La reconnaissance automatique de phonèmes est-elle réellement adaptée pour l'analyse de la parole spontanée ?

Vincent P. Martin, Colleen Beaumard, Charles Brazier, Jean-Luc Rouas, Yaru Wu

La reconnaissance automatique de phonèmes est-elle réellement adaptée pour l'analyse de la parole spontanée ?

Vincent P. Martin[1] Colleen Beaumard[2,3] Charles Brazier[2]
Jean-Luc Rouas[2] Yaru Wu[4]
(1) DDP Research Unit, Department of Precision Health, LIH, L-1445 Strassen, Luxembourg
(2) Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence, France
(3) Univ. Bordeaux, CNRS, SANPSY, UMR 6033, F-33000 Bordeaux, France
(4) Université de Caen, CRISCO/UR4255, France
vincentp.martin@lih.lu, {colleen.beaumard, charles.brazier,jean-luc.rouas}@labri.fr, yaru.wu@unicaen.fr

RÉSUMÉ

La transcription phonémique automatique de la parole spontanée trouve des applications variées, notamment dans l'éducation et la surveillance de la santé. Ces transcriptions sont habituellement évaluées soit par la précision de l'identification des phonèmes, soit par leur segmentation temporelle. Jusqu'à présent, aucun système n'a été évalué simultanément sur ces deux tâches. Cet article présente l'évaluation d'un système de transcription phonémique du français spontané (corpus Rhapsodie) basé sur Kaldi. Ce système montre de bons résultats en identification des phonèmes et de leurs catégories, avec des taux d'erreur de 19,2% et 13,4% respectivement. Il est cependant moins performant en segmentation, manquant en moyenne 40% de la durée des phonèmes et 34% des catégories. Les performances s'améliorent avec le niveau de planification de la parole. Ces résultats soulignent le besoin de systèmes de transcription phonémique automatique fiables, nécessaires à des analyses plus approfondies de la parole spontanée.

ABSTRACT

Is automatic phoneme recognition suitable for spontaneous speech analysis ?

Automatic phonemic transcription of spontaneous speech has a wide range of applications, particularly in education and health monitoring. These transcriptions are usually evaluated either by the accuracy of phoneme identification or by their temporal segmentation. To date, no system has been evaluated simultaneously on both tasks. This article presents the evaluation of a Kaldi-based phonetic transcription system on spontaneous French in the Rhapsodie database. The system performed well in phoneme and category identification, with error rates of 19.2% and 13.4% respectively. However, its segmentation performances are low, missing on average 40% of phoneme duration and 34% of categories. Performance improves with the level of speech planning. These results underline the need for reliable automatic phonetic transcription systems, necessary for more in-depth analyses of spontaneous speech.

MOTS-CLÉS : Reconnaissance Automatique de la Parole, Reconnaissance de phonèmes, Segmentation de phonèmes.

KEYWORDS: Speech recognition, Phoneme recognition, Phoneme segmentation.

1 Introduction

1.1 Contexte

Depuis les modèles fondateurs des années 1970 basés sur des modèles statistiques (Baker, 1975; Jelinek, 1976), le domaine de la transcription automatique de la parole a considérablement évolué, jusqu'au modèles d'apprentissage profond bout-en-bout (Alharbi *et al.*, 2021). Alors que les systèmes de transcription étaient initialement basés sur la modélisation des phonèmes, les modèles les plus récents et les plus performants fournissent directement une transcription des mots, à partir d'estimations des mots directement, de portions de mots, ou de caractères (Alharbi *et al.*, 2021). Cependant, un petit sous-ensemble de cas d'usages nécessitent encore une transcription phonémique, afin d'évaluer la précision de la prononciation lors de l'apprentissage d'une langue, de détecter les mots qui sont hors vocabulaire, (par exemple, pour la parole des enfants (Gelin *et al.*, 2021)) ou d'évaluer l'impact des pathologies sur l'articulation (Huckvale *et al.*, 2023; Beaumard *et al.*, 2023). Le domaine de la transcription automatique de la parole en phonèmes est partagé entre deux sous-applications différentes : d'un côté, l'estimation correcte de la séquence de phonèmes, mesurée par le Pourcentage d'Erreur de Phonèmes (PER) ; de l'autre, la segmentation correcte du fichier audio en phonèmes, délimitant leurs emplacements (généralement mesurée en termes de sensibilité, de spécificité et de score F1).

Concernant l'estimation automatique des phonèmes, le modèle le plus récent pour le français à notre connaissance repose sur le modèle Wav2Vec2 de Meta ré-entraîné plus finement sur Common Voice v13 et publié par Huggingface [1]. Ce système atteint des PER de 5,5% et 4,4% sur Common Voice v13 et Librispeech respectivement, qui sont tous deux des corpus de parole lue. Nous n'avons trouvé aucune évaluation récente concernant les performances de transcription phonémique pour la parole spontanée en français. En ce qui concerne la segmentation des signaux de parole en phonèmes, les dernières approches comprennent l'apprentissage auto-supervisé (Strgar & Harwath, 2023) et les modèles autorégressifs(Kim & Choi, 2023), atteignant des scores F1 autour de 90% sur les corpus TIMIT et Buckeye. Cependant, à notre connaissance, ces systèmes n'ont été évalués que sur l'une de deux tâches de détection ou de segmentation des phonèmes. Aucun système n'a donc été évalué conjointement sur les deux tâches.

1.2 Objectif

Notre objectif est d'évaluer un système standard de transcription de la parole en phonèmes pour différents styles de parole spontanée en français, tant en termes de reconnaissance des phonèmes (taux d'erreur phonémique) que de précision temporelle (rappel, précision et score F1). Cette double évaluation sera utile pour tout type d'analyse des phonèmes extraits automatiquement, que ce soit par exemple pour l'évaluation de la prononciation pour des apprenants de langues ou l'analyse de la parole pathologique.

Ce document est organisé comme suit. Nous introduisons le corpus Rhapsodie, notre modèle et les métriques de performance dans la Section 2. Nous rapportons et discutons les résultats du système conçu dans la Section 3 et concluons dans la Section 4.

1. https ://huggingface.co/Cnam-LMSSC/wav2vec2-french-phonemizer

2 Méthode

2.1 Système de reconnaissance automatique de la parole

Dans cette étude, nous utilisons un système de transcription automatique de la parole entraîné avec la boîte à outils Kaldi (Povey *et al.*, 2011). Il s'agit d'un modèle TDNN-HMM entraîné avec la fonction LF-MMI. Le réseau neuronal est un réseau à délai temporel échantillonné avec 7 couches TDNN, chacune ayant 1024 unités. La valeur de pas temporel est réglée sur 1 pour les trois premières couches, 0 pour la quatrième, et 3 pour les suivantes. Le modèle acoustique est basé sur un vecteur MFCC de haute résolution à 40 dimensions concaténé avec un i-vecteur de 100 dimensions (Gupta *et al.*, 2014). Les données d'apprentissage sont un sous-ensemble des corpus ESTER 1 et 2 (Galliano *et al.*, 2009) (discours radiophoniques). Ce système atteint un taux d'erreur de mots de 13.7% sur le sous-corpus de test d'ESTER (Boyer, 2021), ce qui est proche des performances systèmes état-de-l'art sur le même corpus (taux d'erreurs en mots légèrement inférieur à 12% (Heba, 2021)). Les phonèmes et leur alignement temporel sont obtenus avec la commande `lattice-align-phones`, qui permet d'obtenir une annotation et une segmentation en 35 phonèmes standards.

Source	Description	#enr.	#loc.	Durée	Style
CFPP2000	*Corpus de Français Parlé Parisien*, interviews à propos des quartiers de Paris (Branca-Rosoff & Lefeuvre, 2016)	3	2 H / 5 F	15min.	Semi-spt (3)
Avanzi	Collecté par M. Avanzi pour l'étude intonosyntaxique des phénomènes macrosyntaxiques (Avanzi, 2013)	18	7 H / 15 F	14 min	Spontané (17)
Lacheret	Collecté pour la modélisation continue et fonctionnelle du français (Lacheret-Dujour, 2003)	2	3 H / 1 F	9 min.	Planifié (1), Spontané (1)
Mertens	Collecté pour la modélisation intonosyntaxique du français (Mertens, 1987)	2	4 H / 0 F	10 min	Planifié (1), Semi-spt (1)
C-Prom	Collecté pour étudier les proéminences syllabiques en français (Avanzi & Simon, 2010)	1	1 H / 0 F	3 min.	Planifié (1)
ESLO	*L'Enquête Sociolinguistique à Orléans*, recueillie à Orléans, France en 1968-74 avec un objectif sociolinguistique (Eshkol-taravella *et al.*, 2011)	1	2 H / 0 F	7 min.	Planifié (1)
PFC	*Phonologie du français contemporain*, conversations dirigées entre un sujet et un intervieweur et conversations informelles entre deux personnes appartenant à un réseau social dense, (Durand *et al.*, 2009)	3	2 H / 4 F	14 min.	Spontané (3)
Film	Monologues dans lesquels 7 intervenants différents sont invités, dans un cadre informel, à décrire une courte scène d'un film de Charlie Chaplin collectée pour le projet Rhapsodie	7	4 H / 3 F	9 min.	Spontané (7)
Professionnel	Monologues et dialogues dans un contexte professionnel collectés pour le projet Rhapsodie	3	2 H / 2 F	8 min.	Spontané (3)
Télédiffusion	14 monologues diffusés, dialogues et conversations téléchargés d'Internet pour le projet Rhapsodie	14	22 H / 6 F	67 min.	Planifié (7), Spontané (6)
Tous		54	49 H / 36 F	2h 41m	

TABLE 1 – Description du corpus Rhapsodie : nombre d'échantillons, nombre de locuteurs, durée du corpus. *Semi-spt* : Semi-spontané. Les trois styles de parole spontanée sont ceux fournis dans les métadonnées du corpus.

2.2 Corpus Rhapsodie

Nos analyses ont été réalisées sur le corpus Rhapsodie, un corpus multigenre de français parlé (Lacheret-Dujour *et al.*, 2019). Le corpus contient au total trois heures de parole (∼33000 mots), composées de 54échantillons courts (5 minutes en moyenne). Il inclut des interviews en face à face, des émissions de radio et de télévision, pour un total de 89 locuteurs. Les transcriptions phonétiques sont obtenues en utilisant un outil de conversion automatique graphème-vers-phonème (g2p) (Easyalign (Goldman, 2011) dans Praat (Boersma, 2001)), suivi d'une vérification manuelle (Lacheret *et al.*, 2014). Les pauses ont été détectées automatiquement. Deux enregistrements, D0001 et D1003 (respectivement dans les sous-corpus CFPP2000 et Rhapsodie Professionnel) ont été exclus en raison de leur mauvaise qualité acoustique. Un autre fichier, M2006 (sous-corpus Télédiffusion), a été exclu en raison d'erreurs dans les frontières temporelles de l'annotation phonétique de la vérité terrain. Tous les résultats et statistiques ultérieurs n'incluent pas ces fichiers. Les différentes sources de données utilisées dans le corpus Rhapsodie sont décrites dans le Tableau 1.

Le corpus Rhapsodie contient plusieurs variables pour représenter les caractéristiques discursives de chaque échantillon. Nous nous concentrons dans cette étude sur l'analyse des résultats du système de transcription phonémique automatique en fonction du degré de planification de la parole, tel qu'annoté dans les métadonnées du corpus : parole planifiée, semi-spontanée ou spontanée.

2.3 Vérité terrain : phonèmes et catégories

Les 54 fichiers du corpus représentent un total de 96756 phonèmes, qui ont une durée moyenne de 81.2ms. Pour faciliter l'interprétation des résultats, les 35 phonèmes ont été regroupés en 10 catégories standards : 5 catégories pour les consonnes (occlusives, fricatives, nasales, liquides, glissantes) et 5 catégories pour les voyelles (antérieures arrondies, antérieures non arrondies, centrales, postérieures arrondies, nasales).

2.4 Métriques de performance

Le Pourcentage d'Erreur Phonétique (PER) est la métrique utilisée dans le domaine de la reconnaissance automatique de la parole pour mesurer la précision de la transcription phonémique. Le PER est calculé par la somme du nombre de substitutions (S), d'insertions (I) et de suppressions (D) de phonèmes dans l'hypothèse fournie par le système par rapport au nombre total de phonèmes dans la transcription de référence (N) : $PER = 100 \times (S + I + D)/N$

Une faible valeur de PER indique une bonne précision dans la reconnaissance automatique des phonèmes. Le PER ne prend cependant pas en compte les erreurs dues aux mauvais placement des frontières puisqu'il ne prend en compte que la séquence de symboles phonétiques.

Pour compléter le PER, nous souhaitons également mesurer les performances du système en termes de durée. À cette fin, nous utilisons l'outil *trackeval* (marge d'erreur = 0) qui a été utilisé lors de la campagne d'évaluation ESTER pour estimer les performances de la détection d'événements audio (Galliano *et al.*, 2009). Ici, les événements à identifier correctement sont les phonèmes. Ce faisant, nous avons mesuré trois indicateurs :

— *le score de Rappel*, qui est le rapport de la durée de détection correcte d'un phonème sur la durée totale de cet événement dans le fichier de référence : $R = \hat{d}_{corr}(phon)/d_{ref}(phon)$

— *le score de Précision*, qui est le rapport entre la durée de détection correcte d'un phonème sur la durée totale de détection de ce phonème (y compris les insertions) : $P =$

$\hat{d}_{corr}(phon)/\hat{d}_{corr+ins}(phon)$

— et *le F-score*, une métrique combinant la Précision et le Rappel en une seule métrique : $F = 2 \times (P \times R)/(P + R)$

Une score de *Rappel* élevé indique que le phonème considéré est bien détecté, tandis qu'un score de *Précision* élevé montre que le système détecte le phonème principalement lorsqu'il est réellement présent (peu d'insertions). Idéalement, un bon système de segmentation des phonèmes doit donc avoir à la fois des scores de *Précision* et de *Rappel* élevé, et donc une *Mesure-F* élevée.

Les métriques de PER et de segmentation sont également recalculées sur les catégories phonétiques.

3 Résultats et discussion

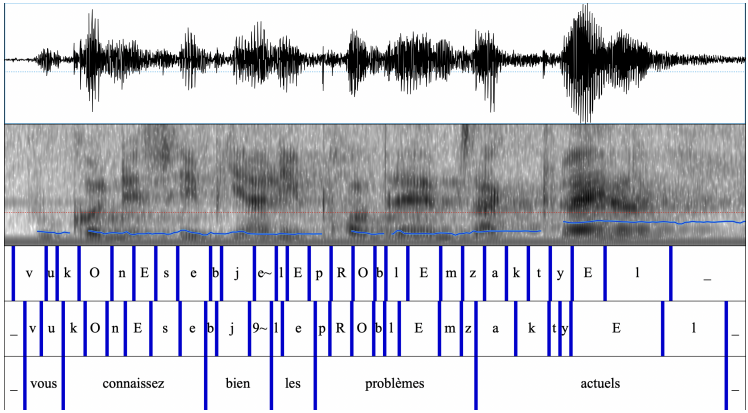

FIGURE 1 – Exemple de sortie du système de reconnaissance phonétique. En haut : résultats de la détection automatique ; au milieu : annotation phonétique de référence ; en bas : transcription en mots.

La Figure 1 montre un exemple de sortie de notre système automatique de reconnaissance phonétique. Cet exemple est un extrait du fichier D1001 du corpus Rhapsodie, provenant de la base de données ESLO (Eshkol-taravella *et al.*, 2011). Cet enregistrement date de 1968 et est un monologue d'un locuteur masculin.

Dans cet exemple, la plupart des phonèmes sont correctement détectés à l'exception de deux substitutions, dans deux couples de phonèmes proches l'un de l'autre : /9~/ a été substitué par /e~/ ; et /e/ par /E/. Étant donné les bonnes performances d'identification des phonèmes, si le système de reconnaissance phonétique se comporte de la même manière pour tous les fichiers, nous espérons obtenir de faibles valeurs de PER, démontrant l'efficacité du système automatique dans l'identification des phonèmes. Ceci est discuté dans la section 3.1.

Cependant, alors que la séquence de phonèmes est correctement identifiée, nous observons des inexactitudes sur les emplacements des frontières des phonèmes, particulièrement à la fin de l'extrait (pour le mot « actuels »). Ces inexactitudes peuvent poser problème dans l'utilisation de cette segmentation automatique pour d'autres analyses, telles que l'analyse prosodique ou l'analyse de la qualité de la voix sur des phonèmes spécifiques. La qualité de la segmentation est discutée dans la section 3.3.

3.1 Performances de reconnaissance des phonèmes

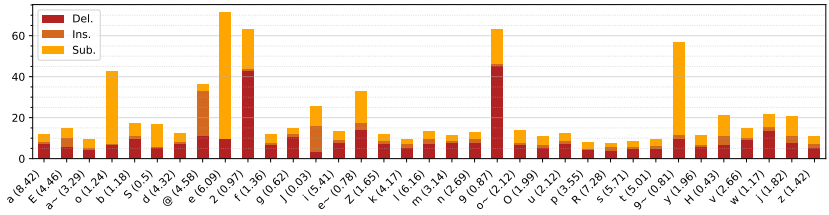

FIGURE 2 – Performances du système automatique de reconnaissance phonétique pour chaque phonème. Les valeurs entre parenthèses indiquent le ratio du nombre d'occurrences du phonème par rapport au nombre total de phonèmes

Les performances de notre système sur chaque phonème sont rapportées dans la Figure 2. Les erreurs les plus communes sont les substitutions, sauf pour /2/ et /9/ qui sont principalement supprimés. En particulier, la voyelle /e/ a un taux de substitution élevé. En effet, dans la parole continue, /e/ est interchangeable avec /E/ par les locuteurs natifs du français. Puisque le système ne permet pas de choix libre entre les deux phonèmes dans le dictionnaire, nous supposons qu'il tend à étiqueter /e/ lorsqu'il rencontre un son similaire à /e, E/. D'autre part, le taux d'insertion élevé pour le schwa /@/ est fortement lié au fait que le système n'est pas informé du caractère facultatif de la voyelle en français.

Lorsque l'on considère les catégories phonétiques (Figure 3), nous observons une réduction drastique du nombre de substitutions, montrant que la plupart de ces erreurs sont faites sur des phonèmes appartenant à la même catégorie. De plus, une observation intéressants peut être faite sur les taux d'erreur observés pour les consonnes : plus la consonne est sonore, plus il est difficile pour le système de l'identifier.

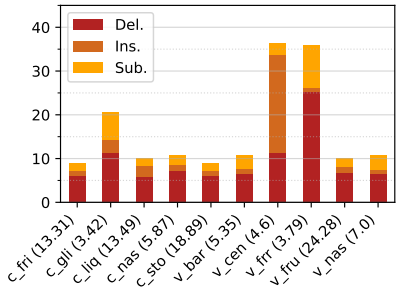

FIGURE 3 – Performances du système sur les catégories de phonèmes. Les valeurs entre parenthèses indiquent le ratio du nombre d'occurrences du phonème appartenant à la catégorie par rapport au nombre total de phonèmes

3.2 Performances de reconnaissance des phonèmes en fonction du style

Les performances de reconnaissance des phonèmes en termes de phonèmes et de catégories de phonèmes en fonction de chaque sous-corpus sont rapportées dans les Tableaux 2 et 3. Pour les deux

unités de mesure, le taux d'erreur diminue avec le degré de préparation de la parole spontanée.

Cependant, puisque la plupart des fichiers du sous-corpus planifié proviennent de la source "Télédiffusion" de la base de données Rhapsodie, nous nous attendions à ce que le système de transcription automatique fonctionne mieux sur ces données que sur les autres sous-corpus, car il est entraîné sur des échantillons du même type (bien que d'une période différente). Ce n'est pas le cas, en partie à cause d'un échantillon de nature plus spontanée que la parole planifiée (D2002, discussion sur un livre) résultant en de mauvaises performances (PER=24,9%). Néanmoins, le système obtient d'assez bonnes performances sur les autres sous-corpus semi-spontanés. Alors que le taux d'erreur augmente avec le degré de spontanéité, les taux d'insertion restent relativement constants, tandis que les taux de substitution augmentent légèrement. Le type d'erreur qui augmente le plus est le taux de délétion (de 3% sur la parole planifiée à 8,5% sur la parole spontanée), atteignant un maximum de 24,9% sur le fichier D2004 de la source Lacheret (locuteur avec un fort accent régional).

Concernant le fait que même pour la parole spontanée, plus de 80% des phonèmes sont détectés correctement (87% pour les catégories phonétiques) et que les erreurs sont principalement dues à des suppressions, nous pourrions penser que la détection automatique des phonèmes peut être adaptée pour l'analyse phonétique de la parole spontanée.

style	# fichiers	# phonèmes	Corr	Sub	Sup	Ins	Err
Tout	54	96756	83,9	9,4	6,7	3,4	19,5
planifié	11	30549	89,2	7,6	3,2	3,0	13,8
semi-spt	4	15302	82,8	9,3	7,9	3,3	20,5
spontané	39	50905	81,0	10,5	8,5	3,6	22,6

TABLE 2 – Erreurs pour la détection des tokens phonétiques pour différents styles de parole

style	# fichiers	# phonèmes	Corr	Sub	Sup	Ins	Err
Tout	54	96756	89,7	3,5	6,7	3,4	13,6
planifié	11	30549	94,9	1,9	3,2	3,0	8,1
semi-spt	4	15302	88,7	3,5	7,8	3.3	14,6
spontané	39	50905	87,0	4,5	8,6	3,6	16,7

TABLE 3 – Erreurs pour la détection des catégories phonétiques pour différents styles de parole

3.3 Performances de segmentation de phonèmes

Le Tableau 4 rapporte les performances de la segmentation des phonèmes selon les métriques détaillées dans la section 2.4.

Sur un total de 7912 secondes à détecter, seules 4746 s. (R=60,0%) sont correctement détectées au niveau du phonème, avec une précision de P=68,2%, conduisant à un score F de 0.62. Cette valeur atteint 5232 s. lorsque l'on considère les catégories phonétiques (66,1%), avec une précision supérieure P = 71,0%, conduisant à un score F correspondant de 0.68.

Concernant l'effet du degré de planification, de manière similaire à ce que nous avons observé dans la section 3.1, ajouter plus de spontanéité dégrade les résultats, à la fois pour les phonèmes (d'un score F1 de 0,67 pour le discours planifié à 0,58 pour le discours spontané) et pour les catégories phonétiques (de F=0,73 à F=0,68).

De plus, notre système fonctionne de manière inégale selon les classes phonétiques : alors qu'il segmente avec de bonnes performances les voyelles nasales (durée cible=823 s., R=71%, P=81%, F=0,76) et les consonnes fricatives (durée cible=1117 s., R=71%, P=78%, F=0,74), il a du mal à estimer les frontières des voyelles centrales (durée cible=330 s., R=60%, P=47%, F=0,53), des voyelles antérieures arrondies (durée cible=595 s., R=30%, P=72%, F=0,42) et des consonnes glissantes (durée cible=170 s., R=57%, P=48%, F=0,52). Les autres catégories phonétiques sont segmentées avec des scores F entre 0,60 et 0,68.

		Phonèmes			Catégories de phonèmes		
	cible	%R	%P	F	%R	%P	F
planifié	2596s	66,5	68,2	0,67	72,3	74,1	0,73
semi-spt	1198s	59,0	64,6	0,62	65,2	71,3	0,68
spontané	60,8	55,5	61,5	0,58	62,5	68,8	0,65
Tous	7912s	60,0	64,4	0,62	66,1	71,0	0,68

TABLE 4 – Évaluation de la segmentation pour la détection des tokens phonétiques pour différents styles de parole

4 Conclusion

Cet article évalue les performances d'un système de reconnaissance automatique des phonèmes pour le français spontané, non seulement en termes de détection de phonèmes mais aussi sur l'identification correcte de leurs frontières. Basé sur le corpus Rhapsodie, qui contient de la parole spontanée de plusieurs sources avec trois degrés de spontanéité, nous avons calculé des métriques d'identification et de segmentation tant au niveau phonémique qu'en fonction de dix catégories phonétiques standards (5 types de voyelles et 5 types de consonnes). Nous avons montré qu'un système de transcription automatique pouvait en même temps obtenir des performances satisfaisantes en identification des phonèmes (PER global de 19,5%, taux d'erreur de 13,6% sur les catégories) et des performances de segmentation insatisfaisantes (F-score de 0,62 et 0,68 pour les phonèmes et les catégories respectivement). Cependant, dans les deux évaluations, de nombreuses disparités ont été observées en fonction du type de parole et des phonèmes considérés.

La prise en compte des catégories de phonèmes améliore les performances dans les deux évaluations, suggérant que les substitutions sont effectuées dans le même groupe phonétique. De plus, toutes les métriques de performance augmentent avec le degré de planification de la parole spontanée considérée. Comme la plupart des systèmes de reconnaissance phonétique sont uniquement évalués sur la parole lue, ces résultats incitent à être très prudents lors de l'utilisation de tels systèmes pour l'analyse linguistique ou prosodique de la parole spontanée.

Remerciements

Cette recherche est financée par l'Agence Nationale de la Recherche (ANR) dans le cadre de l'axe Autonom-Health du PEPR Santé Numérique, convention de subvention n°ANR-22-PESN-0009. VPM a reçu le soutien financier du programme de recherche et d'innovation européen Horizon Europe à travers le projet Marie Skłodowska-Curie MATER (No. 101106577). CB a reçu le soutien financier de la MITI du CNRS (projet PRIME 80 DSM-HEALTH).

Références

ALHARBI S., ALRAZGAN M., ALRASHED A., ALNOMASI T., ALMOJEL R., ALHARBI R., ALHARBI S., ALTURKI S., ALSHEHRI F. & ALMOJIL M. (2021). Automatic speech recognition : Systematic literature review. *IEEE Access*, **9**, 131858–131876. DOI : 10.1109/ACCESS.2021.3112535.

AVANZI M. (2013). *L'interface prosodie/syntaxe en français*.

AVANZI M. & SIMON A. C. (2010). C-PROM : An Annotated Corpus for French Prominence Study. *Speech Prosody*.

BAKER J. K. (1975). *Stochastic modeling as a means of automatic speech recognition.* Carnegie Mellon University.

BEAUMARD C., MARTIN V. P., WU Y., ROUAS J.-L. & PHILIP P. (2023). Automatic detection of schwa in French hypersomniac patients. In *Journée Santé et Intelligence Artificielle (Evènement affilié à PFIA 2023)*.

BOERSMA P. (2001). Praat, a system for doing phonetics by computer. *Glot. Int.*, **5**(9), 341–345.

BOYER F. (2021). *Reconnaissance de Parole Pour Le Français et Intégration Dans Un Système de Compréhension Du Langage Parlé*. Thèse de doctorat, Université de Bordeaux.

BRANCA-ROSOFF S. & LEFEUVRE F. (2016). Le CFPP2000 : constitution, outils et analyses. Le cas des interrogatives indirectes. *Corpus*, (15). DOI : 10.4000/corpus.3043.

DURAND J., LAKS B. & LYCHE C. (2009). Phonologie, variation et accents du français. chapitre Le projet PFC : une source de données primaires structurées, p. 19–61. Hermès.

ESHKOL-TARAVELLA I., BAUDE O., MAUREL D., HRIBA L., DUGUA C. & TELLIER I. (2011). Un grand corpus oral disponible : le Corpus d'Orléans 1968-2012 [A Large available oral corpus : Orleans corpus 1968-2012]. *Traitement Automatique des Langues*, **52**(3), 17–46.

GALLIANO S., GRAVIER G. & CHAUBARD L. (2009). The ester 2 evaluation campaign for the rich transcription of French radio broadcasts. In *Interspeech 2009*, p. 2583–2586. DOI : 10.21437/Interspeech.2009-680.

GELIN L., PELLEGRINI T., PINQUIER J. & DANIEL M. (2021). Simulating Reading Mistakes for Child Speech Transformer-Based Phone Recognition. In *Proc. Interspeech 2021*, p. 3860–3864. DOI : 10.21437/Interspeech.2021-2202.

GOLDMAN J.-P. (2011). Easyalign : an automatic phonetic alignment tool under praat. In *Twelfth Annual Conference of the International Speech Communication Association*.

GUPTA V., KENNY P., OUELLET P. & STAFYLAKIS T. (2014). I-vector-based speaker adaptation of deep neural networks for French broadcast audio transcription. In *ICASSP*, p. 6334–6338. DOI : 10.1109/ICASSP.2014.6854823.

HEBA A. (2021). *Reconnaissance Automatique de La Parole à Large Vocabulaire : Des Approches Hybrides Aux Approches End-to-End*. Theses, Université toulouse 3 Paul Sabatier.

HUCKVALE M., LIU Z. & BUCIULEAC C. (2023). Automated voice pathology discrimination from audio recordings benefits from phonetic analysis of continuous speech. *Biomedical Signal Processing and Control*, **86**, 105201. DOI : https ://doi.org/10.1016/j.bspc.2023.105201.

JELINEK F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, **64**(4), 532–556. DOI : 10.1109/PROC.1976.10159.

KIM H. & CHOI H.-S. (2023). Towards trustworthy phoneme boundary detection with autoregressive model and improved evaluation metric. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 1–5. DOI : 10.1109/ICASSP49357.2023.10096748.

LACHERET A., KAHANE S., BELIAO J., DISTER A., GERDES K., GOLDMAN J.-P., OBIN N., PIETRANDREA P. & TCHOBANOV A. (2014). Rhapsodie : a prosodic-syntactic treebank for spoken French. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, H. LOFTSSON, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, p. 295–301, Reykjavik, Iceland : European Language Resources Association (ELRA).

LACHERET-DUJOUR A. (2003). *La prosodie des circonstants en français parlé*. Collection linguistique (Paris). Paris : Peeters.

LACHERET-DUJOUR A., KAHANE S. & PIETRANDREA P. (2019). *Rhapsodie : A Prosodic and Syntactic Treebank for Spoken French*. John Benjamins.

MERTENS P. (1987). *L'intonation Du Français : De La Description Linguistique à La Reconnaissance Automatique*. Thèse de doctorat.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y., SCHWARZ P., SILOVSKY J., STEMMER G. & VESELY K. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Hilton Waikoloa Village, Big Island, Hawaii, US : IEEE Signal Processing Society.

STRGAR L. & HARWATH D. (2023). Phoneme segmentation using self-supervised speech models. In *IEEE Spoken Language Technology Workshop (SLT)*, p. 1067–1073. DOI : 10.1109/SLT54892.2023.10022827.