion. Section V presents our conclusions and potential future work. To encourage reproducible research, the developed learnable STRF layer, the learned STRF modulations, and the recipes to replicate results are available in an open-source package.[1]

II. MODELS AND METHODS

A. Overview

To learn spectro-temporal representations of sounds, we constructed a learnable front-end model of natural sounds that stays interpretable. The model is composed of an initial fixed frequency analysis of sounds. Then this time-frequency representation of the sound is convoluted with a parametrized layer that controls the parameters of a set of Gabor filters (Sec. II B). We used this layer as a replacement of the first stage of processing in the different neural network architectures to solve each individual task. As this layer adapted to each task under study, our ability to directly read out the parameters helped us quantify what was being learned by the models (Sec. II C).

B. Learnable STRFs

Here, $\Re$, $\Im$, $|\cdot|$, and $[\cdot,\cdot]$ represent the real part, imaginary part, modulus, and concatenation operators, respectively. $\{\cdot\}$ represents a set and $card(\cdot)$ the cardinal of a specific set.

1. First stage of processing

The first audio processing step is the transformation of the audio signal from the time domain into the frequency domain $\mathbf{Y}(t,f)$. Each excerpt of sound given to the network is normalized before spectral analysis (Ulyanov *et al.*, 2016). We computed Mel-filterbanks by filtering sounds with a set of 64 bandpass filters cascaded with a log compression, to mimic the cochlear frequency analysis. All the sounds are sampled at 16 kHZ; thus, the center frequency of the filters spanned [0.0, 8000.0 Hz]. Frames are computed every 10 ms, with a Hamming window of 25 ms. The computations for the Mel-filterbanks of the audio can be performed on-the-fly directly on the graphics processing unit (GPU) thanks to Cheuk *et al.* (2020).

2. Definition of the learnable STRFs

The second step of front-end processing is a set of convolutions between the time-frequency representation of the audio and a set of Gabor filters (Gabor, 1946).

The two-dimensional (2D) Gabor filter kernel g_k is a sine-wave w_k modulated by a 2D Gaussian envelope s_k. Each Gabor filter g_k is expressed based on the set of parameters $(\sigma_t, \sigma_f, \gamma_k, F_k)$ in polar coordinates. We used the following formulation in this work:

$$g_k(t,f) = s_k(t,f) \cdot w_k(t,f), \tag{1a}$$

JASA
https://doi.org/10.1121/10.0005482

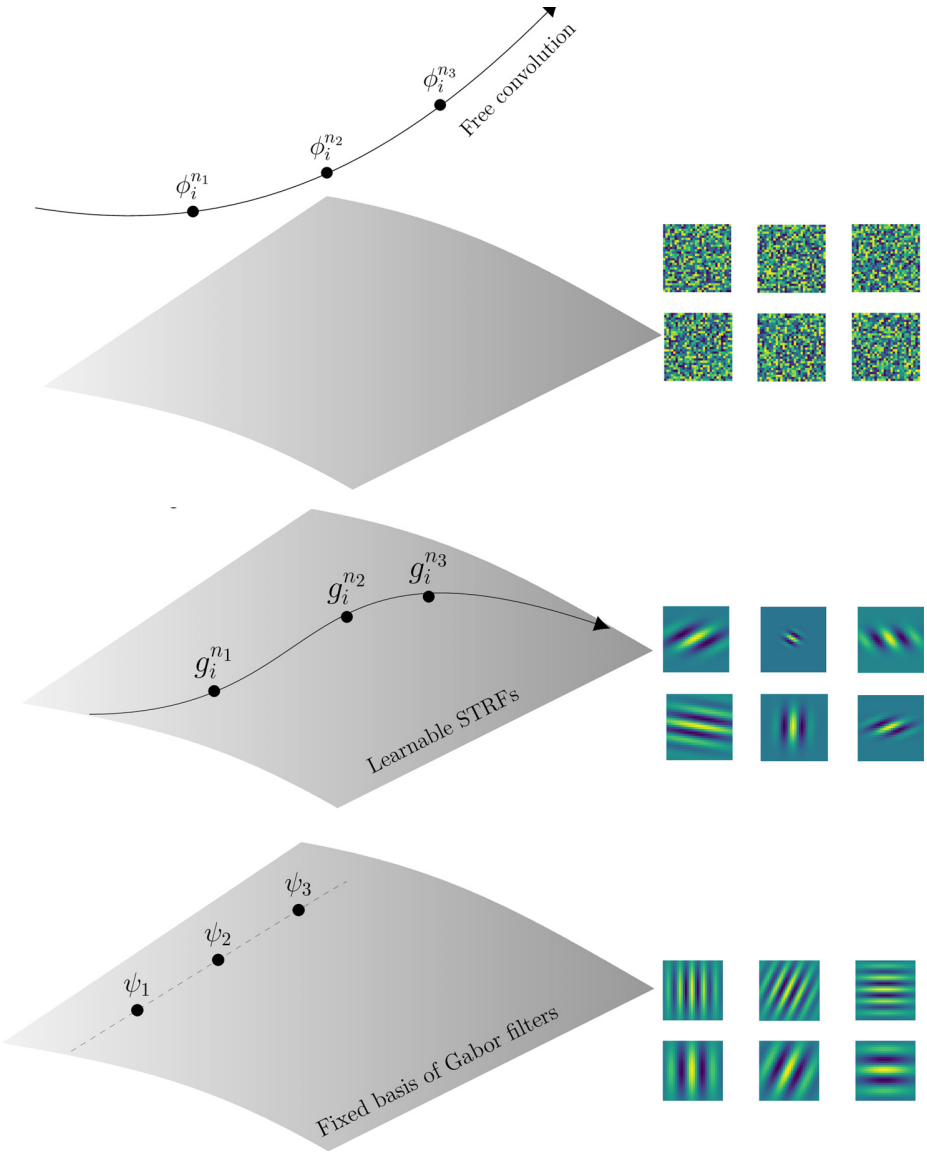


FIG. 1. (Color online) Schematic illustration of the different approaches to obtain spectro-temporal representations of sounds. The three dimensions represent all the functions that can be learned by a convolution. The parametric surface represents the space of Gabor functions. Top: Free learnable convolutions are unconstrained to move anywhere in the space of functions $(\phi_i^{n_k})$ (Młynarski and McDermott, 2018; Ondel *et al.*, 2019). The learned functions remain difficult to interpret. Middle: Learnable STRFs remain in the Gabor space of functions $(g_k^{n_i})$ (this study). The learned filters remain interpretable. Bottom: Fixed basis (ψ_i) are predefined by hand for each task. The 2D Gabor filterbanks are built with various scales and rotations but do not concentrate in specific modulations of interest (Bellur and Elhilali, 2015; Chang and Morgan, 2014; Elie and Theunissen, 2016; Mesgarani *et al.*, 2006; Schädler *et al.*, 2012). The upper index n_k and lower index i represent the n_k-step during learning for the filter indexed i.

$$s_k(t,f) = \frac{1}{2\pi\sigma_{t_k}\sigma_{f_k}} e^{-1/2\left(t^2/\sigma_{t_k}^2 + f^2/\sigma_{f_k}^2\right)}, \tag{1b}$$

$$w_k(t,f) = e^{j\left(2\pi\left(F_k R_{\gamma_k}\right)\right)}, \tag{1c}$$

$$R_{\gamma_k} = t\cos(\gamma_k) + f\sin(\gamma_k). \tag{1d}$$

We obtain a bank of N filters $\{g_k(t,f)\}_{k=0..N-1}$. This bank of filters is convolved with the time-frequency representation $\mathbf{Y}$ to obtain the 3D representation $\mathbf{Z}$,

$$\mathbf{Z}(t,f,k) = \sum_{u,v} \mathbf{Y}(u,v) g_k(t-u, f-v) \in \mathbb{C}. \tag{2}$$

These filters and their parameters can be used in 2D convolution neural networks (Alekseev and Bobe, 2019) in different ways. First, Gabor filters can be used as an initialization [free two-dimensional convolution Gabor initiation (free 2D conv. Gabor Init.)] of a 2D convolution neural network, and the 2D grid is tuned completely by backpropagation. In Chang and Morgan (2014), the authors compared the use of fixed Gabor features and the method free 2D conv. Gabor Init. (GCNN in their paper for automatic speech recognition). In our case, we also used Gabor filters for the learnable STRFs, but the gradient descent is only performed on the set of parameters $(\sigma_{t_k}, \sigma_{f_k}, F_k, \gamma_k)$. Indeed, all the operators to derive the Gabor filters are differentiable almost everywhere with respect to the parameters $(\sigma_{t_k}, \sigma_{f_k}, F_k, \gamma_k)$. The 2D grid instantiated by the Gabor filter used the parameters $(\sigma_{t_k}, \sigma_{f_k}, F_k, \gamma_k)$ in each cell; therefore, the gradients are summed over the 2D grid for each parameter.

Approaches such as Ezzat *et al.* (2007), Mesgarani *et al.* (2006), and Schädler *et al.* (2012) designed the Gabor filterbanks by hand for specific audio tasks. They designed each time a Gabor filterbank tailored by hand for each specific task under study (bottom panel of Fig. 1). Our learnable STRFs learned this Gabor filterbank (middle panel of Fig. 1).

On the other hand, Młynarski and McDermott (2018) and Ondel *et al.* (2019) did not use a prior and structure on

the convolution on top of the time-frequency representation. All the weights need to be learned, and the convolutions remain difficult to interpret (top panel of Fig. 1). This approach was evaluated with free two-dimensional convolution random initiation (free 2D conv. random Init.).

Images and spectrograms extracted from audio have very different properties. On one hand, images have important geometric variability due to perspective projections of 3D scenes under various viewpoints. On the other hand, classes of objects are invariant to image rotations (image rotation is an important data augmentation in computer vision).

These differences of modalities led us to add two main differences with our learnable STRF layer and the GaborNet (Alekseev and Bobe, 2019) introduced in computer vision. First, we introduced two different parameters (σ_t, σ_f) to give more freedom to the model in the temporal and spectral axes. Indeed, spectral and temporal axes play very different roles in the spectrogram. The second main difference is the shape of the receptive field. A 3×3 filter is limiting the computations of temporal modulations with an excerpt of 30 ms. Slow long-range modulations cannot be captured with a small receptive field, yet it was shown that different windows of integration play important parts in speech perception (Poeppel, 2003).

Therefore, each learnable STRF filter takes as input nine Mel-frequency bands and 1.1 s of context, thus yielding a size of 9×111 for each filter.

Finally, the output representation $\mathbf{Z}$ is in the complex domain $\mathbb{C}$. To be used by classic neural network architectures, we concatenated $[\Re(\mathbf{Z}), \Im(\mathbf{Z})]$ to obtain the representation to be fed to the rest of each network. We denoted this specific front-end by "learnable STRFs" in Tables I–V and denoted it by "learned STRFs" once we examined these representations.

C. Descriptive quantifiers of the distribution

We used quantitative measures to describe the structure of the distribution of the learned STRFs. These measures are used in auditory neuroscience to study the characteristics of the transfer function of biological neurons in several species: zebra finch (Depireux *et al.*, 2001; Theunissen *et al.*, 2000), monkeys (Massoudi *et al.*, 2015), and humans (Schönwiesner and Zatorre, 2009).

These measures (Singh and Theunissen, 2003) can also be extracted directly from the sound ensembles and compared to the ones extracted in the brain.

To extract such measures from our models, we read out directly the parameters of the learned STRFs and converted them in Cartesian coordinates (Schädler *et al.*, 2012) with the temporal modulation ω_k and spectral modulation Ω_k: $(\sigma_t, \sigma_f, \omega_k, \Omega_k)$, where $\omega_k = F_k \cos(\gamma_k)$ and $\Omega_k = F_k \sin(\gamma_k)$. We took the same convention as Chi *et al.* (2005) and Singh and Theunissen (2003) for the up-sweep and down-sweep modulations and represented only half the plan due to the symmetry.

We adapted the measures of separability, asymmetry, low-pass coefficient, and starriness coefficients with the interpretable parameters obtained for each supervised learning task. As the learned STRFs self-organized to solve each task, we examined each of these parameters for each task. Each α is estimated with the bootstrap re-sampling method (Efron and Tibshirani, 1994) on the learned STRFs (100 bootstraps).

1. Asymmetry

The distribution of the learned STRFs can show asymmetry preferences. The distribution is considered asymmetric if there are preferences for either down-sweep or up-sweep learned STRFs

$$\alpha_{\text{asymmetry}} = \frac{card(\{g_k \text{s.t.} \omega_k > 0\})}{card(\{g_k\})} = \frac{card(\{g_k \text{s.t.} \omega_k > 0\})}{N}. \quad (3)$$

If $\alpha_{\text{asymmetry}} \approx 0$, the distribution of STRFs filters $\{g_k(t,f)\}_{k=0..N-1}$ is considered symmetric. If $\alpha_{\text{asymmetry}} > 0$, there are more up-sweeps than down-sweeps. For instance, zebra finches exploit these degrees of freedom during their calls. The distribution of down-sweeps of zebra finch calls differs between male and female (Theunissen *et al.*, 2000).

2. Low-pass coefficient and starriness

It has been observed in Singh and Theunissen (2003) that most energy in the modulation power spectrum was concentrated in low spectral and temporal modulations for natural sounds. In addition, the higher spectral and temporal modulations were not distributed uniformly but were mostly along the axes. We derived two coefficients to quantify these phenomena with the learned STRFs:

$$\alpha_{\text{low}} = \frac{card(\{g_k \text{s.t.} |\omega_k| < \Delta_t, \Omega_k < \Delta_f\})}{N}, = \frac{N_{\text{low}}}{N}. \quad (4)$$

For the temporal modulation low limit, we opt, as did Singh and Theunissen (2003), for $\Delta_t = 16$ Hz. The spectral modulation low limit is set to $\Delta_f = 0.08$ cycle/octave. These parameters were chosen deliberately low as in Singh and Theunissen (2003) to observe differences between tasks. The parameter α_{star} measures the "starriness" of the distribution. This measure examines portions of the distribution that do not have high joint modulation and is an indicator of the importance of low modulation in either time or frequency, but not both,

$$\alpha_{\text{star}} = \frac{N_{\Delta_t} + N_{\Delta_f} - 2 \times N_{\text{low}}}{N - N_{\text{low}}}. \quad (5)$$

The quantities $N_{\Delta_t} = card(\{g_k \text{s.t.} |\omega_k| < \Delta_t\})$ and $N_{\Delta_f} = card(\{g_k \text{s.t.} \Omega_k < \Delta_f\})$ are the regions near the axes.

https://doi.org/10.1121/10.0005482

3. Separability

To obtain a separability measure from the learned STRFs, we approximated the 2D-distribution $\mathcal{P}(\omega,\Omega)$ of the filters with kernel density estimation with Gaussian filters. Then we evaluate if the normalized 2D-distribution $\mathcal{P}$ can be factorized into a product of two independent functions, $\mathcal{P}(\omega,\Omega) = G(\omega)\cdot F(\Omega)$. To quantify the separability, we calculated as Singh and Theunissen (2003) the singular value decomposition of the $\mathcal{P}(\omega,\Omega)$ obtained from each task

$$\mathcal{P}(\omega,\Omega) = \sum_{i=1}^{n} \lambda_i g_i(\omega)\cdot h_i(\Omega), \lambda_1 > \lambda_2 > \cdots > \lambda_n. \quad (6)$$

Then we computed the ratio of first singular value relative to the sum of all singular values

$$\alpha_{\text{sep}} = \frac{\lambda_1}{\sum_{i=1}^{n}\lambda_i}. \quad (7)$$

If $\alpha_{\text{sep}} \approx 1$, the distribution of the learned STRFs can be considered separable. Indeed, the magnitudes of the singular values λ_i of the decomposition provide information about the stretching/shrinking in the corresponding directions.[2] Complete separability suggests that there are fully independent temporal and spectral processing stages in the brain (Depireux *et al.*, 2001; Flinker *et al.*, 2019).

4. Measuring distance between tasks based on the learned STRF filters and optimal transport

The α measures provide some descriptors allowing some comparison between the learned distributions. However, they only look at one view and aspect of the learned distributions at a time. There is no clear way to measure the distances between each task based on the α. Besides, these α measures do not take into account the learned Gaussian envelope parameters $(\sigma_{t_k},\sigma_{f_k})$. Here, the goal is to obtain a quantitative metric able to compare the distributions obtained from each task. Usually, researchers fall back on the Mahalanobis distance or an approximation of the Kullback–Leibler (KL) divergence to compare observations of two sets of points, yet either these metrics make modeling assumptions about the data (approximation of the underlying density functions that generated the data), or it is impossible to compare sets of points with different cardinals.

The non-parametric, natural, and most powerful way to compare distributions is to use optimal transport distances (Peyré and Cuturi, 2019). Instead of computing distance between two individual items at a time, optimal transport is concerned with the problem of moving simultaneously several items (i.e., a distribution) from one configuration onto another. We compared the different tasks by comparing the learned STRFs using the regularized version Sinkhorn distance (Cuturi, 2013; Flamary and Courty, 2017). Especially, this regularized version of the optimal transport distance allows fast computation of distances and multiple assignments between points. Optimal transport distances require a metric space to find the transport between the sets of points. We made the choice to compare two individual learned STRFs with the Euclidean distance $||.||$. We normalized along each axis/parameter to not privilege for a specific parameter variability. Based on each task we tackled in this work, we obtained a distribution of normalized learned parameters $\{\sigma_{t_k},\sigma_{f_k},\omega_k,\Omega_k\}_{task}$ with the size n_{task} being the total number of filters used for this task. Therefore, equipped with the Euclidean distance to compare the individual filters, we can obtain the cost matrix between two tasks $M_{(task_a,task_b)} \in \mathbb{R}^{n_{task_a}\times n_{task_b}}$. We did not privilege any learned STRFs to build the distribution; therefore, we attributed equal weight to each individual filter $w_{task} = (1/n_{task})1_{n_{task}}$. This allows us to compare the different tasks if we have several models due to cross-validation (urban and bird) or fewer filters for a specific task (bird). If we denote, by $\langle .,. \rangle_F$, the norm of Froebenius between matrices, the regularized distance d_λ between two tasks is defined as

$$\begin{aligned} &d_\lambda = \min_P \langle P, M\rangle_F - \lambda\cdot h(P),\\ &\text{s.t. } P1_{n_{task_a}} = w_{task_a},\\ &P^T 1_{n_{task_b}} = w_{task_b},\\ &P \in \mathbb{R}_+^{n_{task_a}\times n_{task_b}},\\ &h(P) = -\sum_{i,j} P_{i,j}\log(P_{i,j}),\\ &\lambda = 10^{-3}. \end{aligned} \quad (8)$$

Therefore, we were able to obtain a proxy on how close two different tasks $task_a$ and $task_b$ are to each other based on the Sinkhorn distance $d_\lambda(\{\sigma_{t_k},\sigma_{f_k},\omega_k,\Omega_k\}_{task_a}, \{\sigma_{t_k},\sigma_{f_k},\omega_k,\Omega_k\}_{task_b})$. Based on the distances between all tasks, we built a hierarchical cluster tree and represent these distances with a dendrogram (see Fig. 5).

III. EXPERIMENTAL SETUP

We compared the learnable STRF layer with state-of-the-art systems that have recently been introduced to solve each task as well more classic baselines for each task. We tried to keep the experimental methods as close as possible to the methods used in the reported experiments of state-of-the-art systems.

A. Speech activity detection

The goal of speech activity detection is to segment a given stream audio into portions of speech or non-speech. We choose this task, as it allows us to examine what are the exact spectro-temporal modulations that make standout speech in an audio stream with silences and background noises (Mesgarani *et al.*, 2006).

We conduct experiments with two challenging datasets with different characteristics.

The AMI database (McCowan *et al.*, 2005) is a meeting dataset in English recorded with multiple microphones in three different rooms. There are 180 different speakers in the dataset. Here, we focus on the *AMI.SpeakerDiarization.MixHeadset* protocol, as we are working only on single channel feature analysis. We denoted by "speech AMI" the experiments and the distribution of learned STRFs on this dataset and this task. The CHiME5 database (Barker *et al.*, 2018) is a dataset recorded at home during parties. Here, we focus also on single channel feature analysis with the *CHiME5.SpeakerDiarization.U01* protocol. We denoted by "speech CHIME5" the experiments and the distribution of learned STRFs on this dataset and this task.

We compared different input front-ends to tackle this task. We evaluated the learnable STRFs (64 filters) with a contraction layer (CL) as well the free 2D convolution with a CL. The CL is a convolution layer taking the outputs at each time step of the learnable STRFs to reduce the number of dimensions of output tensor of the learnable STRF layer $\mathbf{Z}$ [see Eq. (2)]. We compared these techniques with classic signal processing baselines used in speech processing: Mel-filterbanks with 64 filters and Mel-frequency cepstral coefficients (MFCC) with 19 coefficients, with their deltas and their delta-deltas. We also compared them with the more recent parametrized neural network SincNet. SincNet is composed of parametrized sinc functions, which implement 80 bandpass filters (to replace directly more classic input spectral representations), and three temporal convolution/pooling layers. All the input front-ends are then fed to a stack of two layers of BiLSTM layers of dimension 128 and two forward layers of dimension 32 before a final decision layer. The learning rate is controlled by a cyclical scheduler, each cycle lasting for 21 epochs. Data augmentation is performed directly on the waveform using additive noise based on the MUSAN database (Snyder *et al.*, 2015) with a random target signal-to-noise ratio ranging from 5 to 20 dB. To evaluate speech activity detection, we used the detection error rate (DetER),

$$\text{DetER} = \frac{T_{\text{false alarm}} + T_{\text{missed detection}}}{T_{\text{total speech}}}.$$

We also reported the missed detection rate (%) and false alarm rate (%). We used the implementation of the metrics from `pyannote.metrics` (Bredin, 2017), and all experiments were run with `pyannote.audio` (Bredin *et al.*, 2020).

We ran an additional analysis for the speech activity detection task to compare the use of $\Re(\mathbf{Z})$, $\Im(\mathbf{Z})$, $|\mathbf{Z}|$ and $[\Re(\mathbf{Z}), \Im(\mathbf{Z})]$ (see Table V in the Appendix).

B. Speaker verification

The goal of the speaker verification task in speech processing is to accept or reject the hypothesis that a given speaker pronounced a given sentence. To do so, we learned an embedding function of any speech sequence of variable length. We examined this task, as it is believed that spectro-temporal modulations encode specifically the speaker information (Elliott and Theunissen, 2009; Lei *et al.*, 2012).

We followed the same procedure as Coria *et al.* (2020) to conduct experiments with the two versions of the VoxCeleb databases: VoxCeleb2 (Chung *et al.*, 2018) is used for training, and VoxCeleb1 (Nagrani *et al.*, 2017) is split into two parts for development and test sets. We compared two different input front-ends for the speaker verification task. We compared the learnable STRFs (64 filters) with a CL and the SincNet front-end, as described in the speech activity detection setup. Each model is trained with the additive angular margin loss ($\alpha = 10, m = 0.05$) with stochastic gradient descent with a learning rate of 0.01. We compared the different speaker verification approaches with the equal error rate (EER). We also measured the performance of each approach when using the S-normalization. We also reported the baseline performance of the I-vector system trained on VoxCeleb1 combined with probabilistic linear discriminant analysis (PLDA). We denoted by "speaker" the experiments and the distribution of learned STRFs on this dataset and this task.

C. Urban sound classification

The problem of urban sound classification is to classify short excerpts of audio sounds into broad categories (e.g., car horns, air conditioners, drilling). We investigated the use of the learnable STRFs for urban sound classification, especially to test the use of spectro-temporal modulations for types of sounds other than animal (human or bird) vocalizations (Młynarski and McDermott, 2018).

We followed the same evaluation procedure as Salamon and Bello (2017) to evaluate the experiments with the UrbanSound8K database (Salamon *et al.*, 2014). The dataset is composed of 8732 excerpts of urban sounds from ten categories (air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren, street music) and split into ten separate folds. To compare with previous approaches, each model is evaluated by cross-validation on the ten folds. We reported the mean, minimum, and maximum of the accuracy across the ten folds. We used the code-base from Arnault *et al.* (2020) for the training and evaluations of the two approaches. The state-of-the-art approach is the use of Mel-filterbanks with the CNN10 architecture from Kong *et al.* (2020). For the learnable STRFs approach, the first convolution layer of the CNN10 architecture (free 2D convolution with size 3×3 with 64 filters) is replaced by the learnable STRFs layer (64 filters) on top of the Mel-filterbanks described in Sec. II. The models are trained with the RAdam optimizer (Liu *et al.*, 2019) with LookAhead (Zhang *et al.*, 2019). We also reported the results from Salamon and Bello (2017) as baseline. We denoted by "urban" the experiments and the distribution of learned STRFs on this dataset and this task.

D. Zebra finch call type classification

Finally, we examined the zebra finch call type classification task (Elie and Theunissen, 2016). The goal of this

task is to classify short excerpts of sounds into call type categories for the zebra finch bird. Indeed, it has been found by Elie and Theunissen (2016) that several properties of the acoustic space allow the separation, to some extent, of the call types in the repertoire of zebra finches. We tried to stay as close as possible to the experimental protocol of Elie and Theunissen (2016). The dataset is composed of 3433 excerpts of zebra finches' calls from 11 categories ("Wsst or aggressive call," "begging calls," "distance call," "distress call," "long tonal call," "nest call," "song," "tet call," "thuk call," "tuck call," "whine call") produced by adults and chicks. The calls were segmented to keep only the first 3 s of each excerpt, and if the file was too short, the sound was zero-padded.

Each set of features and model was evaluated with a random cross-validation procedure that took into account the nested format of the database. Eighty percent of the birds were kept for training and 20% for testing. Fifty different permutations of excluded birds were obtained to generate 50 training and validation data sets. To compare the approaches, we computed the mean, minimum, and maximum of the accuracy over the permutations.

We ran four different baselines for this task based on two different input features and two types of classifiers. We extracted the features introduced by Elie and Theunissen (2016): predefined acoustical features (PAF) and the modulation power spectrum (MPS). The PAFs are composed of 23 parameters extracted from the spectral envelope, temporal envelope, and fundamental frequency [mean, minimum, maximum, standard deviation (SD) of the F0; mean of F1; mean of F2; mean of F3; saliency; root mean square (RMS) energy; maximum of the amplitude; mean, SD, skewness, kurtosis, entropy, first, second, and third quartiles of the frequency power spectrum; mean, SD, skewness, kurtosis, entropy of the temporal envelope]. The MPS representation is the amplitude spectrum of the 2D Fourier transform applied on the spectrum representation of the sound waveform. The MPS extracts the spectro-temporal modulations in a fine-grained fashion and sums the contribution along the frequency axis. We tested both these input features with linear discriminant analysis (LDA) and random forest (RF) classifiers as in Elie and Theunissen (2016).

Finally, we evaluated the potential of the learnable STRFs (24 filters) for this task. We combined the learnable STRFs with a simple linear layer to directly output the decision layer. The models were trained with the Adam optimizer (Kingma and Ba, 2014). We denote by "bird" the experiments and the distribution of learned STRFs on this dataset and this task.

IV. RESULTS AND DISCUSSION

First, we analyzed the quantitative performances to perform the tasks for the learnable STRFs for the different audio benchmarks. Then we examined and compared, qualitatively and quantitatively, the statistics of the learned STRF representations.

A. Quantitative performance on audio benchmarks

Overall, the performances of the learnable STRFs are on par for all tasks with the different baselines. There is no skip connection between the Mel-filterbanks and the rest of each neural network that has been considered. This means that these learned STRFs are in some way useful to perform each task, as this layer acts as a filter. A degradation of performance means that it might be not fully sufficient to use spectro-temporal modulations to perform this specific task.

The objective results for the speech activity detection task are shown for all models in Table I. Overall, learnable front-end approaches with injected prior improved over the classic signal processing baselines, Mel-filterbanks, and MFCCs, yet the approaches with free 2D convolution were not capable of improving over the classic signal processing baselines and had the worst performance, even for the convolution initialized with Gabor filters. The best-performing models for this task were the ones trained with the learnable STRFs, and they outperformed all the baselines. They improved over the state-of-the-art model with SincNet on the AMI dataset and matched the performance on the CHIME5 dataset. Therefore, adding prior for spectro-temporal modulations was beneficial for speech activity detection. The closest work to our knowledge around speech activity detection is Vuong *et al.* (2020), where they derived a layer that learned the spectro-temporal modulation especially for voice type discrimination in an industrial environment. The main difference from our work is that they relied on the expression of the discrete implementation of the Hilbert transform. They also reported that parametrized neural networks were better than free convolutions. They also used a long receptive field along the time axis and a small receptive field along the frequency axis.

The results for the speaker verification task are reported in Table II. We found that the SincNet that was designed initially for speaker recognition (Ravanelli and Bengio, 2018) got better results than the learnable STRFs + CL. The S-normalization improved both systems. This result differs from previous findings reported by Lei *et al.* (2012) that the spectro-temporal modulations were useful for speaker recognition. One difference, which could explain this discrepancy, is the use of Bayesian models after the different features [hidden Markov model-Gaussian mixture model (HMM-GMM)]. Indeed, the X-vector (Snyder *et al.*, 2015) was designed based on the latest advances of deep learning research to tackle the speaker recognition task and was validated initially on spectral representations of the audio. Our results suggest that spectro-temporal modulations are not fully sufficient to distinguish speakers. Harmonic structure was found useful for the speaker verification and recognition task (Imperl *et al.*, 1997). One of the hypotheses is that the learning of the global harmonic structure is more difficult with the output learnable STRF layer than directly with the Mel-filterbanks. The Gabor filters applied on log Mel-spectrograms are capable of capturing local harmonic

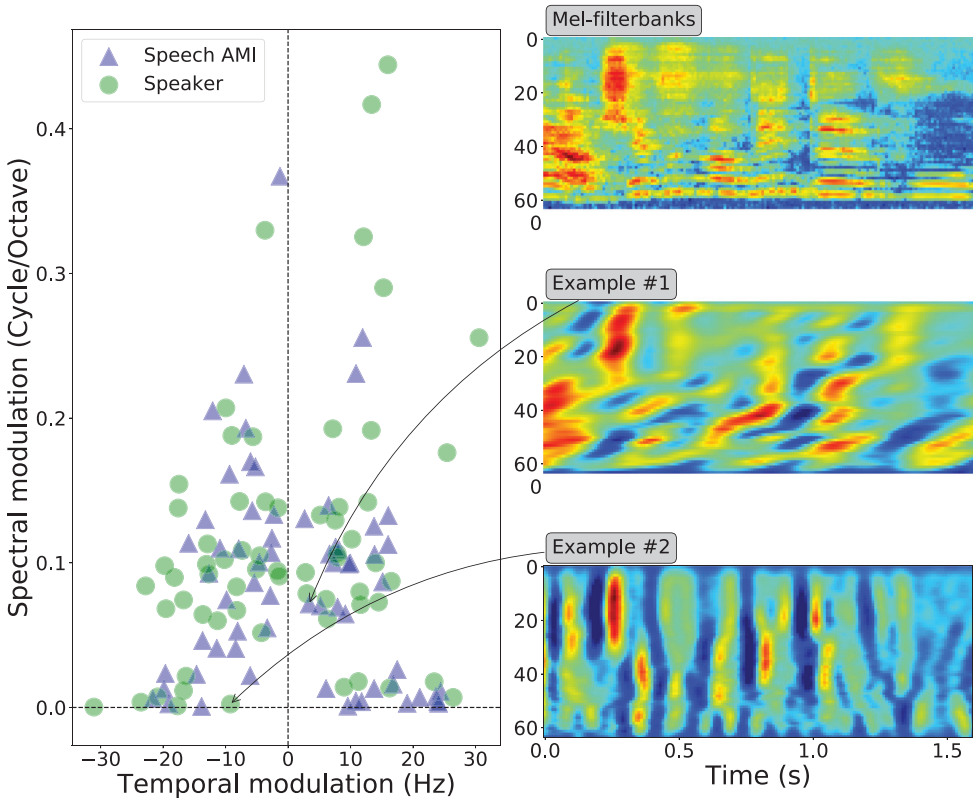


FIG. 2. (Color online) Left: Temporal and spectral modulation populations learned to tackle speech activity detection (speech AMI) on the AMI dataset and speaker verification (speaker). Top right: Mel-filterbanks representation of a sentence pronounced by a female speaker. Middle and bottom right: Output examples computed by the convolution of specific learned STRF kernels with the input Mel-filterbanks displayed in the first row.

dependencies, yet, as the receptive field along the frequency axis remained small (nine frequency bands), the Gabor filters cannot capture the global harmonic structure. The global harmonic structure is left to be learned by the rest of the models after the learnable STRFs. In contrast, specific global harmonic convolutions were introduced in Lostanlen (2017) and capture long dependencies across multiple octaves.

The performances for the urban sound classification task are reported in Table III. The accuracy of the learnable STRFs is above the baseline approach from Salamon and Bello (2017) and is on par (slightly below) with the CNN10 architecture using Mel-filterbanks (Kong *et al.*, 2020). It was found previously by Espi *et al.* (2015), that the use of different sizes of the spectral representation increased the performance of deep learning models for acoustic event detection. This suggests that the varying sizes of the focus on the Mel-filterbank representations boost the performances, both in time and frequency. In our case, the model learned to focus through the fitting of the (σ_f, σ_t) parameters.

TABLE I. Speech activity detection results for the different approaches described. The CL is a convolution layer reducing the size of the tensor dimension after the convolution (free 2D convolution or learnable STRFs) on the Mel-filterbanks. The free 2D convolution had the same grid size as the learnable STRFs (9×111). Each input front-end is then fed to a two-layer BiLSTM and two feed-forward layers. The best scores for each metric overall are in bold. MD, missed detection rate; FA, false alarm rate; DetER, detection error rate. For all metrics, lower is better.

	AMI database			CHIME5 database		
Input front-end	DetER	MD	FA	DetER	MD	FA
Mel-filterbanks	7.7	2.6	5.1	24.1	2.8	21.3
MFCC	6.3	2.7	3.5	19.6	1.6	18.0
SincNet (Ravanelli and Bengio, 2018)	6.0	**2.4**	3.6	**19.2**	1.7	17.6
Free 2D conv. random Init. + CL	8.0	3.0	5.0	26.5	0.6	25.9
Free 2D conv. Gabor Init. + CL	7.9	2.5	5.3	26.4	**0.2**	26.1
Learnable STRFs + CL	**5.8**	**2.4**	**3.4**	**19.2**	3.1	**16.1**

TABLE II. Speaker verification results for the different approaches described. The CL is a convolution layer reducing the size of the tensor dimension after the convolution (learnable STRFs). The X-vector [Snyder *et al.* (2018)] is used after each input front-end. We evaluated the performance of the speaker verification with and without S-normalization (Coria *et al.*, 2020). We also reported the baseline performance of the I-vector combined with PLDA (?). The best scores for each metric overall are in bold. For the EER, lower is better.

Metric	EER	EER with S-normalization
Baseline		
I-vectors + PLDA (?)[a]	8.8	—
Input front-end		
SincNet (Coria *et al.*, 2020)	**3.9**	**3.5**
Learnable STRFs + CL	6.4	6.1

[a]This result is directly extracted from Coria *et al.* (2020) and was not replicated for this study.

https://doi.org/10.1121/10.0005482

TABLE III. Urban sound classification results for the different approaches described. The best score for the mean accuracy over the 10 folds overall is in bold. The CNN10 architecture from Kong *et al.* (2020) is used after each input front-end. Higher is better.

	Accuracy (%)	
	Mean	(Minimum–maximum)
Baseline		
SB-CNN (Salamon and Bello, 2017)[a]	79	(71–85)
Input front-end		
Free 2D conv. 3 × 3 (Kong *et al.*, 2020)	**84**	(76–93)
Learnable STRFs	82	(74–90)

[a]This result is directly extracted from Salamon and Bello (2017) and was not replicated for this study.

TABLE IV. Zebra finch call type classification results for the different approaches. The best scores for each metric overall are in **bold**. PAF, predefined acoustical features; MPS, modulation power spectrum; LDA, linear discriminant analysis. RF, random forest. The learnable STRF input front-end is combined with a simple linear model to output directly the decisions. Higher is better.

	Accuracy (%)	
	Mean	(Minimum–maximum)
Chance level	17	6–23
Features + model		
PAF (Elie and Theunissen, 2016) + LDA	57	(43–71)
PAF (Elie and Theunissen, 2016) + RF	59	(47–68)
MPS (Elie and Theunissen, 2016) + LDA	41	(23–53)
MPS (Elie and Theunissen, 2016) + RF	**69**	(49–84)
Input front-end		
Learnable STRFs	43	(23–73)

Finally, the results for the zebra finch call type classification are shown in Table IV. On one hand, the PAF features depended slightly on the model used after for classification (LDA 57% to RF 58%), while the MPS had the worst performance overall with a linear model, such as LDA, and the MPS had the best performance overall when combined with the RF (going from 41% to 69%). The learnable STRF models were decoded with a simple linear layer, so the closest baseline is the combination of the MPS with the LDA. The learnable STRFs perform below the PAF features and the MPS with RF. The performance with a combination of features in the MPS with RF suggests that the model with learnable STRFs could benefit greatly from adaptive neural trees (Tanno *et al.*, 2019) to perform the task. In addition, this encourages the use of co-occurrences or anti-occurrences of the spectro-temporal patterns in models as in Młynarski and McDermott (2018, 2019), since the routing in RF implies measurement of joint patterns in the feature space of the MPS.

B. Description of the learned filters

First, we observed that the learned STRFs organized differently for each task, both the modulations (ω, Ω) (see Fig. 3) and the size of the Gaussian envelopes through (σ_t, σ_f) (see Fig. 6 in the Appendix). Within the space allowed by the Nyquist theorem and the size of the convolutions, all the learned STRFs concentrated in low spectral and temporal modulations (see Fig. 3). We also observed, as did Singh and Theunissen (2003), that higher spectral modulations were found at low temporal modulations (and vice versa). We found that the Gaussian envelopes of the learned

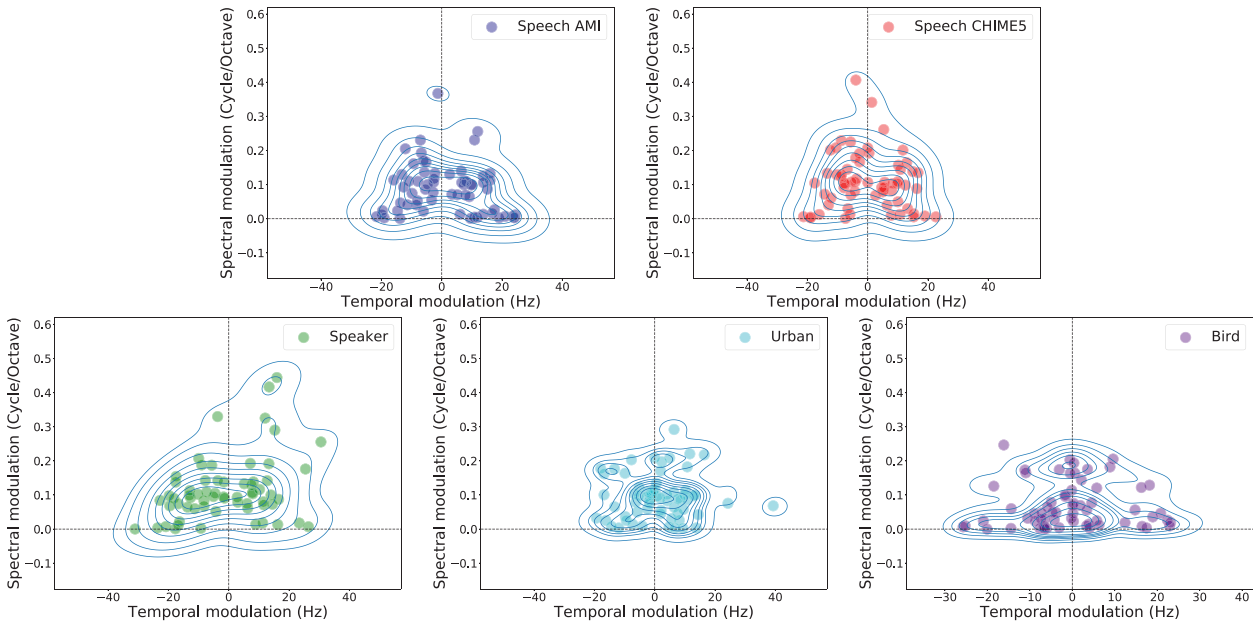


FIG. 3. (Color online) Temporal and spectral modulation of the learned STRFs to tackle speech activity detection on the AMI dataset (speech AMI) and on the CHIME5 (speech CHIME5), speaker verification on VoxCeleb (speaker), urban sound classification on Urban8k (urban), and zebra finch call type classification (bird). We displayed only a subset of the learned STFRs of the bird and urban tasks for clarity. We also plotted the bi-variate distributions using kernel density estimation for each task.

https://doi.org/10.1121/10.0005482

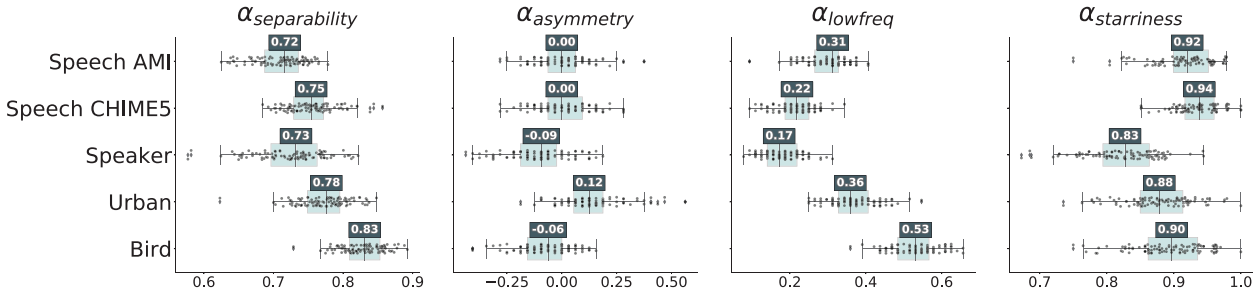


FIG. 4. (Color online) Separability, asymmetry, low-pass, starriness coefficients. Four quantifiers measured different aspects of learned distribution for the different tasks under study: speech activity detection on the CHIME5 dataset (Speech CHIME5) and on the AMI (Speech AMI), urban sound classification on Urban8k (Urban), speaker verification on VoxCeleb (Speaker), and zebra finch call type classification (Bird). We display the median value for each α and task above each box-plot.

STRFs can be characterized more by a continuum of values, not in a set of specific values. The Gaussian envelopes are more concentrated in the low values and exhibit preferences depending on the task for temporal or spectral shapes. Finally, the distributions of the learned STRF modulations and Gaussian envelopes of speech tasks on AMI and CHIME5 datasets and the speaker task look more similar than the bird and urban ones. We quantified the learned parameters with the distributional parameters that measured modulation (Sec. II C) and the optimal transport distance between distributions (Sec. II C 4).

First, the separability index $\alpha_{separability}$ showed that most learned STRFs are quite separable and that the tasks related to human vocalizations (speech and speaker) were less separable than the other ones (Fig. 4). The results for the separability for speech tasks are consistent with speech perception in humans, as found very recently by Flinker *et al.* (2019). There seems to be independence for the processing of spectral and temporal modulations. We also found that all modulations have quite high $\alpha_{starriness}$ indexes. Similar results were found in Singh and Theunissen (2003) for the separability and the starriness for the ensembles of sounds of speech corpora, zebra finch vocalizations, and environmental sounds. Schädler *et al.* (2012) evaluated the use of high joint spectral and temporal modulation and also found that they were degrading the performance for speech recognition tasks.

In addition, the learned STRFs for the speech did not show preferences for up- or down-sweep modulations ($\alpha_{asymmetry} \approx 0.0$), while the speaker and bird tasks exhibit slight preferences for down-sweeps and the urban for up-sweeps. The result for the bird task differed from Singh and Theunissen (2003). This could be explained by the fact that Singh and Theunissen (2003) used a quantification of these parameters with an ensemble of sounds. The information about the specific characteristic of an individual zebra finch is mixed with the information of the call type. This suggests that a fully interpretable supervised approach might allow deciphering of the different factors and contributions that influenced the acoustic properties of vocalizations. Finally, the bird task focused more on the low frequency modulations ($\alpha_{lowfreq} = 0.53$) than the other tasks ($\alpha_{lowfreq} \leq 0.35$). We also observed that the learned STRFs of the speaker task moved away from the low spectral modulations and yielded the lowest low-pass coefficient ($\alpha_{lowfreq} = 0.19$). Especially, Elliott and Theunissen (2009) also found that the removal of spectral modulations between 3 and 7 cycles/kHz significantly increases the gender mis-identifications of female speakers. In addition, the results for the speech on the AMI and CHIME5 datasets and the speaker are very similar to the ones found directly in the auditory cortex neurons in awake monkeys (Massoudi *et al.*, 2015) and in awake humans (Hullett *et al.*, 2016; Schönwiesner and Zatorre, 2009). Hullett *et al.* (2016), Massoudi *et al.* (2015), and Schönwiesner and Zatorre (2009) measured responses of natural sounds directly in the superior temporal gyrus and found specific spectral modulation selectivity for 0.4 ± 0.55 cycle/octave and specific temporal modulation 16 ± 11 Hz, and most of the modulations were concentrated along the axes with high separability.

Finally, we examined the structure obtained from the hierarchical clustering based on the distances between tasks (see Sec. II C 4 for the full description). We obtained the clustering tree in Fig. 5. We observed that the learned STRFs of the different tasks organized in a meaningful disposition. The learned STRFs for speech on the CHIME5 and AMI are the closest to each other. Then we found that another human vocalization task, speaker, is closer to the speech ones. On the other hand, the bird task organized far away from both the urban and the human vocalization tasks (speech and speaker). In future work, this method could be used to discover automatically organization trees based only on acoustic properties of spectro-temporal modulations and test these predictions against what is known about the phylogeny, acoustic environment, and con-species living nearby (McCracken and Sheldon, 1997).

V. CONCLUSIONS AND FUTURE WORK

In summary, we examined the use of a parametrized neural network front-end to learn spectro-temporal modulations optimal for different behavioral tasks. This front-end, the learnable STRFs, yielded performances close to published state-of-the-art using an engineering-oriented neural

https://doi.org/10.1121/10.0005482

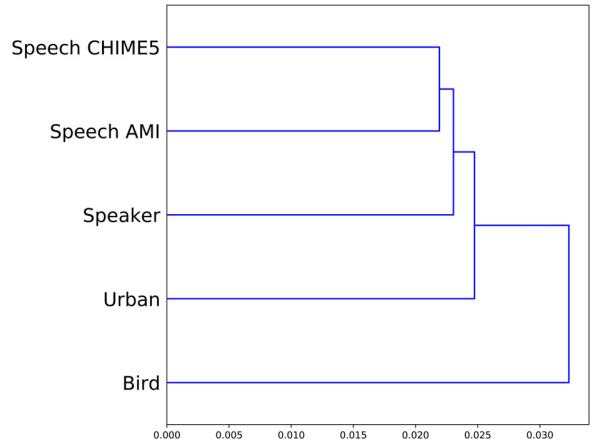


FIG. 5. (Color online) Hierarchical clustering of the tasks: speech activity detection on the CHIME5 dataset (Speech CHIME5) and on the AMI (Speech AMI), urban sound classification on Urban8k (Urban), speaker verification on VoxCeleb (Speaker), and zebra finch call type classification (Bird). The distance between tasks is computed between the learned STRF filters of each task with the Sinkhorn distance (we used the Euclidean distance between each filter, and the regularization parameter of the Sinkhorn distance is $\lambda = 10^{-3}$).

network for speaker verification, urban sound classification, and zebra finch call type classification and obtained the best results on two datasets for speech activity detection. As our front-end is fully interpretable, we found markedly different spectro-temporal modulations as a function of the task, showing that each task relies on a specific set of modulations. These task-specific modulations were globally congruent with previous work based on three approaches: spectro-temporal analysis of different audio signals (Elliott and Theunissen, 2009), analysis of trained neural networks (Schädler *et al.*, 2012), and analysis of the auditory cortex (Hullett *et al.*, 2016; Santoro *et al.*, 2017). In particular, for the speech activity detection task, we observed the same modulation distributions as the ones found directly by the human auditory cortex listening while listening to naturalistic speech [Hullett *et al.* (2016)]. The modulations also displayed generic characteristics across tasks, namely, a predominance of low frequency spectral and temporal modulations and a high degree of "starriness" and "separability," corresponding to the fact that filters tend to remain close to either the temporal or spectral axis, with low occupation of joint spectral and temporal responses. This is consistent with Singh and Theunissen (2003).

Several avenues of extensions are possible for this work, based on what is known in auditory neuroscience. First, this work only modelled the final outcome of plasticity after each task had been fully learned, starting from a random initialization. Yet, the same model could be used to address a range of issues relevant to changes occurring during task learning (top-down plasticity) or due to modification of the distribution of audio input (bottom-up plasticity). Recent work by Bellur and Elhilali (2015) has investigated the online adaptation of modulations in analytical models and witnessed several improvements in terms of engineering performance, suggesting that this is also an interesting avenue in terms of behavioral modeling. Second, the analyses from Hullett *et al.* (2016) showed that neurons not only have spectro-temporal selectivity, but they are also topographically distributed along the posterior-to-anterior axis in the superior temporal gyrus. In future work, it would be interesting to reproduce such topography by using an auxiliary self-organizing map objective in addition to the task-specific loss

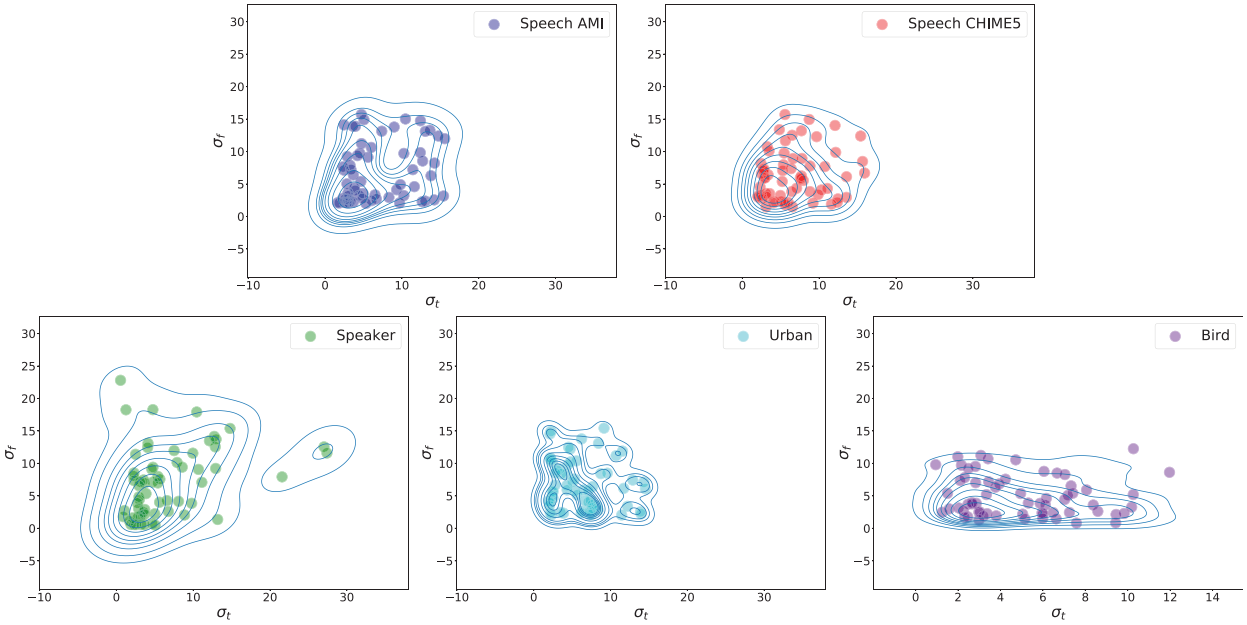


FIG. 6. (Color online) Gaussian envelopes (σ_t, σ_f) of the learned STRFs to tackle speech activity detection on the AMI dataset (Speech AMI) and on the CHIME5 (Speech CHIME5), speaker verification on VoxCeleb (Speaker), urban sound classification on Urban8k (Urban), and zebra finch call type classification (Bird). We displayed only a subset of the learned STFRs of the bird and urban tasks for clarity. We also plotted the bi-variate distributions using kernel density estimation for each task.

function for the STRFs. Finally, despite their wide use in auditory neuroscience, the spectro-temporal modulations do not provide a complete picture of computations in the auditory cortex (Williamson *et al.*, 2016). A potential extension of our work would be to add an extra layer able to express co-occurrences and anti-occurrences of pairs of spectro-temporal receptive fields as in Młynarski and McDermott (2018, 2019). Such an extra layer would provide a learnable and interpretable extension to spectro-temporal representations.

To conclude, we emphasize that neuroscience-inspired parametrized neural networks can provide models that are both efficient in terms of behavioral tasks and interpretable in terms of auditory signal processing.

ACKNOWLEDGMENTS

This work is funded in part from the Agence Nationale pour la Recherche (Grant Nos. ANR-17-EURE-0017 Frontcog, ANR-10-IDEX-0001-02 PSL*, and ANR-19-P3IA-0001 PRAIRIE 3IA Institute). A.C.B.L. was funded through Neuratris, and E.D. in his Ecole des Hautes Etudes en Sciences Sociales (EHESS) role by Facebook AI Research (Research Gift) and CIFAR (Learning in Minds and Brains). The university [EHESS, CNRS, INRIA, Ecole Normale Supérieure (ENS)-Paris Sciences et Lettres] obtained the datasets reported in this paper, and the experiments were run on its computer resources.

APPENDIX: IMPORTANCE OF THE REPRESENTATION OF THE LEARNABLE STRFS

We performed an additional analysis of speech activity detection of the choice of representations **Z** from the learnable STRFs used in the subsequent neural network (Fig. 6). The performance of the real part and the imaginary part and absolute values of the filter output are compared. The results are presented in Table V. In comparison with the concatenation of the real and imaginary parts, the performances obtained for each part were in the same range on the AMI dataset but were lower on the CHIME5 dataset. As in Schädler *et al.* (2012), this indicates that phase information contained in the real and imaginary parts is important for the learnable STRFs.

TABLE V. Speech activity detection results for the different uses of the **Z** for the learnable STRFs. The CL is a convolution layer reducing the size of the tensor dimension after the convolution (learnable STRFs). Each input front-end is then fed to a two-layer BiLSTM and two feed-forward layers. The best scores for each metric overall are in bold. MD, missed detection rate. FA, false alarm rate. DetER, detection error rate. For all metrics, lower is better.

Input front-end (learnable STRFs + CL)	AMI database			CHIME5 database		
	DetER	MD	FA	DetER	MD	FA
Real part $\Re(\mathbf{Z})$	5.9	2.4	3.5	20.1	2.6	17.5
Imaginary part $\Im(\mathbf{Z})$	5.9	**2.2**	3.7	22.1	**1.0**	21.1
Magnitude $\lvert\mathbf{Z}\rvert$	5.9	**2.2**	3.7	19.8	3.1	16.7
Concatenation $[\Re(\mathbf{Z}), \Im(\mathbf{Z})]$	**5.8**	2.4	**3.4**	**19.2**	3.1	**16.1**

[1]https://github.com/bootphon/learnable-strf (Last viewed 7/7/2021).
[2]For more information on separability, see https://bartwronski.com/2020/02/03/separate-your-filters-svd-and-low-rank-approximation-of-image-filters/ (Last viewed 7/7/2021).

Alekseev, A., and Bobe, A. (**2019**). "Gabornet: Gabor filters with learnable parameters in deep convolutional neural network," in *Proceedings of the 2019 International Conference on Engineering and Telecommunication (EnT)*, November 20–21, Dolgoprudny, Russia, pp. 1–4.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., Chen, J., Chen, J., Chen, Z., Chrzanowski, M., Coates, A., Diamos, G., Ding, K., Du, N., Elsen, E., Engel, J., Fang, W., Fan, L., Fougner, C., Gao, L., Gong, C., Hannun, A., Han, T., Johannes, L. V., Jiang, B., Ju, C., Jun, B., LeGresley, P., Lin, L., Liu, J., Liu, Y., Li, W., Li, X., Ma, D., Narang, S., Ng, A., Ozair, S., Peng, Y., Prenger, R., Qian, S., Quan, Z., Raiman, J., Rao, V., Satheesh, S., Seetapun, D., Sengupta, S., Srinet, K., Sriram, A., Tang, H., Tang, L., Wang, C., Wang, J., Wang, K., Wang, Y., Wang, Z., Wang, Z., Wu, S., Wei, L., Xiao, B., Xie, W., Xie, Y., Yogatama, D., Yuan, B., Zhan, J., and Zhu, Z. (**2016**). "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *Proceedings of the 33rd International Conference on Machine Learning*, New York, June 20–22, pp. 173–182.

Arnault, A., Hanssens, B., and Riche, N. (**2020**). "Urban sound classification: Striving towards a fair comparison," arXiv:2010.11805.

Barker, J., Watanabe, S., Vincent, E., and Trmal, J. (**2018**). "The fifth 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018)*, September 2–6, Hyderabad, India.

Bellur, A., and Elhilali, M. (**2015**). "Detection of speech tokens in noise using adaptive spectrotemporal receptive fields," in *Proceedings of the 2015 49th Annual Conference on Information Sciences and Systems (CISS)*, March 18–20, Baltimore, MD, pp. 1–6.

Bredin, H. (**2017**). "pyannote.metrics: A toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems," in *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH 2017)*, August 20–24, Stockholm, Sweden.

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., and Gill, M.-P. (**2020**). "Pyannote.Audio: Neural building blocks for speaker diarization," in *Proceedings of ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 4–8, Barcelona, Spain, pp. 7124–7128.

Chang, S.-Y., and Morgan, N. (**2014**). "Robust CNN-based speech recognition with Gabor filter kernels," in *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016)*, September 8–12, San Francisco, CA.

Cheuk, K. W., Anderson, H., Agres, K., and Herremans, D. (**2020**). "nnaudio: An on-the-fly GPU audio to spectrogram conversion toolbox using 1D convolutional neural networks," IEEE Access **8**, 161981–162003.

Chi, T., Ru, P., and Shamma, S. A. (**2005**). "Multiresolution spectrotemporal analysis of complex sounds," J. Acoust. Soc. Am. **118**(2), 887–906.

Chung, J. S., Nagrani, A., and Zisserman, A. (**2018**). "Voxceleb2: Deep speaker recognition," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018)*, September 2–6, Hyderabad, India, pp. 1086–1090.

Coria, J. M., Bredin, H., Ghannay, S., and Rosset, S. (**2020**). "A comparison of metric learning loss functions for end-to-end speaker verification," in *Statistical Language and Speech Processing*, edited by L. Espinosa-Anke, C. Martín-Vide, and I. Spasić (Springer International Publishing, Cham, Switzerland), pp. 137–148.

Cuturi, M. (**2013**). "Sinkhorn distances: Lightspeed computation of optimal transport," in *Proceedings of Advances in Neural Information Processing*

Systems 26 (NIPS 2013), December 5–10, Lake Tahoe, NV, pp. 2292–2300.
Depireux, D. A., Simon, J. Z., Klein, D. J., and Shamma, S. A. (**2001**). "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," J. Neurophysiol. **85**(3), 1220–1234.
Edraki, A., Chan, W.-Y., Jensen, J., and Fogerty, D. (**2019**). "Improvement and assessment of spectro-temporal modulation analysis for speech intelligibility estimation," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH 2019)*, September 15–19, Graz, Austria, pp. 1378–1382.
Efron, B., and Tibshirani, R. J. (**1994**). *An Introduction to the Bootstrap* (CRC, Boca Raton, FL).
Elhilali, M., Chi, T., and Shamma, S. A. (**2003**). "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," Speech Commun. **41**(2), 331–348.
Elie, J. E., and Theunissen, F. E. (**2016**). "The vocal repertoire of the domesticated zebra finch: A data-driven approach to decipher the information-bearing acoustic features of communication signals," Anim. Cogn. **19**(2), 285–315.
Elliott, T. M., and Theunissen, F. E. (**2009**). "The modulation transfer function for speech intelligibility," PLoS Comput. Biol. **5**(3), e1000302.
Espi, M., Fujimoto, M., Kinoshita, K., and Nakatani, T. (**2015**). "Exploiting spectro-temporal locality in deep learning based acoustic event detection," EURASIP J. Audio Speech Music Process. **2015**(1), 1–12.
Ezzat, T., Bouvrie, J., and Poggio, T. (**2007**). "Spectro-temporal analysis of speech using 2-D Gabor filters," in *Proceedings of the Eighth Annual Conference of the International Speech Communication Association (INTERSPEECH 2007)*, August 27–31, Antwerp, Belgium.
Flamary, R., and Courty, N. (**2017**). "POT: Python optimal transport," https://pythonot.github.io/ (Last viewed 7/7/2021).
Flinker, A., Doyle, W., Mehta, A., Devinsky, O., and Poeppel, D. (**2019**). "Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex," Nat. Hum. Behav. **3**(4), 393–405.
Francis, N. A., Elgueda, D., Englitz, B., Fritz, J. B., and Shamma, S. A. (**2018**). "Laminar profile of task-related plasticity in ferret primary auditory cortex," Sci. Rep. **8**(1), 16375.
Fritz, J., Shamma, S., Elhilali, M., and Klein, D. (**2003**). "Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex," Nat. Neurosci. **6**(11), 1216–1223.
Gabor, D. (**1946**). "Theory of communication. part 1: The analysis of information," J. Inst. Electr. Eng. Part III Radio Commun. Eng. **93**(26), 429–441.
Hullett, P. W., Hamilton, L. S., Mesgarani, N., Schreiner, C. E., and Chang, E. F. (**2016**). "Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli," J. Neurosci. **36**(6), 2014–2026.
Imperl, B., Kačič, Z., and Horvat, B. (**1997**). "A study of harmonic features for the speaker recognition," Speech Commun. **22**(4), 385–402.
Jääskeläinen, I. P., Ahveninen, J., Belliveau, J. W., Raij, T., and Sams, M. (**2007**). "Short-term plasticity in auditory cognition," Trends Neurosci. **30**(12), 653–661.
Kell, A. J., and McDermott, J. H. (**2019**). "Deep neural network models of sensory systems: Windows onto the role of task constraints," Curr. Opin. Neurobiol. **55**, 121–132.
Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., and McDermott, J. H. (**2018**). "A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy," Neuron **98**(3), 630–644.
Kingma, D. P., and Ba, J. (**2014**). "Adam: A method for stochastic optimization," arXiv:1412.6980.
Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., and Plumbley, M. D. (**2020**). "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 2880–2894.
Koumura, T., Terashima, H., and Furukawa, S. (**2019**). "Cascaded tuning to amplitude modulation for natural sound recognition," J. Neurosci. **39**(28), 5517–5533.
Lei, H., Meyer, B. T., and Mirghafori, N. (**2012**). "Spectro-temporal Gabor features for speaker recognition," in *Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 25–30, Kyoto, Japan, pp. 4241–4244.
Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., and Han, J. (**2019**). "On the variance of the adaptive learning rate and beyond," in *Proceedings of the International Conference on Learning Representations*, May 6–9, New Orleans, LA.
Lostanlen, V. (**2017**). "Convolutional operators in the time-frequency domain," Ph.D. thesis, Université Paris Sciences et Lettres, Paris, France.
Massoudi, R., Van Wanrooij, M. M., Versnel, H., and Van Opstal, A. J. (**2015**). "Spectrotemporal response properties of core auditory cortex neurons in awake monkey," PLoS One **10**(2), e0116118.
McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., Post, W., Reidsma, D., and Wellner, P. (**2005**). "The AMI meeting corpus," in *Proceedings of Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research*, August 30–September 2, Wageningen, Netherlands, pp. 137–140.
McCracken, K. G., and Sheldon, F. H. (**1997**). "Avian vocalizations and phylogenetic signal," Proc. Nat. Acad. Sci. U.S.A. **94**(8), 3833–3836.
McDermott, J. H. (**2018**). "Audition," in *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, Vol. 2 (Wiley, New York), pp. 1–57.
Mesgarani, N., Slaney, M., and Shamma, S. A. (**2006**). "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," IEEE Trans. Audio Speech Lang. Process. **14**(3), 920–930.
Meyer, A. F., Williamson, R. S., Linden, J. F., and Sahani, M. (**2017**). "Models of neuronal stimulus-response functions: Elaboration, estimation, and evaluation," Front. Syst. Neurosci. **10**, 109.
Młynarski, W., and McDermott, J. H. (**2018**). "Learning midlevel auditory codes from natural sound statistics," Neural Comput. **30**(3), 631–669.
Młynarski, W., and McDermott, J. H. (**2019**). "Ecological origins of perceptual grouping principles in the auditory system," Proc. Natl. Acad. Sci. U.S.A. **116**(50), 25355–25364.
Nagrani, A., Chung, J. S., and Zisserman, A. (**2017**). "Voxceleb: A large-scale speaker identification dataset," in *Proc. Interspeech*, pp. 2616–2620.
Ondel, L., Li, R., Sell, G., and Hermansky, H. (**2019**). "Deriving spectro-temporal properties of hearing from speech data," in *Proceedings of ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 12–17, Brighton, UK, pp. 411–415.
Peyré, G., and Cuturi, M. (**2019**). "Computational optimal transport: With applications to data science," Found. Trends Mach. Learn. **11**(5), 355–607.
Pillow, J., and Sahani, M. (**2019**). "Editorial Overview: Machine Learning, Big Data, and Neuroscience," Curr. Opin. Neurobiol. **55**, iii–iv.
Poeppel, D. (**2003**). "The analysis of speech in different temporal integration windows: Cerebral lateralization as 'asymmetric sampling in time,'" Speech Commun. **41**(1), 245–255.
Ravanelli, M., and Bengio, Y. (**2018**). "Speaker recognition from raw waveform with sincnet," in *Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT)*, December 18–21, Athens, Greece, pp. 1021–1028.
Saddler, M. R., Gonzalez, R., and McDermott, J. H. (**2020**). "Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception," bioRxiv 2020.11.19.389999.
Salamon, J., and Bello, J. P. (**2017**). "Deep convolutional neural networks and data augmentation for environmental sound classification," IEEE Signal Process. Lett. **24**(3), 279–283.
Salamon, J., Jacoby, C., and Bello, J. P. (**2014**). "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM International Conference on Multimedia*, November 3–7, Orlando, FL, pp. 1041–1044.
Santoro, R., Moerel, M., De Martino, F., Valente, G., Ugurbil, K., Yacoub, E., and Formisano, E. (**2017**). "Reconstructing the spectrotemporal modulations of real-life sounds from fmri response patterns," Proc. Natl. Acad. Sci. U.S.A. **114**(18), 4799–4804.
Schädler, M. R., Meyer, B. T., and Kollmeier, B. (**2012**). "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," J. Acoust. Soc. Am. **131**(5), 4134–4151.
Schönwiesner, M., and Zatorre, R. (**2009**). "Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI," Proc. Natl. Acad. Sci. U.S.A. **106**(34), 14611–14616.

Shamma, S. A. (**1996**). "Auditory cortical representation of complex acoustic spectra as inferred from the ripple analysis method," Network Comput. Neural Syst. **7**(3), 439–476.

Singh, N. C., and Theunissen, F. E. (**2003**). "Modulation spectra of natural sounds and ethological theories of auditory processing," J. Acoust. Soc. Am. **114**(6), 3394–3411.

Snyder, D., Chen, G., and Povey, D. (**2015**). "Musan: A music, speech, and noise corpus," arXiv:1510.08484.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (**2018**). "X-vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 15–20, Calgary, Canada, pp. 5329–5333.

Stevens, S. S., Volkmann, J., and Newman, E. B. (**1937**). "A scale for the measurement of the psychological magnitude pitch," J. Acoust. Soc. Am. **8**(3), 185–190.

Tanno, R., Arulkumaran, K., Alexander, D., Criminisi, A., and Nori, A. (**2019**). "Adaptive neural trees," in *Proceedings of the 36th International Conference on Machine Learning*, June 9–15, Long Beach, CA, pp. 6166–6175.

Theunissen, F., Sen, K., and Doupe, A. (**2000**). "Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds," J. Neurosci. **20**(6), 2315–2331.

Thoret, E., Andrillon, T., Léger, D., and Pressnitzer, D. (**2020**). "Probing machine-learning classifiers using noise, bubbles, and reverse correlation," bioRxiv 2020.06.22.165688.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. (**2016**). "Instance normalization: The missing ingredient for fast stylization," arXiv:1607.08022.

Vuong, T., Xia, Y., and Stern, R. M. (**2020**). "Learnable spectro-temporal receptive fields for robust voice type discrimination," in *Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH 2020)*, October 25–29, Shanghai, China, pp. 1957–1961.

Williamson, R. S., Ahrens, M. B., Linden, J. F., and Sahani, M. (**2016**). "Input-specific gain modulation by local sensory context shapes cortical and thalamic responses to complex sounds," Neuron **91**(2), 467–481.

Woolley, S. M., Fremouw, T. E., Hsu, A., and Theunissen, F. E. (**2005**). "Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds," Nat. Neurosci. **8**(10), 1371–1379.

Yarkoni, T., and Westfall, J. (**2017**). "Choosing prediction over explanation in psychology: Lessons from machine learning," Perspect. Psychol. Sci. **12**(6), 1100–1122.

Zeghidour, N., Usunier, N., Synnaeve, G., Collobert, R., and Dupoux, E. (**2018**). "End-to-end speech recognition from the raw waveform," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018)*, September 2–6, Hyderabad, India, pp. 781–785.

Zhang, M., Lucas, J., Ba, J., and Hinton, G. E. (**2019**). "Lookahead optimizer: k steps forward, 1 step back," in *Proceedings of Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, December 8–14, Vancouver, Canada, Vol. 32, pp. 9597–9608.

A.4 Learning Strides in convolutional neural networks

LEARNING STRIDES IN CONVOLUTIONAL NEURAL NETWORKS

Rachid Riad *
ENS, INRIA, INSERM, UPEC, PSL Research University
rachid.riad@ens.fr

Olivier Teboul & David Grangier & Neil Zeghidour
Google Research
{neilz, oliviert, grangier}@google.com

ABSTRACT

Convolutional neural networks typically contain several downsampling operators, such as strided convolutions or pooling layers, that progressively reduce the resolution of intermediate representations. This provides some shift-invariance while reducing the computational complexity of the whole architecture. A critical hyperparameter of such layers is their stride: the integer factor of downsampling. As strides are not differentiable, finding the best configuration either requires cross-validation or discrete optimization (e.g. architecture search), which rapidly become prohibitive as the search space grows exponentially with the number of downsampling layers. Hence, exploring this search space by gradient descent would allow finding better configurations at a lower computational cost. This work introduces DiffStride, the first downsampling layer with learnable strides. Our layer learns the size of a cropping mask in the Fourier domain, that effectively performs resizing in a differentiable way. Experiments on audio and image classification show the generality and effectiveness of our solution: we use DiffStride as a drop-in replacement to standard downsampling layers and outperform them. In particular, we show that introducing our layer into a ResNet-18 architecture allows keeping consistent high performance on CIFAR10, CIFAR100 and ImageNet even when training starts from poor random stride configurations. Moreover, formulating strides as learnable variables allows us to introduce a regularization term that controls the computational complexity of the architecture. We show how this regularization allows trading off accuracy for efficiency on ImageNet.

1 INTRODUCTION

Convolutional neural networks (CNNs) (Fukushima, 1980; LeCun et al., 1989) have been the most widely used neural architecture across a wide range of tasks, including image classification (Krizhevsky et al., 2012; He et al., 2016a; Huang et al., 2017; Bello et al., 2021), audio pattern recognition (Kong et al., 2020), text classification (Conneau et al., 2017), machine translation (Gehring et al., 2017) and speech recognition (Amodei et al., 2016; Sercu et al., 2016; Zeghidour et al., 2018). Convolution layers, which are the building block of CNNs, project input features to a higher-level representation while preserving their resolution. When composed with non-linearities and normalization layers, this allows for learning rich mappings at a constant resolution, e.g. autogressive image synthesis (van den Oord et al., 2016). However, many tasks infer high-level low-resolution information (identity of a speaker (Muckenhirn et al., 2018), presence of a face (Chopra et al., 2005)) by integrating over low-level, high-resolution measurements (waveform, pixels). This integration requires extracting the right features, discarding irrelevant information over several downsampling steps. To that end, pooling layers and strided convolutions aggressively reduce the resolution of their inputs, providing several benefits. First, they act as a bottleneck that forces features to focus on information relevant to the task at hand. Second, pooling layers such as low-pass filters (Zhang, 2019)

*This work was conducted while interning at Google.

improve shift-invariance. Third, a reduced resolution implies a reduced number of floating-point operations and a higher receptive field in the subsequent layers.

Pooling layers can usually be decomposed into two basic steps: (1) computing local statistics densely over the whole input (2) sub-sampling these statistics by an integer striding factor. Past work has mostly focused on improving (1), by proposing better alternatives to max and average pooling that avoid aliasing (Zhang, 2019; Fonseca et al., 2021), preserve the important local details (Saeedan et al., 2018), or adapt to the training data distribution (Gulcehre et al., 2014; Lee et al., 2016). Observing that integer strides reduce resolution too quickly (e.g. a $(2, 2)$ striding reduces the output size by 75%), Graham (2014) proposed fractional max-pooling, that allows for fractional (i.e. rational) strides, allowing for integration of more downsampling layers into a network. Similarly, Rippel et al. (2015) introduce spectral pooling which, by cropping its inputs in the Fourier domain, performs downsampling with fractional strides while emphasizing lower frequencies.

While fractional strides give more flexibility in designing downsampling layers, they increase the size of an already gigantic search space. Indeed, as strides are hyperparameters, finding the best combination requires cross-validation or architecture search (Zoph & Le, 2017; Baker et al., 2017; Tan et al., 2019), which rapidly become infeasible as the number of configurations grows exponentially with the number of downsampling layers. This led Zoph & Le (2017) not to search for strides in most of their experiments. Talebi & Milanfar (2021) and Jin et al. (2021) proposed a neural network that learns a resizing function for natural images, but the scaling factor (i.e. the stride) still required cross-validation. Thus, the nature of strides as hyperparameters — rather than trainable parameters — hinders the discovery of convolutional architectures and learning strides by backpropagation would unlock a virtually infinite search space.

In this work, we introduce DiffStride, the first downsampling layer that learns its strides jointly with the rest of the network. Inspired by Rippel et al. (2015), DiffStride casts downsampling in the spatial domain as cropping in the frequency domain. However, and unlike Rippel et al. (2015), rather than cropping with a fixed bounding box controlled by a striding hyperparameter, DiffStride learns the size of its cropping box by backpropagation. To do so, we propose a 2D version of an attention window with learnable size proposed by Sukhbaatar et al. (2019) for language modeling. On five audio classification tasks, using DiffStride as a drop-in replacement to strided convolutions improves performance overall while providing interpretability on the optimal per-task receptive field. By integrating DiffStride into a ResNet-18 (He et al., 2016a), we show on CIFAR (Krizhevsky et al., 2009) and ImageNet (Deng et al., 2009) that even when initializing strides randomly, our model converges to the best performance obtained with the properly cross-validated strides of He et al. (2016a). Moreover, casting strides as learnable parameters allows us to propose a regularization that directly minimizes computation and memory usage.

2 METHODS

We first provide background on spatial and spectral pooling, and propose DiffStride for learning strides of downsampling layers. We focus on 2D CNNs since they are generic enough to be used for image (LeCun et al., 1989; Krizhevsky et al., 2012; He et al., 2016a) and audio (Amodei et al., 2016; Kong et al., 2020) processing (taking time-frequency representations as inputs). However, these methods are equally applicable to the 1D (e.g. time-series) and 3D (e.g. video) cases.

2.1 NOTATIONS

Let $x \in \mathbb{R}^{H \times W}$, its Discrete Fourier Transform (DFT) $y = \mathcal{F}(x) \in \mathbb{C}^{H \times W}$ is obtained through the decomposition on a fixed set of basis filters (Lyons, 2004):

$$\mathcal{F}(x)_{mn} = \frac{1}{\sqrt{HW}} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} x_{hw} e^{-2\pi i \left(\frac{mh}{H} + \frac{nw}{W}\right)}, \forall m \in \{0, \dots, H-1\}, \forall n \in \{0, \dots, W-1\}. \tag{1}$$

The DFT transformation is linear and its inverse is given by its conjugate $\mathcal{F}(.)^{-1} = \mathcal{F}(.)^*$. The Fourier transform of a real-valued signal $x \in \mathbb{R}^{H \times W}$ being *conjugate symmetric* (Hermitian-symmetry), we can reconstruct x from the positive half frequencies for the width dimension and

omit the negative frequencies ($y_{mn} = y^*_{(H-m) \bmod H, (W-n) \bmod W}$). In addition, the DFT and its inverse are differentiable with regard to their inputs and the derivative of the DFT (resp. inverse DFT) is its conjugate linear operator, i.e. the inverse DFT (resp. DFT). More formally, if we consider $\mathcal{L} : \mathbb{C}^{H \times W} \longrightarrow \mathbb{R}$ as a loss taking as input the Fourier representation y, we can compute the gradient of $\mathcal{L}$ with regard to x, by using the inverse DFT:

$$x \in \mathbb{R}^{H \times W}, y = \mathcal{F}(x), \frac{\partial \mathcal{L}}{\partial x} = \mathcal{F}^*(\frac{\partial L}{\partial y}) = \mathcal{F}^{-1}(\frac{\partial L}{\partial y}). \tag{2}$$

We denote by L the total number of convolution layers in a CNN architecture and each layer is indexed by l. The $\circ$ symbol represents the element-wise product between two tensors, $\lfloor . \rfloor$ is the floor operation and $\otimes$ the outer product between two vectors. S represents the stride parameters, and sg is the stop gradient operator (Bengio et al., 2013; Yin et al., 2019), defined has the identity function during forward pass and with zero partial derivatives.

2.2 Downsampling in convolutional neural networks

A basic mechanism for downsampling representations in a CNN is strided convolutions which jointly convolve inputs and finite impulse response filters and downsample the output. Alternatively, one can disentangle both operations by first applying a non-strided convolution followed by a pooling operation that computes local statistics (e.g. using an average, max (Boureau et al., 2010)) before downsampling. In both settings, downsampling does not benefit from the global structure of its inputs and can discard important information (Hinton, 2014; Saeedan et al., 2018). Moreover, and as observed by Graham (2014), the integer nature of strides only allows for drastic reductions in resolution: a 2D-convolution with strides $S = (2, 2)$ reduces the dimension of its inputs by 75%. Furthermore, stride configurations are cumbersome to explore as the number of stride combinations grows exponentially with the number of downsampling layers. This means that cross-validation can only explore a limited subset of the stride hyperparameter configurations. This limitation is likely to translate into lower performance, as Section 3.2 shows that an inappropriate choice of strides for a ResNet-18 architecture can account for a drop of $> 18\%$ in accuracy on CIFAR-100.

2.3 Spectral pooling

Energy of natural signals is typically not uniformly distributed in the frequency domain, with signals such as sounds (Singh & Theunissen, 2003), images (Ruderman, 1994) and surfaces (Kuroki et al., 2018) concentrating most of the information in the lower frequencies. Rippel et al. (2015) build on this observation to introduce *spectral pooling* which alleviates the loss of information of spatial pooling, while enabling fractional downsizing factors. Spectral pooling also preserves low frequencies without aliasing, a known weakness of spatial/temporal convnets (Zhang, 2019; Ribeiro & Schön, 2021).

We consider an input $x \in \mathbb{R}^{H \times W}$ and strides $S = (S_h, S_w) \in [1, H) \times [1, W)$. First, the DFT is computed $y = \mathcal{F}(x) \in \mathbb{C}^{H \times W}$ and for simplicity we assume that the center of this matrix is the DC component — the zero frequency. Then, a bounding box of size $\lfloor \frac{H}{S_h} \rfloor \times \lfloor \frac{W}{S_w} \rfloor$ crops this matrix around its center to produce $\tilde{y} \in \mathbb{C}^{\lfloor \frac{H}{S_h} \rfloor \times \lfloor \frac{W}{S_w} \rfloor}$. Finally, this output is brought back to the spatial domain with an inverse DFT: $\tilde{x} = \mathcal{F}^{-1}(\tilde{y}) \in \mathbb{R}^{\lfloor \frac{H}{S_h} \rfloor \times \lfloor \frac{W}{S_w} \rfloor}$. In practice, x is typically a multi-channel input (i.e. $x \in \mathbb{R}^{H \times W \times C}$) and the same cropping is applied to all channels. Moreover, since x is real-valued and thanks to Hermitian symmetry (see Section 2.1 for more details), only the positive half of the DFT coefficients are computed, which allows saving computation and memory while ensuring that the output $\tilde{x}$ remains real-valued.

Unlike spatial pooling that requires integer strides, spectral pooling only requires integer output dimensions, which allows for much more fine-grained downsizing. Moreover, spectral pooling acts as a low-pass filter over the entire input, only keeping the lower frequencies i.e. the most relevant information in general and avoiding aliasing (Zhang, 2019). However, and similarly to spatial pooling, spectral pooling is differentiable with respect to its inputs but not with respect to its strides. Thus, one still needs to provide S as hyperparameters for each downsampling layer. In this case,

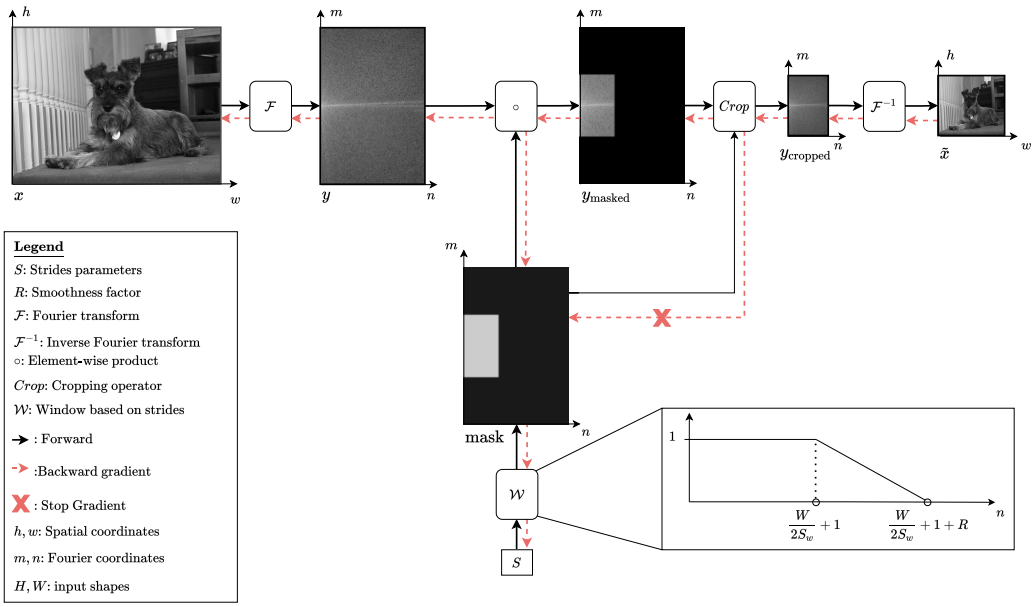


Figure 1: DiffStride forward and backward pass, using a single-channel image. We only compute the positive half of DFT coefficients along the horizontal axis due to conjugate symmetry. The zoomed frame shows the horizontal mask $\text{mask}^{w}_{(S_w,W,R)}(n)$. Here $S = (S_h, S_w) = (2.6, 3.1)$.

the search space is even bigger than with spatial pooling since strides are not constrained to integer values anymore.

2.4 DiffStride

To address the difficulty of searching stride parameters, we propose DiffStride, a novel downsampling layer that allows spectral pooling to learn its strides through backpropagation. To downsample $x \in \mathbb{R}^{H \times W}$, DiffStride performs cropping in the Fourier domain similarly to spectral pooling. However, instead of using a fixed bounding box, DiffStride learns the box size via backpropagation. The learnable box $\mathcal{W}$ is parametrized by the shape of the input, a smoothness factor R and the strides. We design this mask $\mathcal{W}$ as the outer product between two differentiable 1D masking functions (depicted in the lower right corner of Figure 1), one along the horizontal axis and one along the vertical axis. These 1D masks are directly derived from the adaptive attention span introduced by Sukhbaatar et al. (2019) to learn the attention span of self-attention models for natural language processing. Exploiting the conjugate symmetry of the coefficients, we only consider positive frequencies along the horizontal axis, while we mirror the vertical mask around frequency zero. Therefore, the two masks are defined as follows:

$$\text{mask}^{h}_{(S_h,H,R)}(m) = \min\left[\max\left[\frac{1}{R}(R + \frac{H}{2S_h} - |\frac{H}{2} - m|), 0\right], 1\right], m \in [0, H] \tag{3}$$

$$\text{mask}^{w}_{(S_w,W,R)}(n) = \min\left[\max\left[\frac{1}{R}(R + \frac{W}{2S_w} + 1 - n), 0\right], 1\right], n \in [0, \frac{W}{2} + 1] \tag{4}$$

where $S = (S_h, S_w)$ are the strides and R an hyperparameter that controls the smoothness of the mask. We build the 2D differentiable mask $\mathcal{W}$ as the outer product between the two 1D masks:

$$\mathcal{W}(S_h, S_w, H, W, R) = \text{mask}^{h}_{(S_h,H,R)} \otimes \text{mask}^{w}_{(S_w,W,R)} \tag{5}$$

We use $\mathcal{W}$ in two ways: (1) we apply it to the Fourier representation of the inputs via an element-wise product, which performs low-pass filtering (2) we crop the Fourier coefficients where the mask is zero (i.e. the output has dimensions $\lfloor \frac{H}{S_h} + 2 \times R \rfloor \times \lfloor \frac{W}{S_w} + 2 \times R \rfloor$).

The first step is differentiable with respect to strides S, however the cropping operation is not. Therefore, we apply a stop gradient operator (Bengio et al., 2013) to the mask before cropping. This

Algorithm 1: DiffStride layer

Inputs : Input $x \in \mathbb{R}^{H \times W}$, strides $S = (S_h, S_w) \in [1, H) \times [1, W)$, smoothness factor R.

Output: Downsampled output $\tilde{x} \in \mathbb{R}^{\lfloor \frac{H}{S_h} + 2 \times R \rfloor \times \lfloor \frac{W}{S_w} + 2 \times R \rfloor}$

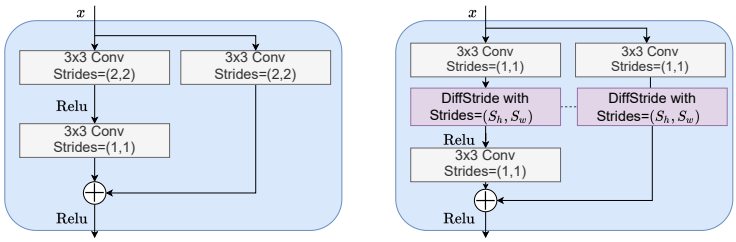
$y \longleftarrow \mathcal{F}(x)$ ▷ Project input to the Fourier domain.

$\text{mask} \longleftarrow \mathcal{W}(S_h, S_w, H, W, R)$ ▷ Construct the mask. See Equation 5.

$y_{\text{masked}} \longleftarrow y \circ \text{mask}$ ▷ Apply the mask as a low-pass filter.

$y_{\text{cropped}} \longleftarrow Crop(y_{\text{masked}}, sg(\text{mask}))$ ▷ Crop the tensor with the mask after stopping gradients.

$\tilde{x} \longleftarrow \mathcal{F}^{-1}(y_{\text{cropped}})$ ▷ Return to the spatial domain.



(a) Residual block with a strided convolution.

(b) Residual block with a shared DiffStride layer.

Figure 2: Comparison side by side of the shortcut blocks in classic ResNet architectures with strided convolutions, and with DiffStride that learns the strides of the block.

way, gradients can flow to the strides through the differentiable low-pass filtering operation, but not through the non-differentiable cropping. Finally, the cropped tensor is transformed back into the spatial domain using an inverse DFT. All these steps are summarized by Algorithm 1 and illustrated on a single channel image in the Figure 1.

During training we constrain strides $S = (S_h, S_w)$ to remain in $[1, H) \times [1, W)$. When x is a multi-channel input (i.e. $x \in \mathbb{R}^{H \times W \times C}$), we learn the same strides S for all channels to ensure uniform spatial dimensions across channels. In spatial and spectral pooling, strides are typically tied along the spatial axes (i.e. $S_w = S_h$), which we can also do in DiffStride by sharing a single learnable stride on both dimensions. However, our experiments in Section 3 show that learning specific strides for the vertical and horizontal axis is beneficial, not only when processing time-frequency representations of audio, but also — more surprisingly— when classifying natural images. Adding an hyperparameter R to each downsampling layer would conflict with the goal of removing strides as hyperparameters. Thus, not only we use a single R value for all layers, but we found no significant impact of this choice and all our experiments use $R = 4$. While we focus on the 2D case, using a single 1D mask allows deriving DiffStride in 1D CNNs, while performing the outer product between three 1D masks allows applying DiffStride to 3D inputs.

2.4.1 Residual Block with DiffStride

Unlike systems that only feed outputs of the l^{th} layer to the $(l+1)^{th}$ (Krizhevsky et al., 2012), ResNets (He et al., 2016a;b) introduce skip-connections that operate in parallel to the main branch. ResNets stack two types of blocks: (1) identity blocks that maintain the input channel dimension and spatial resolution and (2) shortcut blocks that increase the output channel dimension while reducing the spatial resolution with a strided convolution (see Figure 2a). We integrate DiffStride into these shortcut blocks by replacing strided convolutions by convolutions without strides followed by DiffStride. Besides, sharing DiffStride strides between the main and residual branches ensures that their respective outputs have identical spatial dimensions and can be summed (See Figure 2b).

2.4.2 Regularizing computation and memory cost with DiffStride

The number of activations in a network depends on the strides and learning these parameters gives control over the space and time complexity of an architecture in a differentiable manner. This contrasts with previous work, as measures of complexity such as the number of floating-point operations

(FLOPs) are typically not differentiable with respect to the parameters of a model and searching for efficient architectures is done via high-level exploration (e.g. introducing separable convolutions (Howard et al., 2017)), architecture search (Howard et al., 2019; Tan & Le, 2019) or using continuous relaxations of complexity (Paria et al., 2020).

A standard 2D convolution with a square kernel of size k^2 and C' output channels has a computational cost of $k^2 \times C \times C' \times H \times W$ when operating on $x \in \mathbb{R}^{H \times W \times C}$. Its memory usage— in terms of the number of activations to store— is $C' \times H \times W$. Considering a fixed number of channels and kernel size, both the computational complexity and memory usage of a convolution layer are thus linear functions of its input size $H \times W$. This illustrates our argument made in Section 1 that downsampling does not only improve performance by discarding irrelevant information, but also reduces the complexity of the upper layers. More importantly, in the context of DiffStride the input size $H^l \times W^l$ of layer l is determined as follows: $H^l \times W^l = \lfloor \frac{H^{l-1}}{S_h^{l-1}} + 2 \times R \rfloor \times \lfloor \frac{W^{l-1}}{S_w^{l-1}} + 2 \times R \rfloor$, which is differentiable with respect to the strides at the previous layer S^{l-1}. Furthermore, it also depends on spatial dimensions at the previous layer $H^{l-1} \times W^{l-1}$, which themselves are a function of S^{l-2}. By induction over layers, the total computational cost and memory usage are proportional to $\sum_{l=1}^{l=L} \prod_{i=1}^{l} \frac{1}{S_h^i \times S_w^i}$. Since in the context of DiffStride the kernel size and number of channels remain constant during training, we can directly regularize our model towards time and space efficiency by adding the following regularizer to our training loss:

$$\lambda J((S^l)_{l=1}^{l=L}) = \lambda \sum_{l=1}^{l=L} \prod_{i=1}^{l} \frac{1}{S_h^i \times S_w^i}, \tag{6}$$

where λ is the regularization weight. In Section 3.2, we show that training on ImageNet with different values for λ allows us to trade-off accuracy for efficiency in a smooth fashion.

3 Experiments

We evaluate DiffStride on eight classification tasks, both on audio and images. For each comparison, we keep the same architecture and replace strided convolutions by convolutions with no stride followed by DiffStride. To avoid the confounding factor of downsampling in the Fourier domain, we also compare our approach to the spectral pooling of Rippel et al. (2015), which only differs from DiffStride by the fact that its strides are not learnable.

3.1 Audio classification

Experimental setup We perform single-task and multi-task audio classification on 5 tasks: acoustic scene classification (Heittola et al., 2018), birdsong detection (Stowell et al., 2018), musical instrumental classification and pitch estimation on the NSynth dataset (Engel et al., 2017) and speech command classification (Warden, 2018). The statistics of the datasets are summarized in Table A.1. The audio sampled at 16 kHz is decomposed into log-compressed mel-spectrograms with 64 channels, computed with a window of 25 ms every 10 ms.

A 2D-CNN, based on (Tagliasacchi et al., 2019) takes these spectrograms as inputs and alternates blocks of strided convolutions along time ((3×1) kernel) and frequency ((1×3) kernel). Each strided convolution is followed by a ReLU (Glorot et al., 2011) and batch normalization (Ioffe & Szegedy, 2015). The sequence of channels dimensions is defined as $(64, 128, 256, 256, 512, 512)$ and the strides are initialized as $((2, 2), (2, 2), (1, 1), (2, 2), (1, 1), (2, 2)$ for all downsampling methods. The output of the CNN passes through a global max-pooling and feeds into a single linear classification layer for single-task, and multiple classification layers for multi-task classification. As examples vary in length, we train models on random 1 s windows with ADAM (Kingma & Ba, 2015) and a learning rate of 10^{-4} for 1 M batches, with batch size 256. Evaluation is run by splitting full sequences into 1 s non-overlapping windows and averaging the logits over windows.

Results Table 1 summarizes the results for single-task and multi-task audio classification. In both settings, DiffStride improves over strided convolutions and spectral pooling, with strided convolutions only outperforming DiffStride for acoustic scene classification in the single task setting. Table 2 shows the strides learned by the first layer of DiffStride, which downsamples mel-spectrograms

Setting	Single-task			Multi-task		
Task	Strided Conv.	Spectral	DiffStride	Strided Conv.	Spectral	DiffStride
Acoustic scenes	**99.1** $\pm$ 0.2	98.6 $\pm$ 0.1	98.6 $\pm$ 0.2	**97.7** $\pm$ 0.4	**97.7** $\pm$ 0.7	**97.7** $\pm$ 0.3
Birdsong detection	78.8 $\pm$ 0.3	79.7 $\pm$ 0.3	**81.3** $\pm$ 0.1	77.3 $\pm$ 0.2	77.8 $\pm$ 0.3	**78.6** $\pm$ 0.5
Music (instrument)	72.6 $\pm$ 0.3	72.9 $\pm$ 0.5	**75.4** $\pm$ 0.0	69.8 $\pm$ 0.4	70.4 $\pm$ 0.4	**73.0** $\pm$ 0.8
Music (pitch)	91.8 $\pm$ 0.1	90.1 $\pm$ 0.0	**92.2** $\pm$ 0.1	89.4 $\pm$ 0.3	87.6 $\pm$ 0.7	**89.9** $\pm$ 0.3
Speech commands	87.3 $\pm$ 0.1	88.5 $\pm$ 0.3	**90.5** $\pm$ 0.3	83.5 $\pm$ 0.6	83.9 $\pm$ 0.4	**86.2** $\pm$ 0.8
Mean Accuracy	85.0 $\pm$ 9.3	86.0 $\pm$ 9.2	**88.3** $\pm$ 8.7	83.5 $\pm$ 10.0	83.5 $\pm$ 9.6	**85.0** $\pm$ 8.9

Table 1: Test accuracy (% $\pm$ sd over 3 runs) for audio classification in the single (one model per task) and multi-task (one model for all tasks) settings.

	Learned Strides		Equivalent cut-off frequencies	
	Time	Frequency	Time (Hz)	Frequency (Cyc/Mel)
Acoustic scenes	1.89 $\pm$ 0.05	1.99 $\pm$ 0.03	26.25 $\pm$ 0.63	0.2448 $\pm$ 0.009
Birdsong detection	1.91 $\pm$ 0.02	1.96 $\pm$ 0.01	25.83 $\pm$ 0.36	0.2500 $\pm$ 0.000
Music (Instrument)	1.29 $\pm$ 0.06	2.12 $\pm$ 0.01	38.33 $\pm$ 1.57	0.2292 $\pm$ 0.009
Music (Pitch)	1.32 $\pm$ 0.10	1.61 $\pm$ 0.07	37.50 $\pm$ 2.72	0.3021 $\pm$ 0.018
Speech commands	1.97 $\pm$ 0.00	1.95 $\pm$ 0.01	25.00 $\pm$ 0.00	0.2500 $\pm$ 0.000
Multi-task model	1.46 $\pm$ 0.01	1.79 $\pm$ 0.03	34.17 $\pm$ 0.30	0.2708 $\pm$ 0.0074

Table 2: Learned strides (% $\pm$ sd over 3 runs) of the first layer for the single and multi-task models. The sampling rate of the input spectrogram being known (10 ms), we can convert the strides to upper cut-off frequencies (i.e. the maximum frequency kept by the lowpass-filter).

along frequency and time axes. Learning allows the strides to deviate from their initialization ($(2, 2)$) and to adapt to the task at hand. Converting strides to cut-off frequencies shows that the learned strides fall in a range showed by behavioral studies and direct neural recordings (Hullett et al., 2016; Flinker et al., 2019) to be necessary for e.g. speech intelligibility at 25 Hz (Elliott & Theunissen, 2009). Moreover, DiffStride learns different strides for the time and frequency axes. Table A.7 shows the benefits of learning a per-dimension value rather than sharing strides. Another notable phenomenon is the per-task discrepancy on NSynth, with the pitch estimation requiring faster spectral modulations (as represented by a higher cutt-off frequency along the frequency axis). Finally, multi-task models do not converge to the mean of strides, but rather to a higher value that passes more frequencies not to negatively impact individual tasks.

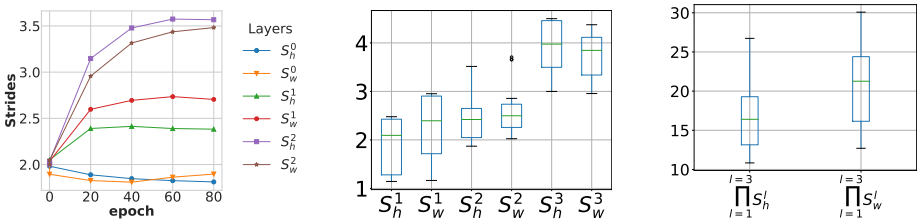
3.2 Image classification

Experimental setup We use the ResNet-18 (He et al., 2016a) architecture, comparing the original strided convolutions (see Figure 2a) to spectral pooling and DiffStride (both as in Figure 2b). We randomly sample 6 striding configurations for the three shortcut blocks of the ResNet-18, each stride being sampled in $[1, 3]$, with $(2, 2, 2)$ being the configuration of the original ResNet of He et al. (2016a). The horizontal and vertical strides are initialized equally at start. These random configurations simulate cross-validation of stride configurations to: (1) showcase the sensitivity of the architecture to these hyperparameters, (2) test our hypothesis that DiffStride can benefit from learning its strides to recover from a poor initialization. On Imagenet, as inputs are bigger than CIFAR we also allow the first ResNet-18 identity block to learn its strides which are 1 by default.

We first benchmark the three methods on the two CIFAR datasets (Krizhevsky, 2009). CIFAR10 consists of 32×32 images labeled in 10 classes with 6000 images per class. We use the official split, with 50,000 images for training and 10,000 images for testing. CIFAR100 uses the same images as CIFAR10, but with a more detailed labelling with 100 classes. We also compare the ResNet-18 architectures on the ImageNet dataset (Deng et al., 2009), which contains 1,000 classes. The models are trained on the official training split of the Imagenet dataset (1.28M images) and we report our results on the validation set (50k images). Here, we evaluate performance in terms of top-1 and top-5 accuracy. We train on all datasets with stochastic gradient descent (SGD) (Bottou et al., 1998) with a learning rate of 0.1, a batch size of 256 and a momentum (Qian, 1999) of 0.9.

	CIFAR10			CIFAR100		
Init. Strides	Strided Conv.	Spectral	DiffStride	Strided Conv.	Spectral	DiffStride
(2, 2, 2)	91.4 ± 0.2	92.4 ± 0.1	**92.5** ± 0.1	66.8 ± 0.2	**73.7** ± 0.1	73.4 ± 0.5
(2, 2, 3)	90.5 ± 0.1	92.2 ± 0.2	**92.8** ± 0.1	63.4 ± 0.5	**73.7** ± 0.2	73.5 ± 0.0
(1, 3, 1)	90.0 ± 0.4	91.1 ± 0.1	**92.4** ± 0.1	64.9 ± 0.5	70.3 ± 0.3	**73.4** ± 0.2
(3, 1, 3)	85.7 ± 0.1	90.9 ± 0.2	**92.4** ± 0.1	55.3 ± 0.8	69.4 ± 0.4	**73.7** ± 0.4
(3, 1, 2)	86.4 ± 0.1	90.9 ± 0.2	**92.3** ± 0.1	56.2 ± 0.3	69.9 ± 0.2	**73.4** ± 0.3
(3, 2, 3)	82.0 ± 0.6	89.2 ± 0.2	**92.3** ± 0.1	48.2 ± 0.2	66.6 ± 0.5	**73.6** ± 0.4
Mean accuracy	87.7 ± 3.4	91.1 ± 1.1	**92.4** ± 0.2	59.1 ± 6.7	70.6 ± 2.6	**73.5** ± 0.3

Table 3: Accuracies (% $\pm$ sd over 3 runs) on CIFAR10 and CIFAR100. First column represents strides at each shortcut block, $(2, 2, 2)$ being the configuration of (He et al., 2016a). For reference, the state-of-the-art on CIFAR10 (CIFAR100) is (Dosovitskiy et al., 2020) ((Foret et al., 2020)) with an accuracy of 99.5% (96.1%).



(a) Trajectories of the different strides $(S^l)_{l\in(1,2,3)}$ for a single run.

(b) Distribution of learned strides in ResNet-18 for the different initializations.

(c) Distribution of the final global striding factors for the different initializations.

Figure 3: Learning dynamics of DiffStride on the CIFAR10 dataset.

On CIFAR, we train models for 400 epochs dividing the learning rate by 2 at 200 epochs and again by 2 at 300 epoch, with a weight decay of 5.10^{-3}. For CIFAR, we apply random cropping on the input images and left-right random flipping. On ImageNet, we train with a weight decay of 1.10^{-3} for 90 epochs, dividing the learning rate by 10 at epochs 30, 60 and 80. We apply random cropping on the input images as in (Szegedy et al., 2015) and left-right random flipping.

Results We report the results on the CIFAR datasets and Imagenet in Tables 3 and 4 respectively, with the accuracy of our baseline ResNet-18 (first row, Strided Conv.) being consistent with previous work (Bianco et al., 2018). First, we observe that strides are indeed critical hyperparameters for the performance of a standard ResNet-18 on the three datasets, with the accuracy on CIFAR100 dropping from 66.8% average to 48.2% between the best and worst configurations. Remarkably, spectral pooling is much more robust to bad initializations than strided convolutions, even though its strides are also fixed. However, DiffStride is overall much more robust to poor choices of strides, converging consistently to a high accuracy on the three datasets, with a variance over initializations that is lower by an order of magnitude. This shows that backpropagation allows DiffStride to find a better configuration during training avoiding a cross-validation which would require 6,561 experiments for testing all combinations of strides in $[1, 3]$ on Imagenet. Tables A.5 and A.6 confirm these observations on the EfficientNet-B0 (Tan & Le, 2019) architecture.

Learning dynamics and equivalence classes Figure 3 illustrates the learning dynamics of DiffStride on CIFAR10. Figure 3a plots the strides as a function of the epoch for a run with the baseline (2,2,2) configuration as initialization. The strides all deviate from their initialization while converging rapidly, with the lower layers keeping more information while higher layers downsample more drastically. Interestingly, we discover equivalence classes: despite converging to the same accuracy (as reported in Table 3) the various initializations yield very diverse strides configurations at convergence, both in terms of total striding factor (defined as the product of strides, see Figure 3c) and of repartition of downsampling factors along the architecture (see Figure 3b). We obtain similar conclusions on CIFAR100 and Imagenet (see Figures A.1 and A.2). In the non-regularized case, it could seem counter-intuitive that minimizing the training loss yields positive stride updates,

	Top-1			Top-5		
Init. Strides	Strided Conv.	Spectral	DiffStride	Strided Conv.	Spectral	DiffStride
(1, 2, 2, 2)	68.65 ± 0.26	69.01 ± 0.19	**69.66** ± 0.06	88.5 ± 0.15	88.48 ± 0.02	**89.07** ± 0.03
(1, 1, 3, 1)	69.79 ± 0.15	**69.88** ± 0.05	68.22 ± 0.07	**89.43** ± 0.18	89.15 ± 0.07	88.10 ± 0.08
(1, 3, 1, 3)	68.86 ± 0.28	68.63 ± 0.08	**69.41** ± 0.16	88.64 ± 0.15	88.42 ± 0.01	**88.98** ± 0.04
(2, 2, 2, 3)	63.45 ± 0.09	67.16 ± 0.17	**69.53** ± 0.08	85.09 ± 0.04	87.25 ± 0.06	**89.05** ± 0.05
(2, 3, 1, 2)	65.35 ± 0.03	66.35 ± 0.24	**69.42** ± 0.06	86.27 ± 0.05	86.67 ± 0.15	**88.91** ± 0.05
(3, 3, 2, 3)	57.11 ± 0.11	64.44 ± 0.01	**69.43** ± 0.11	80.42 ± 0.11	85.22 ± 0.09	**89.03** ± 0.02
Mean accuracy	65.53 ± 4.49	67.58 ± 1.88	**69.28** ± 0.50	86.39 ± 3.15	87.53 ± 1.36	**88.85** ± 0.35

Table 4: Top-1 and top-5 accuracies ($\% \pm$ sd over 3 runs) on Imagenet, $(1, 2, 2, 2)$ being the configuration of (He et al., 2016a). For reference, state-of-the-art on Imagenet is (Dai et al., 2021) with a top-1 accuracy of 90.88%.

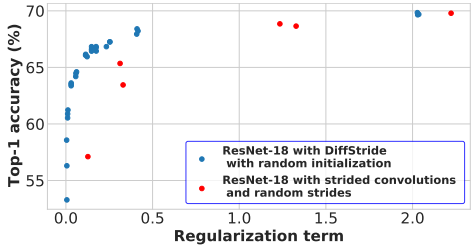


Figure 4: Top-1 accuracy (%) on the Imagenet validation set as a function of the regularization term $J((S^l)_{l=1}^{l=L})$ as defined in equation 6, after training with $\lambda \in [0.1, 10]$.

i.e. dropping more information through cropping. It highlights that loss optimization is a trade-off between preserving information (no striding, no cropping) and downscaling such that the next convolution kernel accesses a wider spatial context.

Regularizing the complexity The existence of equivalence classes suggests that DiffStride can find more computationally efficient configurations for a same accuracy. We thus train DiffStride on ImageNet using the complexity regularizer defined in Equation 6, with λ varying between 0.1 and 10, always initializing strides with the baseline $((1, 1), (2, 2), (2, 2), (2, 2))$. Figure 4 plots accuracy versus computational complexity (as measured by the value of the regularization term at convergence) of DiffStride. For comparison, we also plot the models with strided convolutions with the random initializations of Table 4, showing that DiffStride finds configurations with a lower computational cost for the same accuracy. Some of these are quite extreme, e.g. with $\lambda = 10$ a model converges to strides $((10.51, 32.23), (1.20, 2.68), (1.20, 2.04), (1.96, 4.53))$ for a 58.57% top-1 accuracy. When training a ResNet-18 with strided convolutions using the closest integer strides (i.e. $((11, 32), (1, 3), (1, 2), (2, 5))$), the model converges to a 24.54% top-1 accuracy. This suggests that performing pooling in the spectral domain is more robust to aggressive downsampling, which corroborates the remarkable advantage of spectral pooling over strided convolutions when using poor strides choices in Tables 3 and 4 despite both models having fixed strides.

Limitations Pooling in the spectral domain comes at higher computational cost than strided convolutions as it requires (1) computing a non-strided convolution and (2) a DFT and its inverse (see Table A.2). This could be alleviated by computing the convolution in the Fourier domain as an element-wise multiplication and summation over channels. Further improvements could be obtained by replacing the DFT by a real-valued counterpart, such as the Hartley transform (Zhang & Ma, 2018), which would remove the need for complex-valued operations that may be poorly optimized in deep learning frameworks. We also observe no benefits of DiffStride when training DenseNets (Huang et al., 2017), see Tables A.3 and A.4. We hypothesize that this is due to the limited number of downsampling layers, which reduces the space of stride configurations to a few, equivalent ones when sampling strides in $[1; 3]$. Finally, some hardware (e.g. TPUs) require a static computation graph. As DiffStride changes the spatial dimensions of intermediate representations—and thus the computation graph— between each gradient update, we currently only train on GPUs.

4 CONCLUSION AND FUTURE WORK

We introduce DiffStride the first downsampling layer with learnable strides. We show on audio and image classification that DiffStride can be used as a drop-in replacement to strided convolutions, removing the need for cross-validating strides. As we observe that our method discovers multiple equally-accurate stride configurations, we introduce a regularization term to favor the most computationally advantageous. In future work, we will extend the scope of applications of DiffStride, to e.g. 1D and 3D architectures. Moreover, learning strides by backpropagation opens new avenues in designing adaptive convolutional architectures, such as multi-scale models that would learn to operate at various scales in parallel by using independent branches with separate instances of DiffStride, or by predicting strides parameters of DiffStride on a per-example basis.

5 REPRODUCIBILITY STATEMENT

We describe DiffStride in details in the text as well as with Algorithm 1 and Figure 1. We mention all relevant hyperparameters to reproduce our experiments, as well as describe audio datasets in A.1. Moreover, we will release Tensorflow 2.0 code for training a ResNet-18 with strided convolutions, spectral pooling or DiffStride on CIFAR10 and CIFAR100, with DiffStride being implemented as a stand-alone, reusable Keras layer.

REFERENCES

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pp. 173–182. PMLR, 2016. 1, 2

Bowen Baker, Otkrist Gupta, Nikhil Naik, and Ramesh Raskar. Designing neural network architectures using reinforcement learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=S1c2cvqee. 2

Irwan Bello, William Fedus, Xianzhi Du, Ekin Dogus Cubuk, A. Srinivas, Tsung-Yi Lin, Jonathon Shlens, and Barret Zoph. Revisiting resnets: Improved training and scaling strategies. *ArXiv*, abs/2103.07579, 2021. 1

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013. 3, 4

Simone Bianco, Rémi Cadène, Luigi Celona, and Paolo Napoletano. Benchmark analysis of representative deep neural network architectures. *IEEE Access*, 6:64270–64277, 2018. 8

Léon Bottou et al. Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9):142, 1998. 7

Y-Lan Boureau, Jean Ponce, and Yann LeCun. A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 111–118, 2010. 3

S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 1:539–546 vol. 1, 2005. 1

Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. Very deep convolutional networks for text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 1107–1116, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-1104. 1

Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *arXiv preprint arXiv:2106.04803*, 2021. 9

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009. 2, 7

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 8

Taffeta M Elliott and Frédéric E Theunissen. The modulation transfer function for speech intelligibility. *PLoS computational biology*, 5(3):e1000302, 2009. 7

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi. Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. Technical report, apr 2017. URL http://arxiv.org/abs/1704.01279. 6

Adeen Flinker, Werner K Doyle, Ashesh D Mehta, Orrin Devinsky, and David Poeppel. Spectrotemporal modulation provides a unifying framework for auditory cortical asymmetries. *Nature human behaviour*, 3(4):393–405, 2019. 7

Eduardo Fonseca, Andrés Ferraro, and Xavier Serra. Improving sound event classification by increasing shift invariance in convolutional neural networks. *ArXiv*, abs/2107.00623, 2021. 2

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2020. 8

K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36:193–202, 1980. 1

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, pp. 1243–1252. PMLR, 2017. 1

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 315–323. JMLR Workshop and Conference Proceedings, 2011. 6

Benjamin Graham. Fractional max-pooling. *arXiv preprint arXiv:1412.6071*, 2014. 2, 3

Caglar Gulcehre, Kyunghyun Cho, Razvan Pascanu, and Yoshua Bengio. Learned-norm pooling for deep feedforward and recurrent neural networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 530–546. Springer, 2014. 2

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016a. 1, 2, 5, 7, 8, 9

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pp. 630–645. Springer, 2016b. 5

Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen. TUT Urban Acoustic Scenes 2018, Development dataset, April 2018. URL https://doi.org/10.5281/zenodo.1228142. 6

Geoffrey Hinton. What's wrong with convolutional nets. MIT Brain and Cognitive Sciences-Fall Colloquium Series, Dec 2014a. URL http . . . , 2014. 3

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *ArXiv*, abs/1704.04861, 2017. 6

Andrew G. Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1314–1324, 2019. 6

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017. 1, 9, 15

Patrick W Hullett, Liberty S Hamilton, Nima Mesgarani, Christoph E Schreiner, and Edward F Chang. Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli. *Journal of Neuroscience*, 36(6):2014–2026, 2016. 7

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015. 6

Chen Jin, Ryutaro Tanno, Thomy Mertzanidou, Eleftheria Panagiotaki, and Daniel C. Alexander. Learning to downsample for segmentation of ultra-high resolution images, 2021. 2

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015. 6

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020. 1, 2

A. Krizhevsky. Learning multiple layers of features from tiny images. 2009. 7

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 1, 2, 5

Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009. 2

Scinob Kuroki, Masataka Sawayama, and Shin'ya Nishida. Haptic texture perception on 3d-printed surfaces transcribed from visual natural textures. In *International Conference on Human Haptic Sensing and Touch Enabled Computer Applications*, pp. 102–112. Springer, 2018. 3

Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989. 1, 2

Chen-Yu Lee, Patrick W Gallagher, and Zhuowen Tu. Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree. In *Artificial intelligence and statistics*, pp. 464–472. PMLR, 2016. 2

Richard G. Lyons. *Understanding Digital Signal Processing (2nd Edition)*. Prentice Hall PTR, USA, 2004. ISBN 0131089897. 2

Hannah Muckenhirn, M. Magimai.-Doss, and S. Marcel. Towards directly modeling raw speech signal for speaker verification using cnns. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4884–4888, 2018. 1

Biswajit Paria, Chih-Kuan Yeh, N. Xu, B. Póczos, Pradeep Ravikumar, and I. E. Yen. Minimizing flops to learn efficient sparse representations. *ArXiv*, abs/2004.05665, 2020. 6

Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks : the official journal of the International Neural Network Society*, 12 1:145–151, 1999. 7

Antônio H. Ribeiro and Thomas B. Schön. How convolutional neural networks deal with aliasing. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing*. ICASSP, IEEE, 2021. 3

Oren Rippel, Jasper Snoek, and Ryan P Adams. Spectral representations for convolutional neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, pp. 2449–2457, 2015. 2, 3, 6

Daniel L Ruderman. The statistics of natural images. *Network: computation in neural systems*, 5 (4):517, 1994. 3

Faraz Saeedan, Nicolas Weber, Michael Goesele, and Stefan Roth. Detail-preserving pooling in deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9108–9116, 2018. 2, 3

Tom Sercu, Christian Puhrsch, Brian Kingsbury, and Yann LeCun. Very deep multilingual convolutional neural networks for lvcsr. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4955–4959. IEEE, 2016. 1

Nandini C Singh and Frédéric E Theunissen. Modulation spectra of natural sounds and ethological theories of auditory processing. *The Journal of the Acoustical Society of America*, 114(6):3394–3411, 2003. 3

Dan Stowell, — Mike Wood, — Hanna Pamuła, Yannis Stylianou, and Hervé Glotin. Automatic acoustic detection of birds through deep learning: the first Bird Audio Detection challenge. Technical report, 2018. URL https://arxiv.org/pdf/1807.05812.pdf. 6

Sainbayar Sukhbaatar, Édouard Grave, Piotr Bojanowski, and Armand Joulin. Adaptive attention span in transformers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 331–335, 2019. 2, 4

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015. 8

Marco Tagliasacchi, Beat Gfeller, Félix de Chaumont Quitry, and Dominik Roblek. Self-supervised audio representation learning for mobile devices. *arXiv preprint arXiv:1905.11796*, 2019. 6

Hossein Talebi and Peyman Milanfar. Learning to resize images for computer vision tasks. *ICCV*, 2021. 2

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pp. 6105–6114. PMLR, 2019. 6, 8, 16

Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2820–2828, 2019. 2

Aäron van den Oord, Nal Kalchbrenner, Lasse Espeholt, K. Kavukcuoglu, Oriol Vinyals, and A. Graves. Conditional image generation with pixelcnn decoders. In *NIPS*, 2016. 1

Pete Warden. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. Technical report, 2018. URL https://arxiv.org/pdf/1804.03209.pdf. 6

P Yin, J Lyu, S Zhang, S Osher, YY Qi, and J Xin. Understanding straight-through estimator in training activation quantized neural nets. In *International Conference on Learning Representations*, 2019. 3

Neil Zeghidour, Qiantong Xu, Vitaliy Liptchinsky, Nicolas Usunier, Gabriel Synnaeve, and Ronan Collobert. Fully convolutional speech recognition. *arXiv preprint arXiv:1812.06864*, 2018. 1

Hao Zhang and Jianwei Ma. Hartley spectral pooling for deep learning. *arXiv preprint arXiv:1810.04028*, 2018. 9

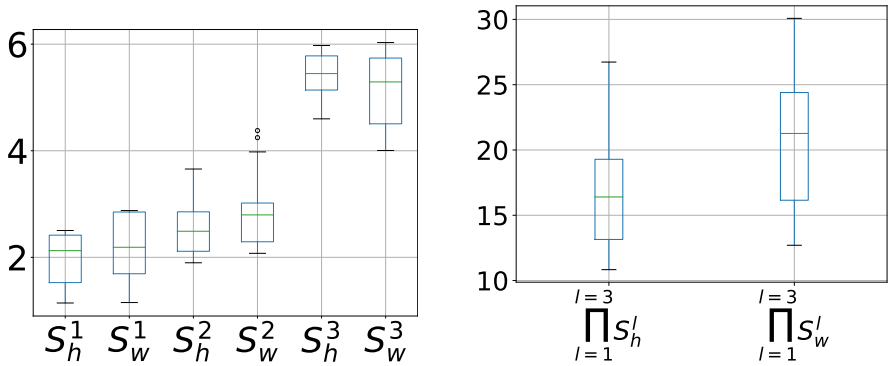
Richard Zhang. Making convolutional networks shift-invariant again. In *ICML*, 2019. 1, 2, 3

Barret Zoph and Quoc V. Le. Neural architecture search with reinforcement learning. 2017. URL https://arxiv.org/abs/1611.01578. 2

A APPENDIX

Table A.1: Datasets used for audio classification. Default train/test splits are always adopted.

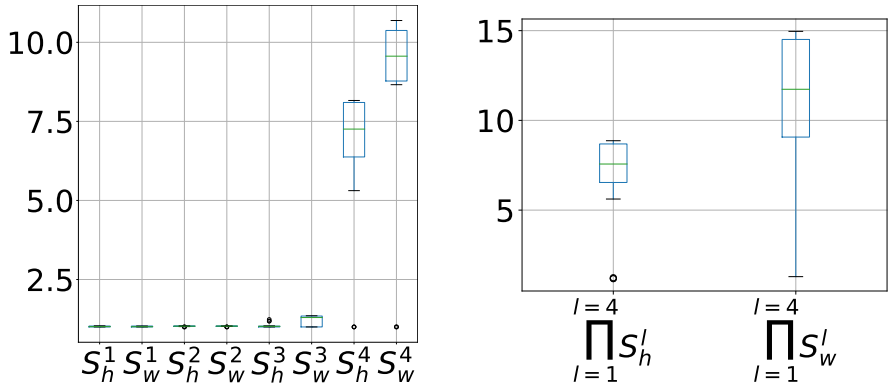
Task	Name	Classes	Train examples	Test examples
Acoustic scenes	TUT Urban 2018	10	7,829	810
Birdsong detection	DCASE2018	2	32,129	3,561
Music (instrument)	Nsynth	11	289,205	12,678
Music (pitch)	Nsynth	128	289,205	12,678
Speech commands	Speech commands	35	84,771	10,700



(a) Distribution of learned strides in ResNet-18 for the different initializations.

(b) Distribution of the final global striding factors for the different initializations.

Figure A.1: Learned strides by DiffStride on the CIFAR100 dataset.



(a) Distribution of learned strides in ResNet-18 for the different initializations.

(b) Distribution of the final global striding factors for the different initializations.

Figure A.2: Learned strides by DiffStride on the Imagenet dataset.

Analysis of strides learned on CIFAR100 and ImageNet In Figure A.1 (Figure A.2), we show the distributions of learned strides and the global striding factor at convergence on CIFAR100 (Imagenet), starting from random stride initializations. On CIFAR100, we observe equivalence classes, i.e. model that learns various stride configurations for a same accuracy. On Imagenet, even though we also observe a significant variance of the global striding factor, models tend to downsample only in the upper layers. Striding late in the architecture comes at a higher computational cost, which furthermore justifies regularizing DiffStride to reduce complexity as shown in Section 3.2.

Time and space complexity in practice While Figure 4 reports theoretical estimates of computational complexity based on stride configurations, both spectral pooling and DiffStride require computing a DFT and its inverse. Moreover, DiffStride requires accumulating gradients with respect to the strides during training. Table A.2 reports the duration and peak memory usage of the multi-task architecture described in 3.1, for a single batch. Replacing strided convolutions with spectral pooling increases the wall time by 32% due to the DFT and inverse DFT, while the peak memory usage is almost unaffected. DiffStride furthermore increases the wall time (by 43% w.r.t strided convolutions) as the backward pass is more expensive. Similarly, it almost doubles the peak memory usage. However, in inference, DiffStride does not need to compute and store gradients w.r.t. the strides, thus the time and space complexity become identical to that of spectral pooling.

		Strided Conv.	Spectral	DiffStride
Training	Time/step	1.0	1.32	1.43
	Peak memory (GB)	1.0	1.02	1.98
Inference	Time/step	1.0	1.32	1.32
	Peak memory (GB)	1.0	1.02	1.02

Table A.2: Per-step time and peak memory usage of Spectral Pooling and DiffStride relative to strided convolutions, on a V100 GPU. During training, a "step" is the forward and backward pass for a single batch, while in inference it only involves a forward pass.

DenseNet experiments on CIFAR We also evaluate DiffStride in DenseNet (Huang et al., 2017), especially the DenseNet-BC architecture with a depth of 121 and a growth rate of 32. The DenseNet architecture halves spatial dimensions during transition blocks. We replace the 2D average pooling in the transition blocks by spectral pooling or DiffStride. The considered architecture for DenseNet has two downsampling steps. We run a similar experiment as in 3.2 with random strides between the dense blocks on the two CIFAR datasets. We observe that initializing strides randomly does not affect the performance of the standard Densenet-BC architecture with average pooling. Consequently, DiffStride does not improve over alternatives.

Init. Strides	Average Pooling	Spectral	DiffStride
(2, 2)	**92.3** $\pm$ 0.2	91.5 $\pm$ 0.2	91.5 $\pm$ 0.1
(1, 2)	**91.8** $\pm$ 0.2	90.5 $\pm$ 0.5	91.1 $\pm$ 0.3
(1, 3)	**92.0** $\pm$ 0.2	91.0 $\pm$ 0.1	91.6 $\pm$ 0.4
(2, 3)	92.0 $\pm$ 0.3	92.1 $\pm$ 0.2	**92.2** $\pm$ 0.1
(3, 1)	91.6 $\pm$ 0.2	91.5 $\pm$ 0.4	**91.7** $\pm$ 0.2
Mean accuracy	**91.9** $\pm$ 0.3	91.3 $\pm$ 0.6	91.6 $\pm$ 0.4

Table A.3: Accuracies (% $\pm$ sd over 3 runs) for CIFAR10 for each downsampling method with the DenseNet-BC architecture. First column represents strides at each transition block, $(2, 2)$ being the configuration of (Huang et al., 2017).

Init. Strides	Average Pooling	Spectral	DiffStride
(2, 2)	**69.1** $\pm$ 0.6	68.6 $\pm$ 0.6	68.2 $\pm$ 0.2
(1, 2)	67.7 $\pm$ 0.3	67.2 $\pm$ 0.3	**68.0** $\pm$ 0.1
(1, 3)	67.8 $\pm$ 0.4	67.3 $\pm$ 0.2	**68.0** $\pm$ 0.7
(2, 3)	67.6 $\pm$ 0.3	**69.0** $\pm$ 0.8	**69.0** $\pm$ 0.2
(3, 1)	68.4 $\pm$ 0.6	69.1 $\pm$ 0.4	**69.7** $\pm$ 0.6
Mean accuracy	68.1 $\pm$ 0.7	68.2 $\pm$ 1.0	**68.6** $\pm$ 0.8

Table A.4: Accuracies (% $\pm$ sd over 3 runs) for CIFAR100 for each downsampling method with the DenseNet-BC architecture. First column represents strides at each transition block, $(2, 2)$ being the configuration of (Huang et al., 2017).

EfficientNet experiments on CIFAR We evaluate DiffStride in an EfficientNet-B0 architecture (Tan & Le, 2019), a lightweight model discovered by architecture search. This architecture has seven strided convolutions. Unlike Tan & Le (2019), we do not pre-train on ImageNet, but rather train from scratch on CIFAR, which explains the lower accuracy of the baseline. As the model has seven downsampling layers, we rescale the images from 32×32 to 128×128, and only sample strides in $[1; 2]$. We run a similar experiment as in 3.2 with random strides on the two CIFAR datasets. Consistently with the results obtained with a ResNet-18, spectral pooling is much more robust to poor strides than strided convolutions, with DiffStride outperforming all alternatives.

Init. Strides	Average Pooling	Spectral	DiffStride
$(1, 2, 2, 2, 1, 2, 1)$	87.2 ± 0.1	90.4 ± 0.2	**91.1** ± 0.0
$(1, 1, 2, 2, 2, 1, 1)$	89.7 ± 0.1	**90.9** ± 0.3	**90.9** ± 0.1
$(1, 2, 2, 2, 2, 2, 1)$	83.7 ± 0.2	90.0 ± 1.0	**90.8** ± 0.1
$(2, 1, 2, 1, 2, 1, 1)$	89.2 ± 0.2	90.4 ± 0.4	**91.1** ± 0.1
Mean accuracy	87.5 ± 2.5	90.4 ± 0.6	**90.9** ± 0.1

Table A.5: Accuracies (% $\pm$ sd over 3 runs) for CIFAR10 for each downsampling method with the EfficientNet-B0 architecture. First column represents strides at each strided convolution, with $(1, 2, 2, 2, 1, 2, 1)$ being the configuration of (Tan & Le, 2019).

Init. Strides	Average Pooling	Spectral	DiffStride
$(1, 2, 2, 2, 1, 2, 1)$	55.2 ± 0.3	66.0 ± 0.6	**66.6** ± 0.5
$(1, 1, 2, 2, 2, 1, 1)$	62.0 ± 0.7	66.4 ± 0.6	**66.6** ± 0.3
$(1, 2, 2, 2, 2, 2, 1)$	46.8 ± 2.0	65.9 ± 0.5	**66.3** ± 0.7
$(2, 1, 2, 1, 2, 1, 1)$	60.4 ± 0.1	65.5 ± 0.1	**67.0** ± 0.1
Mean accuracy	56.1 ± 6.3	65.9 ± 0.5	**66.6** ± 0.5

Table A.6: Accuracies (% $\pm$ sd over 3 runs) for CIFAR100 for each downsampling method with the EfficientNet-B0 architecture. First column represents strides at each strided convolution, with $(1, 2, 2, 2, 1, 2, 1)$ being the configuration of (Tan & Le, 2019).

Ablation study: learning a per-dimension stride or a shared one We perform multi-task audio classification with either learning a single stride value for each DiffStride layer, or a different one for the time and frequency axes. The overall performance across tasks is improved when learning a different stride value for each dimension (See Table A.7).

Setting	Multi-task	
Task	Shared stride	Different strides
Acoustic scenes	97.4 ± 0.3	**97.7** ± 0.3
Birdsong detection	**79.7** ± 1.0	78.6 ± 0.5
Music (instrument)	70.7 ± 0.5	**73.0** ± 0.8
Music (pitch)	86.7 ± 0.5	**89.9** ± 0.3
Speech commands	**86.8** ± 0.1	86.2 ± 0.8
Mean Accuracy	84.3 ± 9.2	**85.0** ± 8.9

Table A.7: Test accuracy (% $\pm$ sd over 3 runs) for audio classification in the multi-task (one model for all tasks) with DiffStride, when learning a single stride value (Shared stride) per layer, or a different one for each dimension (Different strides).

In Table A.1, we report the regular expressions, that allows to parse annotations .

Tab. A.1.: Table of regex rules to parse annotations based on the SpontaneousCHAT protocol, listed in Table 2.6

Token	Value/Regex
QUESTION_MARK	`"?"`
TRAILING_OFF	`"+..."`
SELF_INTERRUPTION	`"+//."`
SUSPENDED_QUESTION	`"+..?"`
PAUSE	`"(.)"`
UNINTELLIGIBLE_WORD	`"xxx"`
FILLER	`r"&-[a-z:]+"`
PHONOLOGICAL_FRAGMENT	`r"&\+\S +"`
NON_LINGUISTIC_ADDITION	`r"&=[A-Za-z+]+"`
WORD	`r"[\w :^\-↫~/@≠]+"`
PURE_WORD	`r"[\w -]+"`
OMITTED_WORD	`r"\([\w -]+\)"`
REPETITION_TAG	`r"\[x[0-9]+\]"`

Colophon

This thesis was typeset with LaTeX 2$_\varepsilon$. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at http://cleanthesis.der-ric.de/.

RÉSUMÉ

Les maladies neurodégénératives sont un enjeu social majeur et une priorité de santé publique dans le monde entier. La maladie de Huntington (MH) est une maladie rare d'origine génétique qui provoque des troubles cognitifs, comportementaux et moteurs dus à des lésions cérébrales, notamment au niveau du striatum. Les personnes porteuses de la mutation génétique de la MH ont une phase pré-symptomatique de plusieurs décennies au cours de laquelle elles n'ont aucun trouble neurologique avant que la phase symptomatique n'apparaisse. Les symptômes de cette maladie ont de nombreuses implications dans le quotidien, avec une perte progressive d'autonomie jusqu'au décès du patient. Cela fait de la MH un modèle potentiel de maladies neurodégénératives qui pourrait conduire au développement de nouveaux outils de suivi clinique. Le suivi médical actuel pour la MH est onéreux et oblige le patient à se rendre régulièrement à l'hôpital, générant une charge humaine et financière importante. L'objectif de cette thèse est de développer et de valider de nouvelles méthodes computationnelles pour le suivi automatique des individus avec la MH, grâce à l'analyse de leurs productions langagières. En effet, la production du langage oral fait appel à diverses compétences cognitives, sociales et motrices, et sa réalisation est influencée par l'état mental et neurologique. Notre hypothèse est qu'à travers l'inspection de la parole et de son contenu nous pouvons évaluer ces différentes compétences et traits. À ce jour, l'analyse des troubles de la parole et du langage pour la MH n'est pratiquée que dans quelques services cliniques et équipes spécialisées, à petite échelle. De plus, la capacité des marqueurs issus du langage oral à prédire les différents symptômes de la MH n'a pas été explorée.

Par conséquent, dans cette thèse, nous avons conçu une batterie complète de tâches qui testent plusieurs niveaux de production de la parole, ainsi qu'un protocole d'annotations complet qui reste analysable par un programme informatique. Cette batterie a été conçu pour obtenir un tableau clinique complet du langage parlé en MH, qui fait varier la cible linguistique, la charge cognitive, le contenu émotionnel, les sujets du discours. Pour accélérer le processus d'annotations, nous avons conçu et développé un logiciel open source pour gérer les campagnes d'annotations. Nous avons pu ainsi collecter la plus grande base de données de productions langagières, à ce jour, avec 125 entretiens annotés pour 3 groupes d'individus: des témoins sains, des individus porteurs du gène qui cause la MH mais sans symptômes cliniques et des individus symptomatiques avec la MH à différents stades. Par ailleurs, nous avons également formalisé et implémenté les voies de communication introduites par H. Clark, qui permettent d'analyser la parole dans des échanges spontanés. Ensuite, pour accélérer et automatiser les annotations, nous avons développé et validé des algorithmes d'apprentissage profond pour reconnaître les tours de parole lors des entretiens et reconnaître les voies de communication directement à partir de l'audio. Enfin, grâce à cette nouvelle base de données, nous avons évalué les capacités des marqueurs issus de la parole à prédire les différents symptômes de la MH. Nous avons notamment découvert que les marqueurs rythmiques et articulatoires, lors des tâches à charge cognitive plus élevée, pouvaient prédire les composantes globales, motrices, fonctionnelles et cognitives de la maladie et corrélaient avec le volume du striatum, la marque neurale de l'évolution de la maladie. Nous avons également proposé de nouvelles méthodologies pour examiner la production de la parole émotionnelle dans la MH. Nous avons ainsi découvert que les individus avec des symptoms cliniques de la MH ont à la fois des troubles vocaux et linguistiques lors de la production de la parole émotionnelle.

MOTS CLÉS

Traitement de la parole naturelle, Maladie de Huntington, Apprentissage Machine, Neurologie, Traitement du signal, Médecine de précision.

ABSTRACT

Neurodegenerative diseases are a major social issue and public health priority worldwide. Huntington Disease (HD) is a rare disease of genetic origin that causes cognitive, behavioural and motor disorders due to brain lesions, in particular in the striatum. People with the genetic mutation of HD have a pre-symptomatic phase of several decades during which they have no neurological disorder before the symptomatic phase occurs. The symptoms of this disease have many implications in the life activities of the patient, with a gradual loss of autonomy, until the death of the patient. This makes HD a potential model of neurodegenerative diseases that could lead to the development of new clinical monitoring tools. The current medical monitoring in HD is expensive and requires the patient to travel regularly to the hospital, generating a significant human and financial burden. The purpose of this thesis is to develop and validate new computational methods for automatically monitoring Huntington's Disease individuals, thanks to the analysis of their spoken language productions. Spoken language production invokes various cognitive, social and motor skills, and its realisation is influenced by the mental state of the individual. Our hypothesis is that through the inspection of the produced speech and its content we can assess these different skills and states. To this date, the analysis of spoken language disorders in HD is only performed in a few clinical departments and specialised research teams, at a small scale without classic clinical validation. In addition, the potential of spoken language markers to predict the different symptoms in HD have not been explored.

Therefore in this thesis, we designed a comprehensive spoken language battery, along with a complete annotation protocol that is parsable by a computer program. This battery measures different parameters to obtain a wide clinical picture of spoken language in HD, that varies the linguistic target, the cognitive load, the emotional content, the topics and the materials of the discourse. To speed up the annotations protocol, we designed and developed open-source software to manage linguistic annotation campaigns. This allowed us to collect what is, to the best of our knowledge, the largest database of fine-grained annotated spoken language productions in HD, with 125 annotated interviews of 3 groups of individuals: healthy controls, premanifest individuals carrying the gene that causes HD and manifest HD at different stages. Besides, we also formalized and implemented the tracks of communication introduced by H. Clark, which allow analyzing the use of spoken language in spontaneous exchanges for HD individuals. Then, to speed up and automate the annotation process, we developed and validated machine learning methods to recognise turn-takings and identify these tracks of communication directly from speech. Finally, thanks to this new database, we assessed the capabilities of spoken language markers to predict the different symptoms in HD. We especially found out that rhythm and articulatory markers extracted from tasks with a cognitive load can predict accurately the global, motor, functional and cognitive components of the disease. We additionally found significant correlations between silence statistics and the volume of the striatum, the neuro-anatomical hallmark of the disease progress. In spontaneous productions, we found that the ratio of tracks of communication was different between HD individuals and other groups. The primary track was diminished, the timing ratio of secondary presentation (filled pauses) also decreased and the timing of incidental elements (ex: vocal noises, audible respiration) greatly increased. We also proposed new methodologies to examine the emotional speech production in HD. Finally, we found out that the manifest individuals with HD have both vocal and linguistic impairments during emotional speech production.

KEYWORDS

Speech and language processing, Huntington's Disease, Neurology, Machine Learning, Signal Processing, Precision Medicine.