Pronunciation and disfluency modeling for expressive speech synthesis

Raheel Qader

N° d'ordre : 00000 ANNÉE 2017

UNIVERSITÉ DE RENNES 1

UNIVERSITE BRETAGNE LOIRE

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Bretagne Loire

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Informatique

École doctorale Matisse

présentée par

Raheel QADER

préparée à l'unité de recherche IRISA – UMR6074
Institut de Recherche en Informatique et Système Aléatoires
École Nationale Supérieure des Sciences Appliquées et de Technologie

---

Pronunciation and disfluency modeling for expressive speech synthesis

Soutenance de thèse envisagée à Lannion le 31 mars 2017

devant le jury composé de :

Simon KING
Full professor at the university of Edinburgh / *Rapporteur*

Spyros RAPTIS
Research director at the Institute for Language and Speech Processing / *Rapporteur*

Yannick ESTÈVE
Professeur des universités à l'Université du Maine / *Examinateur*

Benoît FAVRE
Maître de conférences à Aix-Marseille Université / *Examinateur*

Pascale SÉBILLOT
Professeur des universités à l'INSA Rennes / *Directrice de thèse*

Gwénolé LECORVÉ
Maître de conférences à l'Université de Rennes 1 / *Co-directeur de thèse*

Damien LOLIVE
Maître de conférences à l'Université de Rennes 1 / *Co-directeur de thèse*

Résumé long en français

Le présent résumé est une version condensée en français des éléments présentés en langue anglaise dans cette thèse. L'ordre des sections de ce résumé respecte l'ordre des chapitres de la thèse.

Introduction

Les premiers systèmes de synthèse de la parole ont été utilisés pour aider les personnes malvoyantes, le système lisant le texte de livres. La plupart d'entre eux n'avaient que des fonctionnalités limitées et ne pouvaient produire que des discours robotiques de très faible qualité. De tels systèmes ont cependant été rapidement adoptés par les personnes ayant une déficience visuelle et une augmentation des efforts visant à améliorer la qualité des systèmes s'en est suivie. De remarquables progrès ont ainsi été observés, notamment en raison d'avancées dans le domaine de l'apprentissage automatique, conduisant jusqu'aux systèmes actuels capables de produire une parole intelligible et relativement naturelle. En conséquence, la parole humaine a été remplacée par des discours synthétisés dans diverses applications comme les serveurs de centre d'appel, la lecture d'actualités et la navigation par GPS. Aujourd'hui, le besoin se porte désormais sur l'augmentation de la variabilité et de l'expressivité dans la parole synthétique afin de permettre son emploi dans des contextes interactifs plus ambitieux (lectures de livres, jeux-vidéos, doublage de films...).

L'expressivité est néanmoins un concept complexe. Elle peut être définie comme un indicateur vocal de diverses caractéristiques psychologiques d'un locuteur comme son état émotionnel, son style de parole, sa personnalité et son intention. Par exemple, la parole spontanée est un style de parole très expressif dans lequel les orateurs n'ont pas préparé leur discours auparavant et où la conversation évolue naturellement. En raison de cette complexité, le traitement de ce problème en synthèse de la parole est une tâche difficile. Dans cette lignée, le but principal de ce travail est d'intégrer de l'expressivité dans la synthèse de la parole. Précisément, nous nous intéressons à la parole spontanée et nous concentrons sur deux aspects principaux encore peu étudiés et qui ont un impact

significatif sur l'expressivité : les variantes de prononciation et les disfluences.

Ce manuscrit présente mes deux contributions principales sur ces aspects. La première contribution est une nouvelle méthode d'adaptation de la prononciation qui permet de produire des variantes de prononciation propres à un style. Nous proposons d'effectuer cette adaptation en apprenant automatiquement les variations phonémiques de la parole spontanée à partir d'un corpus de parole conversationnelle. Cet apprentissage s'appuie sur un cadre probabiliste à travers l'emploi de champs aléatoires conditionnels et de modèles de langage. Cette méthode a été validée par des évaluations objectives ainsi que des tests d'écoute. La deuxième contribution de ma thèse est la production automatique de disfluences pour des énoncés n'en contenant originellement pas. Ce travail s'appuie sur le même cadre statistique que nos travaux sur la prononciation. Par ailleurs, comme les travaux sur la production de disfluences sont encore rares, il peut être considéré comme exploratoire. En dehors de la méthode de production proposée, ce travail contribue ainsi également à des aspects techniques du problème comme la préparation des données et l'évaluation des résultats.

Dans la suite de ce résumé, nous listons les éléments, méthodes, résultats introduits et discutés au cours des différents chapitres de cette thèse. Nous commençons par une présentation du domaine, puis un état de l'art. S'ensuivent les détails de chacune de mes contributions et, enfin, un bilan de ce travail de thèse.

Chapitre 1 : Parole et expressivité

Le premier chapitre de la thèse est consacré à l'explication des bases de la parole et de l'expressivité. Nous abordons d'abord les bases du mécanisme de production de la parole humaine, par opposition à la parole synthétique, et les différentes couches d'abstraction du langage, puis nous étudions trois éléments importants de la parole et du langage : la phonétique, la phonologie et la prosodie. Pour chaque aspect, nous mentionnons les éléments les plus importants et comment ils peuvent être liés au problème abordé dans cette thèse. Ensuite, à travers la notion d'expressivité, nous discutons des concepts d'émotions, de styles de parole et d'accents. Enfin, les effets de l'expressivité sur la prononciation et la fluidité du discours oral sont présentés, particulièrement les variations phonémiques et les disfluences qui se produisent en parole spontanée.

Chapitre 2 : Prononciations et disfluences en synthèse de la parole

Dans ce chapitre, l'objectif principal est de décrire les différentes façons d'exploiter l'expressivité pour rendre la parole synthétique plus humaine. Bien que la prise en

compte de l'expressivité soit une problématique à la fois en synthèse et reconnaissance de la parole, les études portant sur la reconnaissance sont plus nombreuses dans la littérature. Ainsi, nous abordons des travaux des deux domaines mais mettons l'accent sur la synthèse de la parole. Le chapitre explique le fonctionnement général d'un système de synthèse de la parole et détaille plus précisément les approches actuellement dominantes pour la réalisation de leur moteur. Les différentes techniques d'apprentissage automatique impliquées dans ces approches sont également traitées. Ensuite, nous décrivons l'état de l'art en modélisation de la prononciation, notamment la conversion graphème-phonème et les traitements dits post-lexicaux, puis nous donnons un panorama des travaux visant la modélisation des disfluences et approfondissons ceux dédiés à leurs prédiction et production en synthèse de la parole.

Chapitre 3 : Données et méthodologie d'évaluation

Dans ce chapitre, nous décrivons les données et la méthodologie d'évaluation qui sont utilisées dans le reste de cette thèse pour produire des variantes de prononciation et des disfluences. Tout d'abord, nous présentons le corpus Buckeye de parole conversationnelle en anglais, corpus qui est la principale source de données pour les deux tâches. Une analyse statistique du corpus en est notamment donnée afin de caractériser empiriquement les phénomènes étudiés. Ensuite, nous dressons la liste des caractéristiques automatiquement extraites du corpus qui sont considérées dans nos travaux. Ces caractéristiques sont d'ordres linguistique, articulatoire et acoustico-prosodique. Enfin, nous traitons des différentes méthodologies d'évaluation objectives et subjectives utilisées pour mesurer l'efficacité de nos propositions.

Chapitre 4 : Production de variantes de prononciation

Les deux principaux objectifs de cette thèse sont de générer des variantes de prononciation et des disfluences de la parole dans le cadre de la synthèse de la parole. Dans ce chapitre, nous présentons nos contributions sur le côté des variantes de prononciation qui jouent un rôle critique pour rendre la parole synthétique plus expressive. Notre but est de fournir une méthode qui est capable d'apprendre et prédire automatiquement de telles variantes. Pour cela, nous proposons une méthode qui permet d'adapter des prononciations dites canoniques, c'est-à-dire telle que données par un dictionnaire, vers un style présentant intrinsèquement beaucoup de variabilité, en l'occurrence la parole spontanée. Cette approche s'appuie sur des champs aléatoires conditionnels effectuant une conversion phonème-phonème et un réordonnancement des hypothèses de prononciation ainsi produites par un modèle phonologique.

Pour le développement de cette méthode, nous considérons plusieurs aspects importants. Notamment, nous étudions la sélection des attributs utiles pour notre tâche et la meilleure conduite à tenir pour l'apprentissage automatique des champs aléatoires conditionnels. Nous déterminons aussi expérimentalement si l'adaptation des prononciations est suffisante pour produire, à elle seule, des signaux de parole jugées comme expressifs, c'est-à-dire sans recourir de surcroît à une adaptation prosodique. Enfin, nous discutons de la manière d'évaluer correctement les résultats. À l'issue de ces investigations, nous montrons que les prononciations spontanées adaptées utilisant une combinaison des caractéristiques linguistiques et prosodiques reflètent effectivement le style spontané, notamment en comparaison des prononciations canoniques initiales. Les résultats des tests d'écoute suggèrent même que les échantillons de parole synthétisés à l'aide de prononciations adaptées sont perçus comme plus intelligibles que ceux qui utilisent des prononciations réalisées par des locuteurs réels. De plus, il a été vérifié que les caractéristiques linguistiques seules fonctionnent bien pour la tâche d'adaptation de la prononciation, sans consignes prosodiques particulières à respecter.

Dans ce chapitre, nous montrons également que la méthode proposée peut être étendue à d'autres tâches similaires d'adaptation. Précisément, nous montrons que la méthode peut être utilisée pour résoudre le problème d'incohérence entre les séquences de phonèmes générées par les convertisseurs graphèmes-phonèmes pendant la synthèse et celles provenant du corpus de parole utilisé par le moteur de synthèse. Cette incohérence conduit généralement à des signaux de parole de mauvaise qualité. Ce travail est réalisé sur un corpus de parole en français. Nous démontrons que la méthode proposée apporte une amélioration en termes de taux d'erreurs sur les phonèmes. Les tests perceptifs ont également montré une amélioration de la qualité de la synthèse vocale lorsque la méthode d'adaptation est incluse dans le processus de phonétisation.

Chapitre 5 : Production de disfluences

Les disfluences sont un autre facteur d'expressivité. Plus généralement, elles apportent de la richesse au langage et à la communication, par exemple en facilitant la compréhension d'un discours par un auditeur, en aidant à la bonne gestion des tours de parole entre interlocuteurs ou en créant une atmosphère amicale. Le problème est que, généralement, l'entrée d'un système de synthèse est un texte avec un style écrit sans aucune sorte de disfluences. Donc, la question principale ici est de savoir comment rendre disfluent le texte écrit, c'est-à-dire où et comment y insérer des disfluences. De plus, selon le style et le contexte de la parole, le degré de disfluence requis peut varier. Par exemple, la parole d'orateurs stressés est *a priori* plus disfluente que celles d'orateurs détendus. Donc, une deuxième question est de savoir comment contrôler le nombre de disfluences insérées.

Dans ce chapitre, plus exploratoire, je propose une nouvelle méthode de génération de disfluences qui est capable d'insérer plusieurs types d'entre elles et de contrôler leurs proportions respectives. Pour ce faire, nous formalisons d'abord le problème comme un processus théorique où le texte initial est transformé itérativement jusqu'à ce que nous atteignions le niveau souhaité de disfluence. Plus précisément, le processus est décomposé en un problème d'étiquetage visant à identifier les portions de texte à éditer et une tâche de génération de langage naturel pour insérer les mots disfluents. Il s'agit d'une nouvelle contribution puisque la plupart des travaux précédents se concentrent sur la génération d'un unique type de disfluence (les pauses), alors que notre méthode est suffisamment générique pour en modéliser et générer plusieurs, à savoir des pauses, des répétitions et des révisions. Nous étudions quelles caractéristiques linguistiques sont utiles à la production des disfluences et comment contrôler le degré de disfluences. Le résultat de ce travail est une preuve de concept sous la forme d'une implémentation du processus fondée sur des champs aléatoires conditionnels et des modèles de langage. Nos expériences ont montré la viabilité de cette implémentation et ouvert des pistes de réflexion pour franchir de nouveaux jalons en terme de qualité des énoncés disfluents produits.

Conclusion et perspectives

Dans cette thèse, nous avons abordé la question de l'expressivité dans la synthèse de la parole. Puisque l'expressivité couvre un vaste domaine, l'accent a été mis sur le discours spontané et sur l'étude des variantes de prononciation et des disfluences. J'ai notamment proposé une nouvelle méthode pour produire des variantes de prononciation en adaptant les séquences de phonèmes canoniques pour imiter le style spontané. J'ai étudié différents facteurs qui influent sur l'efficacité de cette adaptation et montré que la méthode proposée peut être étendue à d'autres tâches d'adaptation. Les évaluations objectives et subjectives montrent de bons résultats dans ces diverses situations. J'ai également contribué a l'état de l'art en proposant une approche exploratoire mais novatrice permettant la production automatique de disfluences. Cette approche s'appuie sur un processus formel du mécanisme de production et une traduction de ce processus en un algorithme et une implémentation expérimentale. Cette implémentation a permis de montrer la viabilité de l'approche proposée.

Plusieurs perspectives sont ouvertes par cette thèse. Dans l'ensemble, la plus directe d'entre elles est la possibilité de combiner les résultats des deux contributions en produisant des disfluences pour un énoncé donné, puis en passant le résultat à travers notre processus d'adaptation de la prononciation. De cette façon, nous pourrions générer une parole synthétique encore plus expressive. Ensuite, les discussions des résultats mon-

trent le besoin de méthodes d'évaluation plus robustes. Dans le travail sur les variantes de prononciation, il serait notamment bon de savoir pondérer l'importance de certaines transformations afin de mieux discriminer les différences entre séquences de phonèmes, que ce soit pour calculer des distances d'édition ou pour analyser des écarts à l'issue de tests d'écoute. Quant aux disfluences, il reste encore à déterminer quelle est la meilleure question à poser aux participants des tests pour produire des résultats discriminants. Une dernière perspective concerne la caractérisation même de la notion d'expressivité. Nous en avons étudié des implications sur les versants phonologique et linguistique du langage. Cependant, la prosodie est aussi un élément critique de la parole expressive. Ainsi, pour obtenir des discours pleinement expressifs, celle-ci devrait également être prise en compte. Encore au-delà, pour aller vers une résolution plus complète et une compréhension approfondie de l'expressivité, il faudrait probablement étudier conjointement les incidences de ces différents facteurs et leurs éventuelles interdépendances.

Contents

Introduction

Interest in building systems that simulate the way humans understand and generate speech has increased immensely in the past decades. Speech technology based systems usually consist of a speech recognizer for speech input, a speech synthesizer for speech output and a language understanding component that serves as a link between the two. Each of these three fields has been studied extensively in the past and an acceptable level of quality has been reached for each of them. However, speech systems are still far from being perfect. In this thesis we concentrate on some ways of improving Text-to-Speech (TTS) systems by making them more expressive.

The early TTS systems were used in aiding visually impaired people, where the system reads some text from a book and converts it into speech [Taylor, 2009]. Most of these early systems had only limited functionalities and could only produce very low quality and robotic speech. Such TTS systems were however quickly adopted by the visually impaired people as they were an easier option than having someone to read a document for them. This adoption led to an increase in the efforts to further improve the quality of TTS systems. Progress in TTS systems has been remarkable in the recent years, mostly due to the emergence of new technologies in the field of speech synthesis and natural language processing. As a result, better TTS systems with more natural and intelligible speech have been developed.

Due to these advancements, in numerous domains and various types of applications where speech plays an important role, human speech has been replaced with synthetic speech including call-center automation, reading news stories, navigation systems and a wide variety of other applications. This has increased the need for more variability and expressiveness in synthetic speech. Thus, most of the recent studies in this domain have been toward making TTS systems more expressive.

The issue here is that expressivity is a complex concept. It can be defined as the vocal indicator of various emotional states [Govind and Prasanna, 2013]. Moreover, emotional states can be extended to psychological characteristics of a speaker such as emotions, speaking style, personality, and intention. For instance, spontaneous speech is a highly expressive speaking style as the speakers have not prepared their speech previously and

the conversation evolves naturally. Due to the complex nature of expressivity, dealing with it in the context of TTS systems has always been a difficult task. Despite this, many different approaches have been proposed to integrate expressivity in TTS systems.

Contribution of the thesis

The main goal of this work is to enable expressivity in speech synthesis. However, as already mentioned, expressivity encompasses a wide area, therefore, we mainly focus on spontaneous speech and we concentrate on two main aspects which are believed to be of significant impact on expressivity: pronunciation variants and speech disfluencies. This PhD manuscript presents my contributions on these aspects, one on each of them. In a more general view, although these contributions are focused towards TTS, some of their elements could be extended to Automatic Speech Recognition (ASR).

The first contribution is a new pronunciation adaptation method. Precisely, this method enables pronunciation variation to improve the performance of TTS systems. It has been particularly applied in the case of spontaneous speech. This contribution is particularly important because, although most of the current systems are able to produce high quality and intelligible speech, they still have a "neutral style" [Pitrelli et al., 2006]. This is mainly because such systems solely rely on standard pronunciations, i.e., extracted or learned from a general dictionary without considering any sort of pronunciation variants. In general, the simplest possible way to introduce pronunciation variants into TTS is to manually add alternative pronunciations in the dictionary [Fukada et al., 1999]. Although this method might work in certain cases, it is definitely not enough to capture all variants; moreover, it requires expert knowledge. Similarly, an expressive or spontaneous speech corpus in TTS can partly introduce some of the variants, however building such a corpus is a costly and time consuming task. A reasonable solution to this problem is thus to adapt standard pronunciations to reflect a specific style, a spontaneous style in our case. In a machine learning perspective, this task corresponds to predicting an adapted sequence of phonemes from an input sequence of canonical phonemes, i.e., standard pronunciations. More precisely, this means predicting if input phonemes should be either deleted, substituted, or simply kept as is, and whether new phonemes should be inserted. In this thesis, we propose to perform pronunciation adaptation by automatically learning phonemic variations of spontaneous speech from a corpus of conversational speech using conditional random fields and language models, and apply them on standard pronunciations to generate alternative ones. The method has been validated by objective and subjective evaluations.

The second contribution of this work is the generation of speech disfluencies for TTS. Similarly to pronunciation variants, speech disfluencies are one of the main char-

acteristics of spontaneous speech. However, understanding disfluencies is not a trivial task since they can be related to various factors such as the psychological state of the speaker and the structure of the speech discourse [Corley and Hartsuiker, 2003, Clark and Fox Tree, 2002]. Moreover, disfluencies impact several aspects of speech such as segment durations, intonation, coarticulation patterns [Shriberg, 1999] and have been found to provide several benefits like faster reaction times and faster word integration [Corley and Hartsuiker, 2003, Fox Tree and Schrock, 2002] (cited by [Dall et al., 2014]). Although the majority of the work in this area has been conducted with the intention of identifying disfluencies for ASR systems [Liu et al., 2006, Honal and Schultz, 2005, Stolcke et al., 1998], integrating disfluencies in TTS systems is also crucial to have more human-like speaking machines. Thus, in this thesis, the strength of our contribution is to propose a disfluency generation approach, which contrary to related work, can generate many different types of disfluencies. Like our work on pronunciation variants, this work uses conditional random fields and language models. Since the works in this area are few, and no clear evaluation metrics have been defined by the community, our work can be considered as exploratory. Hence, apart from the core generation method, this work also contributes to the technical aspect of the problem like data preparation and evaluation.

Outline of the thesis

This thesis is organized as follows: the first two chapters give a general background on speech and expressivity and describe the different techniques to integrate expressivity in TTS systems. The third chapter presents the data and the evaluation metrics that are going to be used for the task of pronunciation variation and disfluency modeling. Finally, the last two chapters present our contributions on the generation of pronunciation variants and speech disfluencies. A more detailed outline is given below.

Chapter 1 The first chapter presents the basics of speech and expressivity. We first start by explaining the human speech production and describe concepts like phonetics, phonology, and prosody. A definition of expressivity and what we consider as expressivity is also given in this chapter. Lastly, the impacts of expressivity on pronunciation and speech fluency are discussed.

Chapter 2 The focus of the second chapter is on describing the different ways of exploiting expressivity to make speech applications more natural and expressive. The chapter explains in detail TTS systems and highlights the different machine learning

techniques used in such systems. It also deals with the problems of integrating pronunciation variants and disfluencies in TTS systems.

Chapter 3 The third chapter presents the data which are used for both pronunciation variant and disfluency generation tasks. Moreover, different objective and subjective evaluation methodologies are explained.

Chapter 4 In the fourth chapter, our contribution on generating pronunciation variants for TTS is presented. Firstly the overall methodology of the proposed approach is presented. Then we explain how the approach can be used to predict variants of spontaneous speech. In addition, we show that the method can be extended to other adaptation tasks, for instance to solve the problem of inconsistency between the phoneme sequences in TTS systems.

Chapter 5 The last chapter of this thesis presents details of the proposed disfluency generation approach. We first, formalize the problem as a theoretical process in which we give details of an iterative disfluency insertion approach. The process of preparing and cleaning the Buckeye corpus is also explained in this chapter. Finally we present several experiments on disfluency generation alongside their objective and subjective evaluation results.

Chapter 1

Speech and expressivity

In order to fully understand the notion of expressivity and what makes speech expressive, a basic understanding of human speech, as opposed to synthetic speech, has to be acquired. In addition, the elements that contribute to expressivity in speech have to be studied as well. Thus, this very first chapter is dedicated to explaining the basics of speech and expressivity. In Section 1.1, we first discuss the human speech production. Then in Section 1.2, we study three important elements of speech and language: phonetics, phonology and prosody. Section 1.3 of this chapter concerns expressivity in speech; concepts like emotions, speaking styles and accents are discussed. Finally in Section 1.4, effects of expressivity on pronunciation and fluency are covered.

1.1 Speech production and structure

Speech is one of the most usual ways that people use to communicate with one another. In addition to conveying a linguistic message, speech also carries other information about emotions, expressions, intention, speaker identity, etc. [Byrnes, 1999]. In order to understand the notion of speech, one has to have a basic knowledge of how speech is produced and of the different types of information it contains. Thus, in the rest of this section, first, the speech production mechanism and then the organization of speech into a multilayer structure is explained.

1.1.1 Speech production mechanism

Speech production is a complicated process involving coordination of several vocal organs [Taylor, 2009]. As illustrated in Figure 1.1, these organs include the lungs, larynx, pharynx, nose and various parts of the mouth—including the tongue—which are collectively known as the vocal tract.

The speech production process is initiated when the air flow, sent from the lungs, passes through the space between the vocal folds, known as *glottis*, and other vocal organs until it exits from the lips. When the vocal folds are stretched, the air flow causes them to vibrate rapidly and creates a periodic sound. The rate of this vibration is known as the fundamental frequency (F0). Sounds created this way are called *voiced* sounds such as vowels /ɑ/, /ɪ/ or some consonants like /v/, /z/. When the vocal folds are relaxed, the air flows through the larynx without any interruption, and with minor modulation by the vocal organs, a non-periodic sound, known as noise is produced. Sounds created this way are called *unvoiced* sounds such as /s/ or /p/.

As soon as the air flow exits the glottis, the properties of the resultant sound is modulated by different vocal organs known as *articulators*. These articulators move in various ways to produce different sounds with different properties [Holmes, 2001]. For example, the blockage and release of air flow using the lips leads to the production of bilabial sounds such as /p/, /b/. Likewise, alveolar sounds like /t/, /d/ are produced when the tongue completely or slightly touches the alveolar ridge. Nasal sounds on the other hand, such as /m/, /n/ are produced when the air stream moves out from the nose instead of the mouth. This model of sound production which uses a source to generate a sound (i.e., the sound produced in the vocal folds) and then shapes or filters it using the articulators is often referred to as the source-filter model.

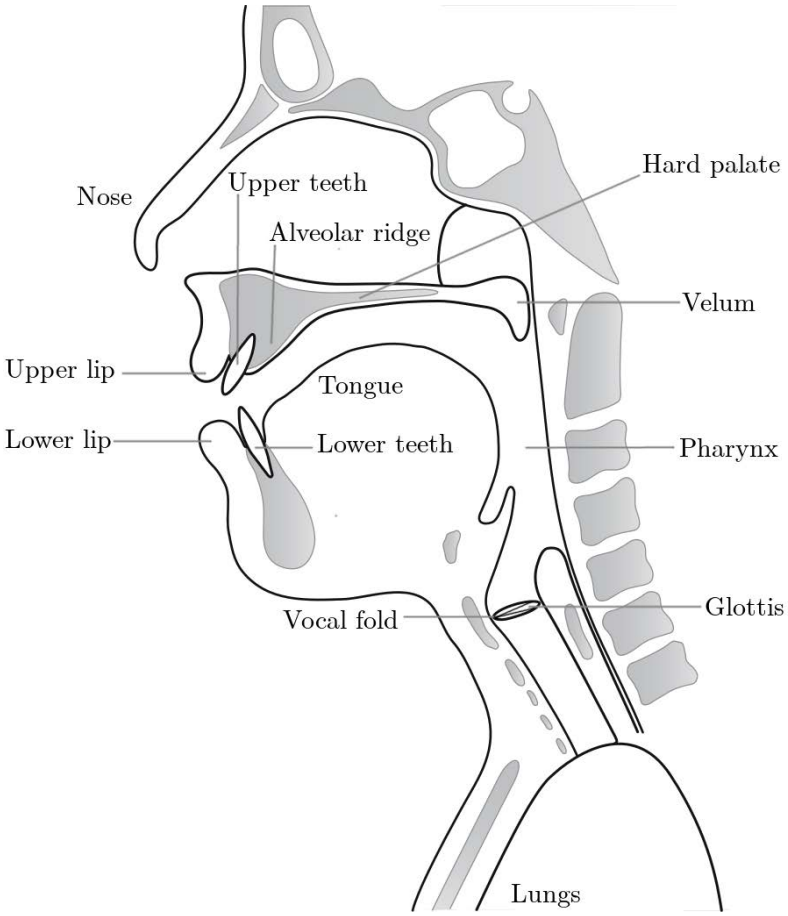


Figure 1.1: A diagram of the vocal organs (articulators) (source: [Benesty et al., 2007]).

1.1.2 Multilayer structure of speech

The main aim of speech is obviously to act as a communication tool in order to share ideas and exchange information. To appreciate how humans use speech to communicate with each other, it is important to know the different parts that form speech.

The first point that has to be made clear is that speech cannot be considered as a single piece of information that is simply made out of a series of distinctive sounds formed by the movement of articulators [Myers et al., 1981]. This is mainly because the comprehension process between a speaker and a listener requires both sides to have a detailed knowledge about the notions on which the speech is built, for instance, the mapping between the words and their corresponding acoustic realizations. Moreover, understanding the semantic side of speech, i.e., to know the real meaning behind the words, and the syntax side, i.e., to know how the flow of words can be correctly arranged in speech, is necessary. A good definition of the speech process which clearly explains these different sides is given in [CASANA, 2013] as follows:

> "The act of speech begins with an intention to communicate. Next, an idea forms, outlining what the speaker wants to say. The words for the desired message are put in correct order, using the correct grammar. Each of the words is comprised of a specific sequence of sounds and syllables that must be ordered together. All of this information is translated from an idea and information into a series of highly coordinated movements of articulatory organs."

As it can be understood from the definition, speech is formed with a complex and iterative process going through several stages. Therefore, it is important to break speech down into some kind of abstract layers in order to be able to study the different types of information it contains. There have been many attempts in the past to represent speech in the form of abstract layers [Comer and Gould, 2010, Marten, 2002]. The following list is often used in the literature to represent speech in the form of layers:

Semantics deals with the meaning of words and sentences as well as the rules in which the meaning of a sentence can be derived from the meaning of its words in a language [Brown and Allan, 2010]. It mainly examines the changes in the meaning of words due to the contextual changes.

Syntax concerns the organization of words in a sentence and the set of rules that organizes words into sentences [Brown and Allan, 2010]. It also studies the principles and processes by which sentences can be constructed in particular languages [Chomsky, 2002].

Phonetics and phonology is the study of speech sounds and their function in a given language [Collins and Mees, 2013]. The generation and classification of speech sounds based on their properties fall under phonetics, while their functions in a language are related to phonology.

Prosody focuses on the rhythmical and tonal features of speech that are layered upon individual phonological segments. It includes stress, pitch, and rythm [Schreiber, 1991]. Prosody plays a critical role in making speech more natural since it carries information about emotions, speaker intention, etc.

Acoustics deals with the physical properties of speech including F0, duration and energy in order to generate the speech waveform.

Each of these layers has a critical role in conveying the meaning behind speech in a communication process. However, what concerns us the most in this thesis are

phonetics, phonology and prosody. These layers, as it has been shown in the literature [Vazirnezhad et al., 2009, Brennan and Schober, 2001, Shriberg, 1999, Greenberg, 1999], have an enormous impact on pronunciation variation and speech disfluencies. As we will see in Chapter 4 and Chapter 5, the information extracted from these three layers can help in generating pronunciation variants and speech disfluencies. In the next section, the key notions about these three layers are provided.

1.2 Phonetics, phonology and prosody

As stated earlier, speech cannot be considered as a mere sequence of sounds produced arbitrarily. There also exists a system that governs all the possible ways these sounds can come together to form meaningful words and sentences. Phonetics and phonology are two branches that deal with the properties of elementary speech sound units and how these units are used in a particular language. The generation and classification of speech sound units based on their properties fall within the branch of *phonetics*, while their functions in a language are related to *phonology*. In addition to phonetics and phonology, *prosody* is also an important aspect of speech, which is more related to larger units of speech such as syllables and plays an important role in making speech natural and expressive.

Important elements of these three domains are covered in this section in order to understand their impact on pronunciation and speech fluency.

1.2.1 Phonemes, phones and allophones

Speech is formulated by combining words into meaningful sentences, each conveying a specific message. Moreover, words are also decomposed into small elementary sounds which are called *phonemes*. A phoneme is a small speech unit that can transform the meaning of words. In other words, the substitution of a phoneme in a word with another, changes the meaning of that word. For instance, substituting the initial phoneme in the word “to” with /d/ will change the word completely as the word becomes “do”.

Additionally, the term *phone* is used to describe the acoustic realization of phonemes. A phone is a single speech sound with unique articulatory properties. One phoneme can be realized in several different ways, each realization being called an *allophone* of that particular phoneme. For instance, the phoneme /p/ has two different realizations in English, one being aspirated [p^h], and the other one unaspirated [p]. In contrast to phonemes, substituting an allophone of a phoneme with another will not result in changing the meaning. Phonemes and allophones are generally written between slashes (/ /) and phones between square brackets ([ ]).

When working with speech sounds it is a common practice to use the International Phonetic Alphabet (IPA) [International Phonetic Association, 1999]. In IPA, each sound is represented with a symbol and all the sounds in any language can be represented using those IPA symbols. Moreover, symbols can be specialized by adding characteristics to them, for instance, to define a phone as stressed or long. This is done with the help of special marks called *diacritics*, indicating a slight change in the sound. For instance, in the French word *bonjour*, the graphemes "on" are represented by the phoneme /õ/ where the tilde is an indication of nasalization. Having such a standard system of sound representation makes working and sharing ideas in this area easier.

Among others, Arpabet is also a widely used phonetic transcription alphabet [Weide, 1998]. Arpabet was developed by the Advanced Research Projects Agency (ARPA) for coding American English phonetic symbols. Arpabet is more machine readable than IPA as every symbol is represented by a sequence of ASCII characters. For instance, the corresponding IPA symbol of /tʃ/ is /ch/ in Arpabet.

1.2.2 Vowels and consonants

Phonemes can be classified into two main categories: vowels and consonants. One major factor that distinguishes vowels from consonants is that they are produced when the flow of air is mostly unconstrained (except for the vocal folds) and with an open mouth, whereas during consonants, there is usually a constriction to air flow somewhere in the vocal tract (e.g., lips, teeth or tongue) [Holmes, 2001]. In certain consonants such as /p/ and /b/ the flow of air is completely blocked by the lips. While consonants can be either voiced or unvoiced, vowel sounds are always voiced. In English and some other languages two vowels can be joined as a result of a glide leading to a slight change in the sound. Such vowels are called diphthongs, like the vowel in "toy" /tɔɪ/ [Ashby and Maidment, 2005]. Although diphthongs are formed from two vowels, they are still considered as single phonemes. In contrast, single vowels are called monophthongs such as the vowel in "teeth" /tiːθ/ [Ashby and Maidment, 2005]. Additionally, some voiced consonants can become similar to vowels. Such consonants are referred to as semi-vowels [Goldberg and Riek, 2000]. Figure 1.2 shows the vocalic triangle where vowels are classified based on the position of the tongue (x-axis) and the opening of the mouth (y-axis), while Figure 1.3 presents the list of consonants classified based on voicing, place, and manner of articulation.

Several studies have been conducted in order to determine the role of vowels and consonants in pronunciation variation. In [Jurafsky et al., 2001], the authors analyzed variations on vowels in three datasets and reported that between 6.3% and 10% of the vowels were affected by a phenomenon called vowel reduction (described in Sec-

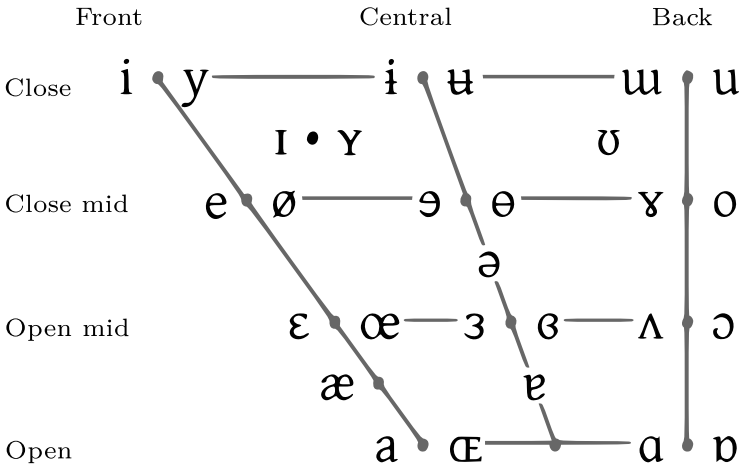


Figure 1.2: Vocalic triangle, where symbols appear in pairs, the one to the right represents a rounded vowel (in which the lips are rounded) (source [International Phonetic Association, 1999]).

tion 1.4.1). Further, a study on Spanish learners of English showed that substitution are very likely for consonants that do not exist in Spanish such as /z/ and /ð/ [You et al., 2005]. Moreover, a significant number of pronunciation variations were observed for vowels, mostly because vowels like /ʌ/ and /ɪ/ have a great tendency to vary. Finally, a study about speech recognition accuracy on Spanish, Italian, and English speakers found out that English speakers had problems in correctly pronouncing some Italian words containing certain diphthongs [Strik et al., 1998].

1.2.3 Syllables

Syllables are considered to be an intermediate unit, sometimes thought to interpose between the phones and the words [Huang et al., 2001]. Contrary to what most people think, syllables are not just mere sequences of chained phonemes but they are completely distinguishable from phonemes in the sense that they have a systematic structure and are tightly connected to the higher tiers of speech such as prosody [Greenberg, 1999]. Syllables are structured into 3 parts which are, from left to right, the onset, the nucleus and the coda. The conjunction of the nucleus and the coda forms the rhyme. The nucleus is mandatory and usually consists of a vowel, while the other two parts are optional and made of consonants and semi-vowels. For instance, the word "kitten" is canonically pronounced as /ki.tən/, where the dot is used to separate syllables. This pronunciation is made of the 2 syllables /ki/ and /tən/, and the second syllable has the following structure: the onset is /t/, the nucleus is /ə/ and the coda is /n/, whereas in the word "my" /maɪ/, there is only one syllable which has the onset /m/, the diphthong nucleus /aɪ/, and no coda.

Syllables can be categorized into *open* and *closed* syllables [Moats, 2004]. Open syllables end with a vowel, like in /maɪ/, while closed syllables end with a consonant as in /tən/. In some languages, for example English, syllables have lexical stress. Stressed syllables are those in which vowels have to be articulated louder or longer or with a higher pitch. In words with more than one stressed syllable, the strongest stress is referred to as *primary stress*, and *secondary stress* for the others [Skandera and Burleigh, 2011].

Several studies have examined the effects of syllables on pronunciation and fluency. For example, stressed syllables have been shown to be less likely to be deleted during spontaneous speech [Dilts, 2013], whereas a study on syllable deletions in the Switchboard corpus showed that in certain cases, some syllables might be completely deleted [You et al., 2005]. For instance, the word "variety" which has four syllables in its canonical form /və.raɪ.ə.ti/; can be reduced to up to two syllables when realized /vraɪ . ti/. In [Vazirnezhad et al., 2009], syllables were analyzed based on their position inside a word, and it was reported that the initial and middle syllables had very low ratios of deletion in spontaneous speech, whereas this ratio was higher in the final syllable. Lastly, the author in [Shriberg, 1999] reported lengthening in syllables when immediately followed by a disfluency.

1.2.4 Coarticulation

Coarticulation refers to a situation where a phonological segment is influenced by the neighboring segments [Hardcastle and Hewlett, 2006]. The consequence is that

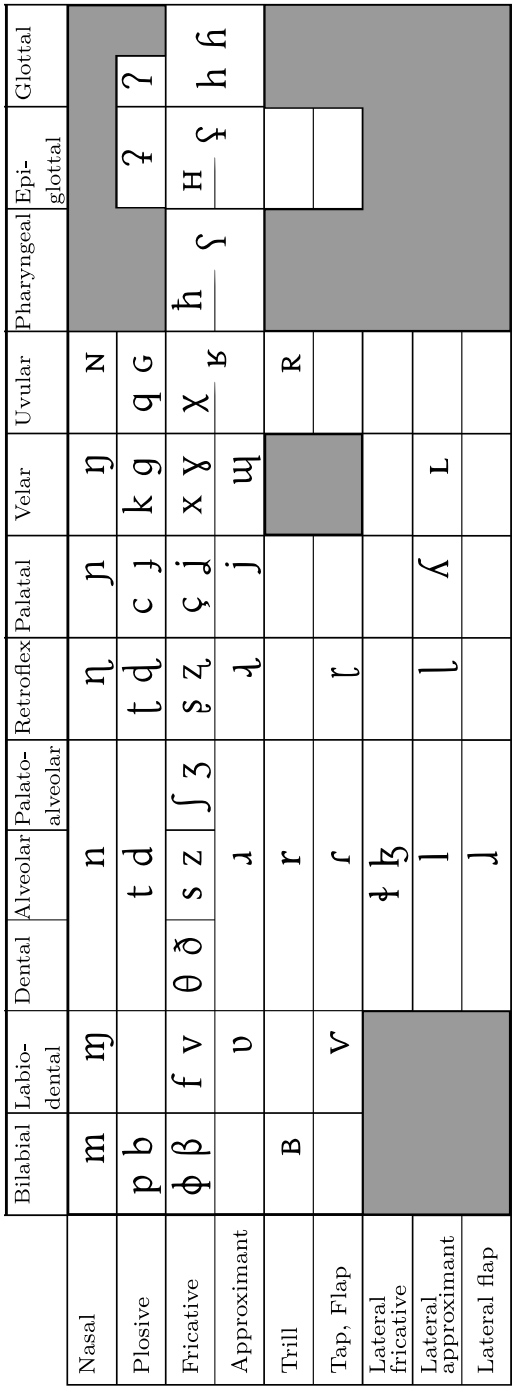


Figure 1.3: IPA consonant chart, where symbols appear in pairs, the one to the right represents a voiced consonant (source: [International Phonetic Association, 1999]).

phonemes are not always realized identically in all environments and can lead to complex acoustic patterns [Taylor, 2009]. The reason of such phenomena is because the articulators are moving constantly and very rapidly, and as they reach a position required to realize a specific phoneme, they have to rapidly move to the next position for realizing the next one. Thus the realization of a specific phoneme is heavily impacted by the neighboring phonemes. As an example, the position of the tongue during the articulation of the consonant /k/ will be placed further forward on the palate (*cf.* Figure 1.1) before a front vowel as in the word ([kiː] "key"), and further backward on the palate before a back vowel as in the word ([kɔː] "caw"). It is important to realize that coarticulation is mainly a physiological process out of the speakers' control and mostly governed by universal rules rather than language-specific rules.

The topics that were discussed in the last four sections mostly concerned the linguistic aspect of speech in the sense that they are independent from any signal, while the next section is about prosody and its impacts on pronunciation variation and disfluencies.

1.2.5 Prosody

According to Huang et al. [2001], prosody can be defined as "a complex form of physical and phonetic effects that is being employed to express attitude, assumptions and attention as a parallel channel in our daily speech communication". Prosody is mostly associated with syllables rather than smaller units like phonemes. Thus, it is often considered as *suprasegmental* information [Rao, 2012].

Prosody has a crucial role in making speech sound natural and more intelligible by varying acoustic parameters of suprasegmental units. This increases the chances of correctly conveying the underlying message to the listener. For instance, by increasing the loudness of certain units of the speech, the speaker can signal their importance. Prosody is used for many purposes such as expressing emotions, emphasizing words, or indicating the end of sentences [Taylor, 2009]. The main acoustic parameters that characterize prosody are the following:

- *Fundamental frequency (F0)* is the rate of the vocal fold vibration. F0 is referred to as pitch in perceptual terms.
- *Duration* is the time interval required to realize a speech signal.
- *Intensity* refers to the amplitude of the sound signal which is also described as the sound strength and measured in decibels (dB).

According to [Delais-Roussarie et al., 2015], prosody has three main elements: accentuation, intonation, and phrasing. Accentuation is the assignment of prosodic prominence to certain syllables. Prominence is mostly related to local modifications of acoustic parameters such as duration, intensity or F0. It includes all non-phonemic lexical properties such as stress in English or tone in Mandarin. As for intonation, it can be considered as the melody of speech, as determined by the variation of F0 over an utterance. Intonation carries different kinds of information through highlighting important parts of speech. For example, a rising intonation at the end of "this is the Paris train" makes the utterance a question rather than a statement. Lastly, prosodic phrasing is the constituent that identifies different chunks in speech and signals grammatical structure. For instance, a falling intonation most of the time denotes the end of a clause or a sentence. Prosodic phrases are generally ended with a silent pause. In addition to these three elements, *speaking rate*, which is basically the number of linguistic units (e.g., syllables) pronounced in a second, plays an important role in prosody.

Prosody has been shown previously to have a systematic effect on pronunciation variation [Greenberg et al., 2002]. For instance, in accented syllables, i.e., prosodically prominent, the nucleus and coda have a greater tendency to be canonically pronounced than for unaccented syllables. Moreover, the nucleus of accented syllables tends to be longer in duration. There is also a greater likelihood that all the phonemes in such syllables will be realized. Similarly, Shriberg [1994], in her analysis of speech disfluencies argues that accented syllables have a high semantic value. Thus less hesitations occur on words bearing such syllables.

In short, prosody is a critical aspect of speech that plays an important role in conveying suprasegmental information which eventually facilitates the understanding process and makes speech more natural.

Up to this point, we discussed the linguistic and prosodic factors that lead to variabilities in speech. In the next section, we will cover different aspects of expressivity in which, phonetics, phonology, and prosody have a huge impact.

1.3 Expressivity in speech

Speech is an acoustically rich signal. It contains not only a linguistic message, but also considerable personal information about the speaker [Bachorowski, 1999]. This information comprises valuable hints about different aspects of expressivity. Examining this information is crucial for a better understanding of human speech. Thus, in this section, we discuss expressivity in speech and give details on its aspects.

It is often believed that expressivity is a direct reflection of the emotional state of a

speaker. Govind and Prasanna [2013] define expressivity as the vocal indicator of various emotional states that is reflected in the speech waveforms. It can also be considered as an extra level of information that is added to speech. This level of information is mainly attributed to uncontrolled internal states including emotions, feelings, attitudes, moods, and psychological states [Beller, 2009]. Most of the time these internal states are impacted by external factors, e.g., conflicts in our lives, thus, making an already complex concept, even more difficult to understand.

Due to this complexity, we will limit our study of expressivity to the sole aspects related to the problem under investigation. These aspects include emotions, speaking styles and accents. Although the main aim of this thesis is to generate pronunciation variants and disfluencies in the context of spontaneous speech, we believe that all these three aspects of expressivity are highly interconnected.

1.3.1 Emotions

Many definitions have been proposed for emotions, but one of the most comprehensive ones is given in [Kleinginna Jr and Kleinginna, 1981]:

> "Emotion is a complex set of interactions among subjective and objective factors, mediated by neural/hormonal systems, which can (a) give rise to affective experiences such as feelings of arousal, pleasure/displeasure; (b) generate cognitive processes such as emotionally relevant perceptual effects, appraisals, labeling processes; (c) activate widespread physiological adjustments to the arousing conditions; and (d) lead to behavior that is often, but not always, expressive, goal directed, and adaptive."

As the definition suggests and as other studies have shown, speech and emotion have a very strong correlation and emotions play a critical role in communication [Iida et al., 2003]. Emotion might affect a speaker's choice of words, i.e., the speaker mostly utters the type of words that reflect his/her emotional state. Emotions are also tightly bound to acoustic characteristics, specifically fundamental frequency, formant frequencies, intensity and duration [Schröder, 2009].

Emotions can be expressed in different forms. They can generally be categorized into positive and negative emotions. Positive emotions include joy, pride, love, relief, hope, compassion while negative ones include anger, anxiety, guilt, shame, sadness, envy, jealousy, and disgust [Adda-Decker et al., 2005]. In a finer way, emotions can be represented as points of a continuous space. Especially, [Russell, 1980] suggests a 2-dimensional representation of emotions where the first dimension stands for pleasure-displeasure and the second for the arousal degree. Figure 1.4 places most known emotions in this space based on their characteristics. The horizontal axis is for pleasure while the vertical axis

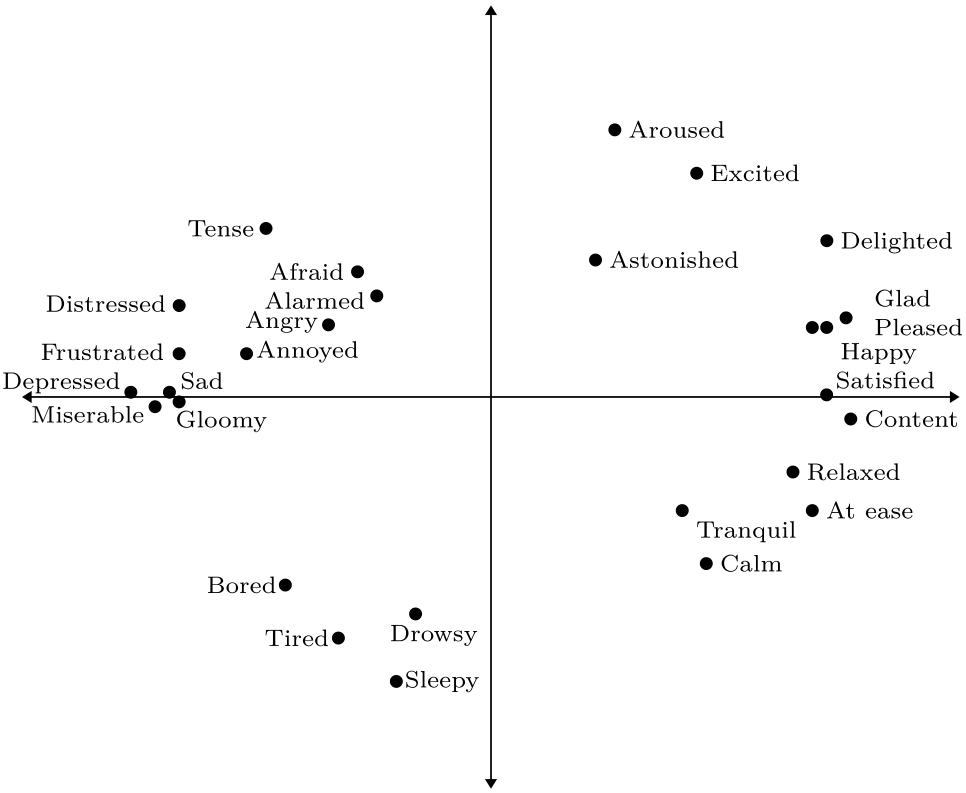


Figure 1.4: 2-dimensional representation of emotions (source: [Russell, 1980]).

is for the arousal degree. For example, anger can be defined as a moderate displeasure (left side) and a neutral arousal (center of the axis), while satisfaction is identified with a high pleasure and a low arousal. Using this representation, the closeness of emotions can be computed as a distance in a Euclidean space.

1.3.2 Speaking styles

The notion of speaking style is still ambiguous to most researchers, as speaking style not only varies from region to region but also from one person to another. However, in general, speaking style can be considered as a deviation from a standard way of speaking and each style has relatively consistent characteristics which enables it to be recognized [Kruschke, 2001]. Moreover, speaking style can be adapted to suit a specific context or environment. According to [Parlikar, 2013], the same speaker can adopt many different speaking styles based on the performed task. For instance, a speaker usually has a relaxed speaking style when talking to a friend in an informal conversation, while the same speaker is likely to change his style of speech into a formal one during a corporate meeting. Thus, in short, speaking style can be considered as a mixture of the way the speaker speaks in general and the context of the speech. The following is

a list of some speaking styles.

- **Prepared speech**: it is a formal type of speech where the speaker has already prepared his speech before articulating it. Due to the preparation process, the number of disfluencies and mistakes is lower than in most other speaking styles. Such kind of speeches are usually formal, thus containing a small number of pronunciation variants as well.
- **Read speech**: the speaker here reads from an already written script, and potentially has rehearsed several times before. Thus, like prepared speech, read speech contains a small number of disfluencies and pronunciation variants.
- **Acted speech**: this speaking style is somehow similar to the previous speaking styles since the speaker reads a script. However, the speaker here is a professional actor who expresses several types of emotions to reflect the type of required acting.
- **Sport comments**: the speaker here gives a commentary of a live sport game. Generally this type of speeches has a standard structure. However, based on the events happening during the game, the content of the speech changes vastly and the speaker might express several types of emotions such as excitement, happiness, etc.
- **Radio or TV interviews**: basically the speaker here hosts a guest and the conversation evolves around some questions answering sessions. Depending on the formality of the interview, the speech can be highly structured, thus being less expressive.
- **Political debate**: it has the form of a discussion between two or more political personalities. In certain cases, the personalities can get aggressive thus expressing several types of emotions like anger, humiliation, etc. In terms of expressivity, this speaking style is close to TV interviews as it has a certain structure, and also bears a great deal of spontaneousness.
- **Spontaneous or conversational speech**: it is an informal, dynamic and unrehearsed type of speech. Since the speaker has not prepared his speech before and the conversation evolves naturally, a vast number of disfluencies and pronunciation variants (based on the accent of the speaker) occur. In addition, speakers might express various types of emotions based on the context of the conversation. Due to this property of spontaneous speech, it is much more expressive than the previously mentioned speaking styles.

Several factors lead to the generation of different speaking styles among speakers. These factors include acoustic-prosodic and phonological variations. Acoustic-prosodic variations include intonation, duration, fundamental frequency and intensity [Laan, 1997] while phonological changes, as suggested by [Adda-Decker and Lamel, 1999], can be related to a variety of factors such as the syllabic structure of words, individual speaker habits, regional dialects and accents. Among these mentioned speaking styles, the spontaneous style is the most complex one as it bears an enormous number of pronunciation variants, disfluencies, and various types of emotions. Because of this, spontaneous speech has received much more attention in the research community. This is also the main topic that is addressed in this thesis, since we try to make TTS more spontaneous through incorporating dedicated pronunciations and integrating disfluencies.

1.3.3 Accents

Accent can be thought of as a particular case of speaking style. As such, it takes part in expressivity in general. Both native and foreign speakers of a language seem to have a specific accent. Native speakers are affected by regional accents (e.g., UK and US English accents), whereas foreign speakers are affected by the patterns which they carry from their own language. As reported by [Arslan and Hansen, 1996], foreign speakers can be identified based on the appearance in their speech of certain patterns which cannot be found in the speech of native speakers. Such accent patterns can be observed in speech through pronunciation variants. Foreign speakers who have acquired the language at an early age are also reported to be able to minimize their accent. Moreover, Arslan and Hansen [1996] define a foreign accent as the patterns of pronunciation features which characterize the speech of a person as belonging to a particular language group. It is also believed that patterns of foreign speakers are more obvious and easier to detect than those of native speakers.

In short, expressivity is an important characteristic of human speech that differentiates speech of individuals from each other and makes speech richer. However, expressivity introduces variability into speech which usually leads to poor performance in speech applications. Moreover, expressivity is a complex concept which is affected by several external factors. Therefore one has to be very specific when dealing with expressivity. Because of this reason, the scope of our work is limited to studying the speaking style aspect of expressivity since we believe that speaking style is highly impacted by accents and emotions. Among the different speaking styles, spontaneous speech was preferred for this work, as it is one of the most variable one, making it a perfect choice for dealing

Table 1.1: Examples of assimilation, elision, epenthesis, reduction, haplology and combination of different phenomena with corresponding phrase/word, canonical and varied forms.

		Word/phrase	Canonical form	Varied form
1	Assimilation	can be	/kæ**n** biː/	/kæ**m** biː/
2	Elision	last month	/læs**t** mənθ/	/læs mənθ/
3	Epenthesis	vanilla ice cream	/vənɪlə aɪskriːm/	/vənɪlə **r**aɪskriːm/
4	Reduction	and	/**æ** nd/	/ənd/
5	Haplology	library	/laɪ.brə.ri/	/laɪ.bri/
6	Combined	bread and butter	/brɛ**d æn**d bʌtə/	/brɛ**b m** bʌtə/

with pronunciation variants and speech disfluencies.

The impacts of expressivity on pronunciation and fluency are discussed in the next section, while a literature review of works exploiting expressivity for TTS systems is presented in Chapter 2.

1.4 Pronunciation and fluency

Expressivity, as we have already discussed, introduces a lot of variability into speech. Such variability can be observed through most of the layers of speech (see Section 1.1.2) such as phonology, prosody, and acoustics. What interests us in this thesis are the variabilities in the phonological layer, particularly those which affect pronunciation and fluency in the context of spontaneous speech. This section covers pronunciation variation and disfluencies and mentions the most important factors leading to their presence in speech.

1.4.1 Pronunciation variation

Pronunciation variation is observed more often in spontaneous speech than in any other speaking styles. Phonetic context, word predictability and prosodic properties of speech are considered as the main reasons for this great deal of variation [Bates and Ostendorf, 2002, Greenberg, 1999, Fosler-Lussier and Morgan, 1998]. Among phonetic factors, assimilation, elision, epenthesis, reduction, and haplology play the biggest role in introducing variations to pronunciation. Hardcastle et al. [2010] define the assimilation phenomenon as:

> "The contextual variability of speech sounds, by which one or more of their phonetic properties are modified and become like those of the adjacent segment."

This definition looks very similar to that of the coarticulation and the two terms are sometimes used interchangeably. However, some researchers distinguish between the two terms clearly, assimilation mostly being accounted for by phonological rules and related to a specific language, while coarticulation being the physiological process and mostly governed by universal rules [Hardcastle et al., 2010]. One can think of their relation as coarticulation being the cause and assimilation the effect [Frawley, 2003]. It is worth mentioning that assimilation does not necessarily occur in spontaneous speech only. However, due to the previously mentioned factors, it is observed more commonly in this speaking style. To better explain assimilation, let us look at an example. In the phrase "can be" (Table 1.1, *line 1*), the /n/ sound usually assimilates to /b/ and becomes /m/. This type of assimilation is called alveolar nasal assimilation. As a bilabial plosive (refer to Figure 1.3) /b/ sound directly follows an alveolar nasal sound /n/, the latter assimilates to the sound /b/ and becomes more like a bilabial nasal /m/ sound in English.

As stated before, pronunciation variation does not occur as a result of assimilation only, but there also exists four other phenomena: elision, epenthesis, reduction and haplology. Elision is the omission of one or more sounds. In the example given in Table 1.1, *line 2*, we can see that the /t/ sound is most of the time not realized during speech. In contrary to elision, epenthesis is the insertion of one or more sounds. The insertion of the /r/ sound in the phrase "vanilla ice cream" (Table 1.1, *line 3*) is an example of epenthesis. On the contrary, reduction happens when a vowel is reduced to a shorter form; for example the /æ/ sound is mostly reduced to /ə/ in the word "and" as shown in Table 1.1, *line 4*. Lastly, haplology is the deletion of successive identical syllables or consonant sound groups. For instance, the second syllable in the example given in Table 1.1, *line 5* has been completely deleted since it has an almost identical pronunciation as the third syllable. Sometimes several phenomena can be applied on the same word or even several successive words and lead to vast changes in speech. In the phrase "bread and butter" (Table 1.1, *line 6*), the word "and" is mostly reduced in British English through reduction, elision, assimilation and becomes only /m/. The effects of these phenomena spread even to the previous word "bread" by transforming the last sound /d/ to a /b/ sound.

Apart from the phonetic context, word predictability also affects pronunciation in spontaneous speech. According to [Bates and Ostendorf, 2002], speakers adjust their articulators to accommodate the importance of the information in their speech. Thus, certain phonemes are hyper-articulated during points of emphasis and reduced at predictable points. For example the word "for" is usually pronounced with a reduced form /fɔ/ rather than the complete form /fɔr/, since it is one of the most predictable words in English. Further, word predictability might also affect the perception of words. In

an experiment conducted in [Lieberman, 1963] (cited by [Fosler-Lussier, 1999]), subjects were asked to recognize examples of words extracted from both predictable and unpredictable contexts. The results showed that predictable words were more difficult for subjects to understand than unpredictable words. This difficulty is probably related to the fact that the examples of predictable words are on average shorter in length since they are pronounced with reduced forms. In another experiment in [Fowler and Housum, 1987], it was shown that when a word is articulated for the second time within the context of a long speech, its duration generally gets shorter than that of the first occurrence of the same word. As a result, the second occurrence of the word gets less intelligible.

Lastly, prosody has also been shown to impact pronunciation to a certain degree. Wightman and Ostendorf [1994] show that speaking rate, stress and phrase boundaries are all directly related to duration. Short durations reflect either a fast speaking rate or a phonetic reduction, while longer durations can be related to a phrase final lengthening or fast speaking rate. Similarly, Fosler-Lussier and Morgan in [Fosler-Lussier and Morgan, 1999] reported that phone deletion rate rises from 9.3% to 13.6% from very slow to very fast speech rate. Moreover, words bearing pitch accent have been reported to be hyper-articulated and to suffer less co-articulation than other words [Chen and Hasegawa-Johnson, 2004]. Lastly, Bates and Ostendorf [2002] showed that word duration and energy have also similar effects on pronunciation.

In short, pronunciation variation is one of the most pervasive characteristics of spontaneous speech in which phenomena like phonetic context, word predictability, and prosody have a major impact.

1.4.2 Speech disfluency

Speech disfluencies can be defined as a phenomenon which interrupts the flow of speech and does not add any propositional content [Tree, 1995]. Disfluencies are very frequent in spontaneous speech, and are among the characteristics that distinguish spontaneous speech from read speech [Stolcke and Shriberg, 1996]. According to [Tree, 1995] approximately 6% of words uttered in a spontaneous context are some form of disfluencies. One of the main reasons why disfluencies appear so frequently in spontaneous speech is related to the thinking process. Basically when the speed of speaking becomes faster than the speed of thinking—particularly in cases where the speaker has not prepared his speech in advance—the speaker tends to use disfluencies until the content resulting from the thinking process is ready [Goto et al., 1999].

Despite the lack of propositional content, disfluencies have several communicative values. As pointed out by [Clark, 2002], disfluencies facilitate synchronization of speech

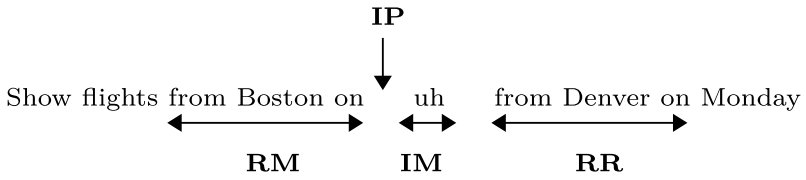


Figure 1.5: Standard structure of disfluencies (source: [Shriberg, 1994]).

between addressees in conversations. Moreover, they improve listening comprehension by creating delays in speech and signal the complexity of the upcoming message [Tree, 2001, Rose, 1998] (cited by [Adell et al., 2012]).

Historically, disfluencies have been considered as irregular phrases by most researchers and therefore have received very little attention. However, some studies have shown that there are actually remarkable regularities in the structure of disfluencies [Shriberg, 1994]. In order to be able to analyze and study disfluencies, first, a standard structure which can encompass all the different types of disfluencies has to be defined. The necessity for such a standard becomes more clear when the different disfluency types are introduced in the next subsection as each one has a different structure. Several structures have been proposed by researchers in the past [Levelt, 1983, Clark, 1996]. These structures are very similar to each other, however the most widely accepted structure and the one that is adopted in this thesis is the structure proposed by Shriberg [1994]. Shriberg suggests some standard terms for different regions of disfluencies as shown in Figure 1.5. The region called Reparandum (RM) refers to the erroneous part of the speech. Some researchers consider the entire erroneous region as RM, while others relate RM only to the mistakenly uttered word such as "Boston" in the given example. The Interruption Point (IP) is the exact place in which the interruption occurs, that is, when the speaker detects a trouble in his speech. The next region is Interregnum (IM) (also referred to as Editing Term (ET) by some researchers) is the start of an editing phase or correction phase. Finally, Repair (RR) is the region in which the speaker corrects his speech.

Several studies have suggested to categorize disfluencies into three main types: pauses, repetitions, and revisions [Mareüil et al., 2005, Shriberg, 1999, Tseng, 1999]. In this section, each of these disfluency types is briefly described along with examples and their functions in speech. It is important to mention that in the literature, several different terminologies can be found to represent these disfluency types. However, in this thesis, we will adopt the aforementioned terminology.

1.4.2.1 Pauses

Pauses are considered to be disfluencies in speech since they do not add any meaning to the spoken utterance, that is, the utterance will still be complete without the pause. Among the disfluency sections (*cf.* Figure 1.5), pauses are always used as an IM. The "uh" in the following utterance is an example of a pause:

Example 1: I will $\overset{IM}{\overbrace{uh}}$ go to supermarket.
$\underset{IP}{\uparrow}$

Generally, pauses are subdivided into silent pauses and filled pauses (FPs) such as "uh" and "um" [Swerts et al., 1996, Duez, 1982]. However, in this thesis, discourse markers like "I mean", "well", "you know" are also considered as pause types. This is because most of the time such discourse markers do not convey any meaning apart from acting like a pause to help the listener understanding how the new speech is linked to what was previously said [Heeman and Allen, 1999].

These mentioned pause types have different functions. Filled pauses can be used to indicate the beginning of a delay to search for a phrase and keep the conversation going on [Clark and Fox Tree, 2002]. As for silent pauses, in an exploratory study [Mahl, 1959] (cited by [Rochester, 1973]), it was found that they tend to be more used when the speaker is anxious. Lastly, according to [Fox Tree and Schrock, 2002], discourse markers like "you know" and "I mean" have several functions such as providing information about the speaker, including anxiety, uncertainty or lack of self confidence. Moreover, they are used to express shared understanding on a topic, usually referred to as positive politeness, which helps decrease the social distance between the speakers and makes speech more casual.

1.4.2.2 Repetitions

Repetitions can consist of one-word repetitions, such as "the the", or multiple-word repetitions, like "I will I will". Repetitions are mostly common in unplanned talks. Shriberg [1994] reports that repetitions can function as a pause for gaining time by repeating words and sometimes help in recovering the flow of the speech after a long pause. Repetitions can also have rhetorical purposes which intensify the effect of an expression or might be used to signal an upcoming problem in the speech [Tseng, 1999]. Example 2 shows an utterance with a single-word repetition and a filled pause in between the repeated words.

Example 2: $\overset{RM}{\overbrace{I}}$ $\underset{IP}{\uparrow}$ $\overset{IM}{\overbrace{uh}}$ $\overset{RR}{\overbrace{I}}$ want to go.

1.4.2.3 Revisions

A revision occurs when the speaker interrupts an utterance, and then begins again by a revised (slightly changed) version of the utterance [Tseng, 1999]. Revisions have no obvious function in speech apart from helping speakers monitoring their speech and interrupting when a trouble is detected as shown in Example 3.

False starts are another form of revisions in which the speaker completely abandons the interrupted utterance and starts a fresh one. An utterance with false starts is given in Example 4 .

Example 3: I think $\overbrace{\textit{she will}}^{RM}$ $\underset{IP}{\uparrow}$ $\overbrace{\textit{I mean}}^{IM}$ $\overbrace{\textit{he will}}^{RR}$ not come today.

Example 4: $\overbrace{\textit{it was}}^{RM}$ $\underset{IP}{\uparrow}$ $\overbrace{\textit{he liked it.}}^{RR}$

To conclude, we can say that although disfluencies have been considered as phenomena interrupting the flow of speech, thus making comprehension more difficult, several studies have reported that disfluencies actually help listeners better understand the content of the speech in many cases [Brennan and Schober, 2001, Fox Tree and Schrock, 1999]. This shows that studying disfluencies and understanding them can lead us to better understand speech and produce more natural and expressive synthetic speech.

1.5 Conclusion

In this chapter, we explained the basic concepts about speech and expressivity which are going to be useful to understand the topics covered in the next chapters. First, the basics of the human speech production mechanism were discussed and the abstract layers of speech were described. Three of these layers which have the highest impact on pronunciation variation and disfluencies were covered: phonetics, phonology and prosody. For each aspect, we mentioned the most important elements and how they can be related to the problem covered in this thesis. Next, we reviewed expressivity which is an important characteristic of human speech and went through the notions of expressivity including emotion, speaking style and accent. Finally, we presented pronunciation and fluency by discussing the pronunciation variations and speech disfluencies that occur in spontaneous speech as a result of expressivity in speech.

In the next chapter, a general background on speech synthesis alongside a review of the works in the area of pronunciation variation and disfluency modeling for TTS will be provided.

Chapter 2

Pronunciation and disfluencies in speech synthesis

In the first chapter, the necessary background about speech and different aspects of expressivity was provided. In this chapter, the main focus is going to be on describing the different ways of exploiting expressivity to have more human-like speaking machines. Although the idea of integrating expressivity, particularly pronunciation variation and disfluencies can be applied to both TTS and ASR, in the literature, studies dealing with ASR have always outnumbered studies on TTS. Because of this reason, the focus of this thesis is going to be mainly on the TTS side with the idea of making it more expressive. Before providing details on the integration of pronunciation variants and speech disfluencies in TTS systems, we first explain how TTS systems work in Section 2.1, and describe the most common techniques used in TTS systems. Then, different machine learning techniques that are used in speech synthesis systems

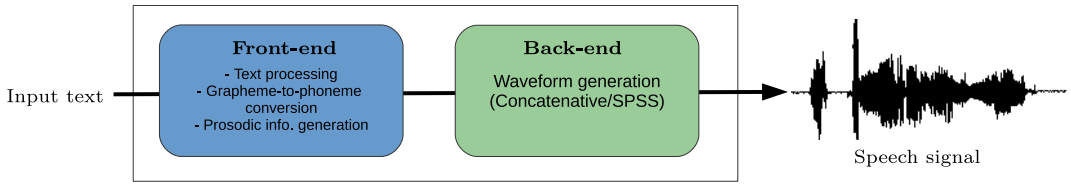


Figure 2.1: Speech synthesis architecture.

are discussed in Section 2.2, while pronunciation modeling with a focus on the problem of pronunciation variation is covered in Section 2.3. Finally, in Section 2.4, integrating speech disfluencies in TTS systems is described.

2.1 General overview of speech synthesis

The speech synthesis problem is generally seen as a two staged process as shown in Figure 2.1. The first stage of the process is usually referred to as the front-end which involves extracting linguistic information from the input text. This step includes transforming the text into a machine readable format, assigning the corresponding phonemes for each letter in the input text and deriving prosodic information. The second stage is the back-end, which is responsible for generating the speech waveform from the linguistic information. The front-end and the back-end are fairly independent, which makes the whole process more flexible. Further details on each process are provided in the following two sections. It is worth mentioning that main sources of the following sections are the books "Text-to-speech Synthesis" [Taylor, 2009], and "Speech Synthesis and Recognition" [Holmes, 2001].

2.1.1 Front-end

As previously mentioned, the front-end is responsible for extracting linguistic information from the input text. This process involves a sequence of steps. The first one is text processing. It includes tokenization and text segmentation in which the input text is split into separate words and the sentence boundaries are identified respectively. It also includes text normalization whereby all alphanumeric characters, numerals, abbreviations, etc. are converted into plain words. In addition, the part-of-speech (POS) tags are also assigned in this step. POS tagging is critical in resolving pronunciation conflicts in some languages, as the pronunciation of certain words changes based on their POS. Once POS tags have been identified, the phrase breaks can be determined. Phrase breaks are particularly important as they determine the phrases and clauses inside a sentence. This information will potentially be used to derive prosodic information.

The next step is to generate the phonemic transcriptions for each word. This is

usually done using a lexicon or a pronunciation dictionary and a Grapheme-To-Phoneme (G2P) converter. The lexicon is used to store the pronunciation of each word explicitly, while the G2P converter is used to generate pronunciations of unknown words. As one of the main scopes of our work is to deal with pronunciations, a detailed description of pronunciation generation techniques is given in Section 2.3.

Finally, the front-end predicts prosodic information such as duration and intonation from the input text. This information can be acquired through expert rules or machine learning. In both cases the goal is to make the output speech more natural.

2.1.2 Back-end

The back-end utilizes the information provided by the front-end to synthesize the speech waveform. Several approaches exist such as articulatory synthesis, formant synthesis, concatenative synthesis and statistical parametric synthesis. Among them, concatenative and statistical parametric speech synthesis are dominant [Taylor, 2009]. A review of these two approaches is provided in the following sections.

2.1.2.1 Concatenative speech synthesis

In this approach, short segments of speech are retrieved from a pre-recorded speech database based on the phonemic transcriptions provided by the front-end. The retrieved segments are then concatenated in an appropriate order to produce the desired utterance. The database for concatenative synthesis is prepared by recording several hours of speech from one speaker and then segmenting them into small units. Then phonetic, acoustic and prosodic features are extracted for every unit and stored alongside the unit. As the synthetic speech is generated from real speech, its quality is very high.

Various types of units can be used such as words, syllables, diphones and phones. Among them, diphones are the most widely used. A diphone unit starts in the middle of one phone and extends to the middle of the next one. In other words, it consists of two half phones. For instance, the word "seen" /s iː n/ can be decomposed into four diphones when surrounded by two silences, as shown in Figure 2.2. Diphones are particularly better than phones for concatenative synthesis since diphones start and end in the middle region of the phones which is considered to be more stable than regions at the edges. Therefore, concatenation at the middle of a phone is known to produce less acoustic artefacts. Diphones are also more appealing than words and syllables because less units are required to cover all possible utterances in a language.

Concatenative synthesis can be divided into two sub-types: diphone synthesis and unit selection synthesis [Schultz and Kirchhoff, 2006]. In diphone synthesis, only one

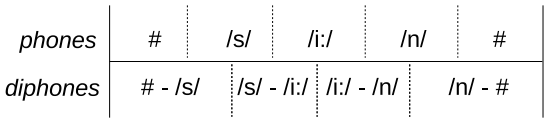


Figure 2.2: Representation of phones and diphones for the word "seen". # represents a silence where the phoneme is not followed/preceded by any phoneme.

instance of each diphone is stored in the database. Then, based on the phonemic transcription provided by the front-end, corresponding diphone units are retrieved from the database and concatenated together. In the case of unit selection, several instances of each diphone are generally present in the database, each with different phonetic, prosodic and acoustic characteristics. This variety of realizations can be used to capture coarticulation and other phonetic variations, thus making the generated speech more natural. During synthesis, an algorithm selects the best units from the database according to two criteria: *target cost* and *concatenation cost*. The former measures the distance between a candidate unit in the database and the desired target unit, usually based on duration, F0, and/or the phonemic context. The latter cost measures how well the candidate unit will be joined to the neighboring units based on acoustic features like F0, amplitude, etc. The total cost for a candidate sequence of units is then computed by summing the target and concatenation costs over all its units with respect to the desired ones. Finally, the sequence which minimizes this score can be found using best path algorithms like Viterbi or A^* [Guennec and Lolive, 2014].

Concatenative synthesis in general has some major drawbacks. First, it requires a very large speech database to cover all the possible diphones in different contexts [Benesty et al., 2007]. Second, it is difficult to control expressivity in the generated speech as the speaking style of TTS databases usually does not vary. Thus, a solution is to have several speech databases for different emotions and speaking styles and to choose one based on the target expressivity. As this is an expensive task, a more flexible but also more difficult option is to take into account expressivity in the front-end. This is the approach that we follow in this thesis by adapting pronunciations and integrating disfluencies in order to reflect a specific speaking style.

2.1.2.2 Statistical parametric speech synthesis

In contrast to concatenative synthesis, statistical parametric synthesis (SPSS) tries to generate speech signals using parametric models [Zen et al., 2009]. More precisely, these parametric models are used to predict speech parameters from which speech is reconstructed using a vocoder. Thanks to its flexibility, SPSS has attracted much more interest than concatenative synthesis in the last years [Hirose and Tao, 2015].

In the last decade, parametric models mostly relied on Hidden Markov Models (HMMs) [Drugman et al., 2008, Zen et al., 2009], while, more recently, deep neural networks (DNNs) have been shown to be another (usually better) solution [Zen et al., 2013]. In HMM-based TTS, the model is dynamically built based on an input sequence of phonemes and their corresponding features, e.g., the right and left phonemes, POS, syllable stress, etc. The resulting HMM is used in a reversed way as usually, i.e., the states are used to generate the speech parameters. In practice, each input phoneme is represented by a 3-state or 5-state HMM, where each state is associated with a GMM over speech parameters. When building the HMMs, these states are retrieved from a set of trained states based on their corresponding features. Retrieval is done using decision trees, where output states may either correspond to the exact desired context or represent the average of tied states, that is the average of relatively similar contexts. This tying mechanism is particularly important at training time when too few data is present to reliably estimate the states. At runtime, the overall method can propose states for any input even for those that have never been observed during the training. This particular property of SPSS provides a great advantage for synthesizing spontaneous speech over unit selection systems, since spontaneous speech contains very rare combinations of phonemes which might not exist in the prerecorded speech database. In such cases, a unit selection system will try to find a sufficiently similar candidate unit for the target phoneme in the database, but this candidate might be far from the desired phoneme. As for SPSS, the missing phoneme is averaged based on several similar phonemes, which provides a better approximation [King, 2010]. Since decision trees are inefficient in modeling complex context dependencies, Zen et al. [2013] suggest to replace the decision trees with DNNs. In more advanced approaches, DNNs have been used to predict speech parameters, thus replacing HMMs, and even directly the raw waveforms, thus also playing the role of a vocoder [van den Oord et al.].

To sum up, statistical parametric synthesis in general has the advantage of being able to generate acceptable quality speech even on a small speech database in contrast to unit selection where a large database is required to produce good quality speech [King, 2011]. Moreover, as the models only predict the speech parameters and not the actual speech signal, it is much easier to control prosodic features and model expressivity [Yamagishi et al., 2005].

In the next section, we will review various machine learning approaches which play an important role in different stages of speech synthesis.

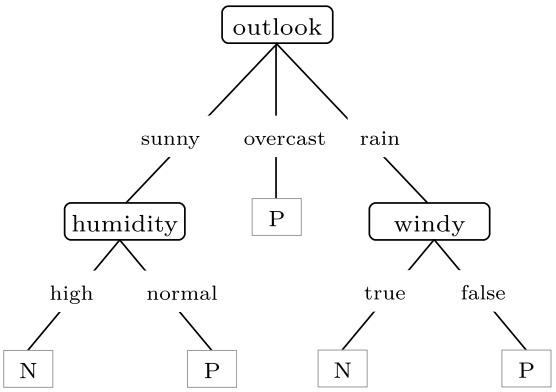


Figure 2.3: An example of a simple decision tree (source: [Quinlan, 1986]).

2.2 Review of machine learning techniques for speech synthesis

Speech synthesis heavily rely on machine learning in various stages. For example, in the front-end, POS tags are learned from data and predicted in the runtime using machine learning algorithms like HMMs or conditional random fields (CRFs). Similarly CRFs and DNNs are widely used in the grapheme-to-phoneme conversion process. In the back-end, as we mentioned, HMMs, DNNs, and decision trees are common in SPSS. This section provides a basic understanding of the machine learning approaches used in the context of TTS in general and compares them in order to determine which is the most adequate for our problem.

2.2.1 Decision trees

Decision trees are one of the most widely used classification techniques. They classify data instances in a tree structure by sorting them down from a root node to leaf nodes. At each node, an attribute of the data instance is evaluated based on its possible values. This process continues until a leaf node is reached which also determines the classification decision for this particular data instance. At training time, the choice of the features to be examined relies on measures like entropy or Gini index. Decision trees can also be represented as if-then rules to make them more human readable. Figure 2.3 shows a simple decision tree used in clustering data instances using three features into two classes, *P* and *N*.

The main reason of the widespread usage of decision trees lies in the fact that they are easy to interpret. However, decision trees can get very complex quickly as the size of the tree grows exponentially with the number of attributes. In addition, overfitting in decision trees is a major problem particularly when the tree has too many nodes

relatively to the amount of training data available.

2.2.2 Hidden Markov models

Hidden Markov models are a powerful statistical method to characterize the observed data samples of discrete-time series [Huang et al., 2001]. Among other domains, HMMs have been successfully used in automatic speech recognition, speech synthesis, language modeling and part-of-speech tagging [Huang et al., 1990, Zen et al., 2009, Stolcke et al., 2002, Kupiec, 1992]. Basically, given an input sequence of observations $\mathbf{x}$, HMMs compute a probability distribution over possible sequences of hidden states (labels) and determine the best label sequence $\mathbf{y}$ as:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \Pr(\mathbf{y}|\mathbf{x}). \tag{2.1}$$

Instead of computing the probability that an observation sequence generates a label sequence $\Pr(\mathbf{y}|\mathbf{x})$, HMMs compute the probability of an observation sequence given the label sequence, using Bayes' rule:

$$\begin{aligned} \hat{\mathbf{y}} &= \arg\max_{\mathbf{y}} \Pr(\mathbf{y}|\mathbf{x}) \\ &= \arg\max_{\mathbf{y}} \frac{\Pr(\mathbf{y})\Pr(\mathbf{x}|\mathbf{y})}{\Pr(\mathbf{x})} \\ &= \arg\max_{\mathbf{y}} \Pr(\mathbf{y})\Pr(\mathbf{x}|\mathbf{y}), \end{aligned} \tag{2.2}$$

where $\Pr(\mathbf{y})$ is the prior probability of a particular label sequence and $\Pr(\mathbf{x}|\mathbf{y})$ is the probability of an observation sequence given a label sequence. The problem here is that $\Pr(\mathbf{y})\Pr(\mathbf{x}|\mathbf{y})$ is still very difficult to compute. Therefore, two simplifying assumptions have to be made: (i) the probability of an observation appearance is only dependent on its own label, (ii) the probability of a label is dependent only on its preceding label. As a result, Equation 2.2 can be rewritten as:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \Pr(\mathbf{y}|\mathbf{x}) \approx \arg\max_{y} \prod_{i=1}^{n} \Pr(x_i|y_i)\Pr(y_i|y_{i-1}), \tag{2.3}$$

where $\Pr(y_i|y_{i-1})$ is the transition probability and represents the probability of a label given its preceding label, and $\Pr(x_i|y_i)$ is the observation probability, which represents the probability of an observation given a particular label.

The transition probabilities can be calculated by counting occurrences of sequences of labels. To compute observation probabilities, a separate model for every possible label is needed, each defining a probability distribution over the set of observations.

Given the transition and observation probabilities, the sequence of labels underlying the sequence of observations can be identified through a decoding process. The most common algorithm used for decoding is Viterbi which conducts a search through all possible label sequences to find the most likely one.

HMMs, as we have already mentioned, have been successfully used in many speech and natural language processing (NLP) tasks. However, their major drawback is that they only capture dependencies between a state and its corresponding observation. This is a major problem for tasks where dependencies between several states in the sequence have to be taken into account, for instance, the task of predicting pronunciation variants.

2.2.3 Conditional random fields

CRFs, just like HMMs, are probabilistic models for labeling sequential data [Lafferty et al., 2001]. They model the conditional probability of a sequence of T labels $\mathbf{y} = (y_1, \ldots, y_T)$ given an input sequence of observations $\mathbf{x} = (x_1, \ldots, x_T)$ as follows:

$$\Pr(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_\theta(\mathbf{x})} \exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K} \theta_k f_k(y_{t-1}, y_t, x_t)\right) , \tag{2.4}$$

where $f_1, \ldots, f_K$ are K so-called feature functions, $\theta_1, \ldots, \theta_K$ are their associated weights estimated on training data such that the error rate on a given development set is minimized, and $Z_\theta(\mathbf{x})$ is a normalization factor.

Feature functions are a powerful mean to combine input information. They typically return 1 when the condition of the feature is met, 0 otherwise. An example of a feature function in the context of pronunciation modeling might be "the output phoneme y_t is /dʒ/ when the POS tag of the previous word x_t is *Noun*":

$$f_1(y_{t-1}, y_t, x_t) = \begin{cases} 1 & if\ y_t = /\text{dʒ}/ \text{ and } x_t = Noun \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

If desired, feature functions can also take advantage of the previous predicted phoneme y_{t-1} to predict y_t. This configuration is referred to as *bigram* configuration, as opposed to *unigram* when only y_t is considered. Unigram and bigram feature functions can be considered together (referred to as *uni+bigram* in the remainder):

$$f_1(y_{t-1}, y_t, x_t) = \begin{cases} 1 & if\ y_{t-1} = /\text{ʊ}/ \text{ and } y_t = /\text{dʒ}/ \text{ and } x_t = Noun \\ 0 & \text{otherwise.} \end{cases} \tag{2.6}$$

When training a CRF model, hundred of thousands of such feature functions (based on the size and sparsity of the data) are created, and for each one, a weight θ_i is

estimated. The weights are learned from the data by computing the gradient of an objective function using an algorithm like Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS). At the runtime, each feature is tested on each observation in the sequence. When a feature function is active, i.e., returns 1, it increases the chances of assigning its label y_t to that particular observation.

One major advantage of CRFs over HMMs is that they can capture dependencies across all the different types of features and not only the label and its observation. For instance, HMMs cannot take into account information about the next observations in the sequence, while CRFs can be configured to consider any of the neighboring observations. Both of our contributions on pronunciation variation and disfluency generation which are presented in Chapter 4 and Chapter 5 are highly based on CRFs.

2.2.4 Artificial neural networks

The last machine learning technique that we will review here are artificial neural networks (ANNs). Due to their functional similarity with biological neurons in our brains, ANNs have received a lot of attention from researchers and have been used extensively in pattern recognition problems [Bishop, 1995]. Neural networks are particularly interesting for speech related problems where many constraints have to be satisfied and evaluated in parallel [Huang et al., 2001]. ANNs, according to [Príncipe et al., 2000] (cited by [Cannas et al., 2006]) can be defined as:

> "Distributed, adaptive, generally nonlinear learning machines built from many different processing elements (PEs). Each PE receives connections from other PEs and/or itself. The interconnectivity defines the topology. The signals flowing on the connections are scaled by adjustable parameters called weights."

ANNs are arranged in layers like a graph. The network shown in Figure 2.4 has an input, a hidden, and an output layer. Each layer in the network has an array of neurons. An observation is represented in terms of numerical values fed into the network through the input layer and flow through the neurons. Each neuron receives an input value, transforms it and transfers the result through its output to the next layer. In most cases, the results are yielded in terms of probabilities from the output layer in which each node represents a possible classification label. Thus, the node with the highest probability is considered as the correct label. As an example, in the context of pronunciation modeling, the input layer neurons might represent a grapheme of a word and its corresponding features, while the output layer neurons represent the possible phonemes of that grapheme.

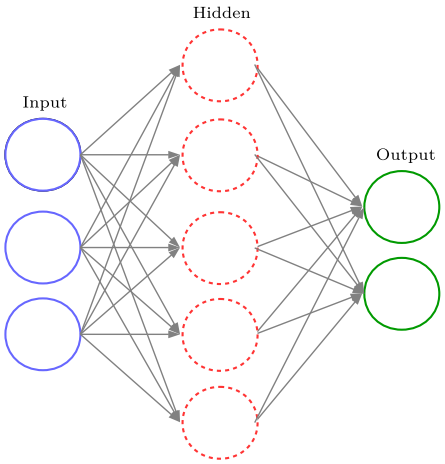


Figure 2.4: Scheme of an artificial neural network.

Based on the arrangement of the layers and the connections between the neurons several typologies can be defined such as DNNs where the network has several hidden layers, Recurrent Neural Networks (RNNs), where the outputs are taken and fed back into the input or the hidden layer neurons, or Long Short Term Memory (LSTM), which are a special kind of RNNs, capable of learning long-term dependencies.

With all these different architectures, ANNs have gained significant popularity within the domain of speech due to their ability to learn and generalize complex patterns of speech [Yin et al., 2015]. However, ANNs are considered to be black boxes since it is very difficult to determine which variables are the most important contributors to a particular output in the trained models, thus, important features cannot be easily identified [Tu, 1996].

In short, each of these mentioned machine learning approaches has pros and cons and is used in different stages of speech synthesis. As we will see in the next two sections, all of these machine learning techniques have been used in the literature for both pronunciation and disfluency modeling. However, based on the descriptions provided earlier, it appears that CRFs and ANNs are the two best approaches for dealing with these two problems. Both approaches have the ability to capture complex relationships in the data. What differentiates them is the fact that it is very difficult to understand the underlying trained model and the importance of features in the case of ANNs, while, important features of CRFs can be easily analyzed based on their assigned weights. This is a critical point, since working with spontaneous speech requires deeper understanding of the problem. Hence, we believe that CRFs have the potential to perform well for modeling pronunciations and disfluencies. The next two sections provides a literature review of pronunciation and disfluency modeling where the aforementioned machine

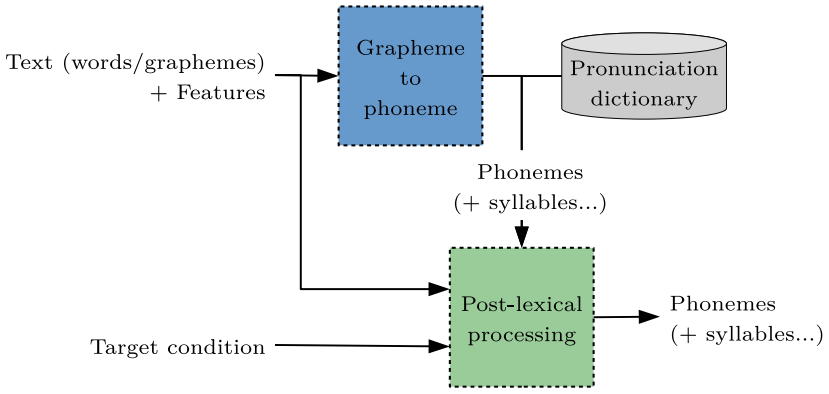


Figure 2.5: Overview of the main modules and data involved in pronunciation modeling.

learning approaches play an important role.

2.3 Pronunciation modeling

The most basic function of pronunciation modeling is to link the orthographic representation of written words with its corresponding phonemic transcription. In the frame of speech synthesis, its main usage is to generate this phonemic transcription from the input text. For example, given the word "chair", a pronunciation model will assign the following phonemes:

$$\text{chair} \rightarrow /\text{ʧɛr}/.$$

The phonemic representations are most of the time based on standard pronunciations, meaning that variants due to regional accents, speaking styles, etc. are not considered. These standard pronunciations will be referred to as *canonical pronunciations* in the rest of this thesis. In more sophisticated approaches, extra features can be considered along with the input words such as POS, etymology, etc. Likewise output phonemes can include information about syllables, stress, etc. Alternatively the output can be the list or the lattice of all possible pronunciations without explicitly determining the one to be chosen. Producing canonical pronunciations is achieved by using a lexicon and a G2P converter.

As illustrated in Figure 2.5, when dealing with the issue of pronunciation variants, generally, an additional step which is usually referred to as *post-lexical processing* is required. A post-lexical processor takes the output from a G2P converter or a pronunciation dictionary in order to modify or rerank the canonical pronunciations such that they reflect a specific target condition, e.g., speaking style, accented speech, emotion, etc.

In the rest of this section, we will first review the G2P conversion process, and then discuss post-lexical processors.

2.3.1 Grapheme-to-phoneme conversion

G2P is one of the most crucial tasks in any speech synthesis system. According to [Taylor, 2009], the objective of G2P conversion is to generate a sequence of phonemes for a given word from its spelling. That means transforming a sequence of graphemes to a sequence of phonemes. G2P conversion is known to be a difficult task in those languages in which graphemes might have different corresponding phonemes in different contexts. For instance, in English, the graphemes "ch" is pronounced as /k/ in the word "chemistry" and as /tʃ/ in the word "chair".

Several approaches exist for performing G2P conversion including knowledge-based, data-driven, and statistical ones. In this section, we provide a review of these approaches.

2.3.1.1 Knowledge-based techniques

The most straightforward G2P technique is to store all the possible pronunciations in a pronunciation dictionary and then to look up in this dictionary to retrieve the phonetic transcription of each input word. This technique has the disadvantage of not being able to predict the pronunciations of out-of-vocabulary (OOV) words.

In a more flexible approach, rule-based techniques—which are based on the idea that the pronunciation of a grapheme can be predicted from its context—can be applied [Pathak and Talukdar, 2013]. Most of the early rule-based systems consisted of hand-written rules. The drawback of this technique is that it requires experts to craft the rules for each language in order to achieve a good performance. In addition, it might perform very poorly in languages such as English where the connection between graphemes and phonemes can be ambiguous [Lafferty et al., 2001].

2.3.1.2 Data-driven techniques

Instead of using hand-written rules, data-driven techniques seek to automatically learn the rules from examples. Such techniques can rely on analogy or decision trees.

The idea behind pronunciation by analogy comes from the studies of how humans learn the pronunciation of new words. When a human is given a new word, he/she learns its pronunciation by comparing it to the nearest known words and adapting or combining their pronunciations [Taylor, 2009, Dedina and Nusbaum, 1991]. For instance, considering the word "fax" as our target new word, a human would automatically think of a similar word such as "tax" and adapt its pronunciation. Pronunciation by analogy

algorithms work by comparing substrings of the unknown word to those extracted from the pronunciation dictionary and find the closest match [Damper and Eastmond, 1997]. Pronunciation chunks are then joined to obtain the final pronunciation.

Decision trees have also been widely used for modeling pronunciation in speech synthesis [Kienappel and Kneser, 2001, Han and Chen, 2004]. They simply predict a phoneme for each input grapheme by asking questions related to the context of the grapheme. More complex configurations can determine the questioned grapheme context in a dynamic way as the tree grows [Pagel et al., 1998]. There also have been studies on using more generalized trees which can take into account phonological structures and stress information [Vazirnezhad et al., 2009]. Such trees have proven to yield better results than traditional ones.

One major problem with the mentioned data-driven techniques is that they do not take into account the information about the previously predicted phonemes. Such information can be extremely useful for pronunciation modeling.

2.3.1.3 Statistical techniques

Statistical techniques are more recent compared to the two other approaches. ANNs, HMMs, CRFs and joint n-gram models are the most widely known statistical techniques for G2P conversion. Most statistical techniques first, align the graphemes and phonemes such that each grapheme corresponds to its phoneme, using, for instance, a Dynamic Time Warping (DTW) algorithm [Pagel et al., 1998]. Models are then trained on the aligned grapheme-phoneme pairs and their related features.

One of the first examples of using statistical techniques in building G2P systems is the NETtalk system [Sejnowski and Rosenberg, 1987] (cited by [Taylor, 2009]). NETtalk relied on a neural network which consisted of 3 layers: an input layer of 203 neurons, a hidden layer of 80 neurons and an output layer of 26 neurons representing the phonemes to be produced. The network only considered seven graphemes at a time, that is the target grapheme and three graphemes from the right and three from the left. The input and output of the system were encoded with various features (voiced, stress, syllable boundary, etc.).

When it comes to using HMMs for pronunciation modeling, one major advantage they offer is that the model is allowed to use the previously predicted phonemes for future decisions. Phonemes are represented as states of a Markov chain and linked to their most likely corresponding graphemes [Karanasou, 2013]. An example of using HMM in G2P is the work of Taylor [2005]. The author suggests that HMMs alone are not sufficient for G2P modeling; however, by adding a preprocessing step, they can be improved. The preprocessing step rewrites some of the graphemes and rearrange them.

For instance, a word like “hate” would be rearranged to “haet” and graphemes “x” were rewritten as “ks”.

In more advanced approaches, CRFs have been shown to offer several advantages over HMMs by relaxing the strong independence assumptions described in Section 2.2.3. The most important point about CRFs are the feature functions. When used in a task like G2P conversion, each feature function takes as input a grapheme that has to be converted to its phonemic representation, and several types of information about the grapheme, such as the word it belongs to, the position of the grapheme in the word, the POS of the word, and the surrounding graphemes. The features and their respective weights are then used to find the optimal sequence of phonemes [Sutton and McCallum, 2006].

Lastly, joint n-gram models use substring pairs of graphemes and phonemes so that the information about both part can be exploited [Jiampojamarn, 2011]. Joint n-gram models can have orders ranging from 1 to 7. A simple search through the pairs would give the most probable sequence using the Viterbi algorithm [Taylor, 2009].

The output of G2Ps may be recomputed by post-lexical processors to further improve the pronunciation or to model variants. An explanation of post-lexical processing can be found in the following section.

2.3.2 Post-lexical processing

Instead of feeding the phonemic transcription from a G2P directly into a speech synthesizer, some TTS systems perform an additional step known as post-lexical processing. The goal of this step is to improve pronunciations by predicting coarticulation phenomena or pronunciation variants. Additionally it might transform pronunciations to reflect a given target expressivity by adapting the results of the G2P to the pronunciation style of an individual speaker, a group of speakers, or a certain accent.

As opposed to the G2P which only considers the graphemic contexts, post-lexical processors take both graphemic and phonemic contexts into account to further modify the generated phonemic transcriptions. Most of the machine learning techniques used for post-lexical processing are the same as the ones used for G2P conversion. However, the difference arises from the perspective that G2P has only one goal, i.e., to generate the phonemic transcription, while as mentioned earlier, post-lexical processing has several goals. One of the main contributions of our work is in the area of pronunciation adaptation which is also a post-lexical processing. Therefore, it is important to have a good background in the previous work conducted in this field.

2.3.2.1 Related work

There has been considerable effort in the past to understand and deal with pronunciation variation for both TTS and ASR. Most of the early work in this area relies on using predefined or automatically extracted phonological rules to derive alternative pronunciations [Tajchman et al., 1995, Giachin et al., 1990, Oshika et al., 1975], whereas, in the recent literature, various machine learning and statistical approaches have been proposed. Among them, decision trees and neural networks have received considerable attention [Vazirnezhad et al., 2009, Chen and Hasegawa-Johnson, 2004, Fosler-Lussier et al., 1999, Riley et al., 1999, Miller, 1998]. Other approaches such as random forests [Dilts, 2013], HMMs [Prahallad et al., 2006], and CRFs [Karanasou et al., 2013] have also been studied to derive alternative pronunciations. Each of those mentioned studies tackles the problem in a different way.

[Miller, 1998] studied post-lexical phonology, which can lead to interspeaker variations, using neural networks. In order to predict the post-lexical pronunciation, the canonical pronunciations were encoded along with prosodic information and then fed into a neural network. Context of the phonemes was also fed into the neural network by using a window of three phonemes (one from left and one from right). For each target phoneme, the neural network outputs a post-lexical phoneme, a silence for deletions, or a diacritic to indicate minor changes in the pronunciation of the canonical phoneme. The system performed best when the variants of a phoneme were few, such as in the case of word-initial vowel glottalization[1], where only two variants are available. However the system struggled when the number of variants were more than two, such as for the phoneme /t/. The main limitation of this study lies in the fact that the author used read speech for training the network, thus the impact of the proposed method on spontaneous speech remains unclear.

In another study, [Vazirnezhad et al., 2009] followed a different technique which consisted of a decision tree and a contextual rule generator to produce post-lexical pronunciations. Given an input phoneme string, the decision tree was used to predict the phonemes that needed to be changed, while the contextual rules were used to generate the post-lexical changes such as substitution, insertion or deletion of phonemes that were susceptible to change. The features that the authors included in the decision tree were the rate of speech, word unigram probabilities, syllable location and stress. These features are known to have a strong impact on pronunciation in spontaneous speech. Their work showed that using extra features in addition to the sole phonemes is useful for post-lexical processing. However, the proposed method uses a very limited number of features.

[1] Glottalization is the closure of vocal folds during the articulation of a sound.

[Jande, 2003] studied how speech rate and speaking style affect pronunciation and phone reduction in Swedish. The goal was to capture the general pronunciation variations in spontaneous speech rather than individual or dialect-related variations. Several rules were extracted to conduct the analysis, e.g., haplology, assimilation, and elision of important phonemes like /r/ and /h/. Common words matching these rules were then extracted, and their realized post-lexical forms were compared to non-reduced (canonical) ones by presenting both to listeners in test utterances. The result of this evaluation showed that the canonical forms were considered more natural when the speech rate is low, while reduced forms were judged as the most natural ones when the speaking rate was medium or high. What this study lacks is a broader evaluation of the speech samples in order to see how the proposed method improves spontaneousness in Swedish.

Lastly, [Bennett and Black, 2003] tried to mimic an individual speaker by predicting a speaker's choice of pronunciation between reduced and canonical forms of English words. The focus was on the most common words which are known to have multiple pronunciations. Words like "the", "a", "to", and "for" are known to have different pronunciations based on their context in English mainly because of vowel reduction. For instance, "to", which has a canonical form of /tu/, is sometimes pronounced with a reduced form as /tə/, likewise "a" and "for" which are canonically pronounced as /eɪ/ and /fɔr/, are reduced and pronounced as /ə/ and /fɝ/ respectively. The word "the" is probably one of the best known cases when it comes to vowel reduction. It is mostly pronounced as /ðə/ when it is followed by a consonant-initial word and is pronounced as /ði/ when followed by a vowel-initial word. However, exceptions occur in many situations. For instance it can be pronounced as /ði/ even before consonant-initial words, e.g., when uttering the phrase "the car" in a context where the mentioned car is meant to be unique in some sense. The evaluation results on the words showed that the prediction for some of the words like "for" and "a" were mostly correct, while the method failed to capture the variations in pronunciation of the other two words, especially "to". Although this study provides some useful insights into pronunciation variation, it remains very limited since the number of considered words is quite small.

2.3.2.2 Useful features

Features that are considered to be the most important to model pronunciation variation can be divided into linguistic-phonological, acoustic-prosodic, and articulatory categories. Linguistic-phonological features can be derived directly from textual data and can be derived from the phoneme, syllable, word and utterance level information. The importance of such features is well known and has already been studied [Vazirnezhad et al., 2009, Bell et al., 2009, 2003]. On the other side, acoustic-prosodic features can

be extracted from speech signals. Usual features are F0, energy, duration, speaking rate, etc. [Bell et al., 2009, 2003, Bates and Ostendorf, 2002]. In the case of TTS, as the speech signal is not present, these features cannot be easily obtained. Instead, they should be predicted from the input text for the desired speech style. In addition to these two feature types, the benefits of using articulatory features have also been examined for this task [Kirchhoff, 1999].

The majority of the mentioned studies for pronunciation variation has either been applied in the context of ASR or has studied variations in a very limited way in the context of TTS. Moreover, most of them have utilized a limited number of features. In [Dilts, 2013], a deep study on the combination of linguistic and prosodic features using random forests is conducted; however, like some other mentioned studies, the approach mainly remains limited as it only focuses on phonetic reductions. In Chapter 4, we will present our contribution on the prediction of pronunciation variants for TTS. What makes our approach different from related studies in this area lies in: (i) exploiting a much larger set of feature types including linguistic, articulatory and prosodic, (ii) predicting all sorts of pronunciation variants as opposed to limiting the study to certain phenomena, (iii) conducting a perceptual test which evaluates both naturalness and spontaneousness of speech samples generated using the proposed method.

2.4 Disfluency modeling

Current TTS systems have already reached a high level of naturalness thanks to the effective use of unit selection and the advancements in SPSS [Adell et al., 2012]. However, the majority of current TTS systems has focused on generating speech that is closer to the way we *read* than the way we *talk*. Despite the efforts that have been put on making TTS systems more expressive, the results are still far from being perfect. In the previous section, we discussed some of the main works in this direction by concentrating on pronunciation modeling. However one cannot expect to have a completely expressive synthetic speech by only considering pronunciation variants. One possible way of further improving TTS systems is to integrate disfluencies. As we already discussed in Section 1.4.2, disfluencies are one of the main characteristics of spontaneous speech. Therefore, we believe that integrating them in TTS will lead to more expressive synthetic speech.

Despite many work on detecting disfluencies in order to improve the accuracy of ASR systems [Liu et al., 2006, Kaushik et al., 2010], the number of studies on generating disfluencies in TTS is very limited. According to [Adell et al., 2008], there are two main reasons for the lack of studies in this area. First, most of the time, the speech database

of unit selection systems does not contain any disfluencies. Second, text analysis models (e.g., POS tagging) expect sentences to have a correct structure, thus, making it difficult for traditional models to perform well on disfluent speech.

Despite these limitations, we can find a few studies on the generation of different types of disfluencies in the literature. In [Sundaram and Narayanan, 2003], the authors studied the automatic insertion of filled pauses ("uh" and "um") in spontaneous speech. The method they proposed works in the following way: in the offline mode, the input words are tagged with their corresponding POS and the most common words that are mostly likely to precede a filled pause are determined using a language model for both "uh" and "um". Then, for each phrase preceding a filled pause, a Finite State Acceptor (FSA) is created. Finally a complete FSA network is created by combination of all FSAs for each occurrence of "uh" and "um". During the online mode, the input sentence is tagged with POS and searched for words that might precede a filled pause. If there is a match, the phrase before the word is extracted and checked against the FSA networks. Lastly if a network accepts the extracted phrase, the corresponding filled pause, i.e., either "uh" or "um" is inserted. Although the proposed algorithm seems to work in certain cases, it also inserts filled pauses in positions where they should not be present. A major problem with this study is that it does not include a true subjective evaluation, thus the quality of the algorithm remains unknown. In addition, POS was the sole feature used in building the FSAs.

In [Adell et al., 2007], the authors concentrated on the place where filled pauses must be placed in the text. The proposed algorithm works by combining language models and decision trees. The tree classifies each word in the text to decide whether it should be followed by a filled pause or not. The decision tree was constructed using several features including POS, language model probabilities, word position in text. The results showed that the proposed system can predict the position of filled pauses with a precision of 96%. Moreover, perceptual tests seem to support this result as most sentences with disfluencies were identified correctly. The main drawback of this study is that it solely concentrates on where the disfluency should be inserted and not on the actual generation of disfluencies. In another similar study [Dall et al., 2014], the authors employed several approaches for automatically predicting interruption point (IP) of filled pauses. Their approach included an n-gram language model, an RNN language model, an interpolated n-gram + RNN language model, a support vector machine, a decision tree, and finally a random insertion. The features they used for the support vector machine and the decision tree include syllable count of words following the IP, phrase boundary associated with the IP, clause boundary associated with the IP, 4-gram log-probability for sentences with "uh", POS associated with words following the IP. Perceptual tests on sentences with and without inserted filled pauses were conducted.

The results showed that the best performing systems were the interpolation of RNN + n-gram language models.

Finally, a more detailed work on disfluencies was conducted in [Andersson et al., 2010] where the prediction of the place and the exact type of lexical and non-lexical fillers were studied. The proposed method included the use of a language model and of the Viterbi algorithm. The language model was used to insert candidate disfluencies and the Viterbi selected the combination of insertions along a sentence that led to the highest overall probability. The perceptual test showed that the synthesized speech with predicted disfluencies were more conversational, with no loss of naturalness on average.

Most of these mentioned studies have several limitations and are mostly experimental. First, most of them concentrate on few types of disfluencies (mostly filled pauses). Second, only few of these studies tackle the problem using statistical approaches which have been shown to perform well for NLP tasks. Lastly, the results are rather poorly evaluated, most of the studies lacking true subjective tests on utterances with predicted disfluencies. In Chapter 5, we propose a complete protocol for generating disfluencies including lexical and non-lexical fillers and repetitions using CRFs and language models. Moreover, the proposed approach is evaluated subjectively in order to test its adequacy.

2.5 Conclusion

The objective of this chapter was to provide a review of literature in the fields of pronunciation and disfluency modeling in the context of TTS. Thus, we first gave a general overview of TTS systems. We briefly mentioned the different components of the front-end and the different techniques used in the back-end. Different machine learning approaches that are most commonly used in TTS were discussed as well. Next, we studied different tasks in pronunciation modeling including grapheme-to-phoneme conversion and post-lexical processing. Finally, in the last section, we provided some insights into modeling disfluencies and surveyed some studies on the prediction of disfluencies.

The focus of the next three chapters will be on explaining the data and our contributions on generating pronunciation variants and disfluencies. Both of these works are applied on a spontaneous speech corpus. Spontaneous speech, as stated earlier, is one of the main causes of variability in speech. In addition, it is highly related to the other aspects of expressivity, i.e., accents and emotions. Hence, these characteristics make spontaneous speech more interesting to work with in the context of pronunciation variants and speech disfluencies.

Chapter 3

Data and evaluation methodology

In this chapter, we will describe the data and the evaluation methodology that is going to be used in the rest of this thesis for generating pronunciation variants and speech disfluencies. Firstly, the Buckeye conversational English corpus is introduced in Section 3.1 which is the main source of data for both tasks. Section 3.2 presents a statistical analysis of the corpus. Next, in Section 3.3, we go through the features which are extracted from the corpus. Finally, Section 3.4 discusses different evaluation methodologies including objective and subjective ones.

3.1 The Buckeye corpus

In this thesis we use the Buckeye corpus of English conversational speech. This corpus consists of 307,000 words collected through interviews with 40 speakers from central Ohio, USA [Pitt et al., 2005]. The proportions of gender and age of the speakers in

the corpus are equally balanced. Each interview lasts about 1 hour making a total of 40 hours of recorded speech. The interviews are conducted through question answering where an interviewer asks questions of general topics to which the speakers have to answer based on their own opinion. The corpus has been orthographically and phonemically transcribed. The phonemic transcription includes the standard pronunciation (*canonical phonemes*) and the one effectively uttered by the speaker (*realized phonemes*). This transcription has been automatically generated, manually checked and corrected. The speech signal and the corresponding start and end times of each phoneme are also included. In addition to lexical items, non speech sounds (silent pauses, filled pauses, cutoffs, lengthenings, etc.) have been identified.

In this work, 20 speakers from the Buckeye corpus are considered. They have been randomly selected under the constraint to maintain the age and gender proportions, in order to avoid having data from only a specific age or gender group. Among the selected speakers, the average number of realized phonemes per speaker is 22,789, and the average number of words is 7,354.

3.2 Statistical analysis of the Buckeye corpus

In Chapter 2, features that are mostly used in the literature to model pronunciation variation and disfluencies were briefly mentioned. In this section we will provide a detailed statistical analysis of such features in the Buckeye corpus. The objective here is to test which types of features mainly impact pronunciation variation and disfluencies in the corpus. The first three sections of this analysis are dedicated to phonemic variations on the word, syllable and phoneme levels, while the last section provides disfluency related statistics.

3.2.1 Word-level pronunciation variations

The total number of words in the selected 20 speakers is around 150,000 words. Among those words, 57% have realized pronunciations different from their canonical forms. This mismatch is mainly due to phonemic variations, i.e., substitution, deletion or insertion of phonemes. The measure which is used to calculate the percentage of this mismatch is called Phoneme Error Rate (PER). The baseline PER between canonical and realized phonemes in the analyzed portion of the data is 28.3%. This strong difference between phonemes suggests that generating pronunciation variants is not a trivial task. In Figure 3.1 we illustrate the average number of different realizations per word in the corpus[1]. As it can be seen, frequent words are by far much more variable than

[1] Frequent words are identified by extracting the most commonly occurring 1000 words in the corpus.

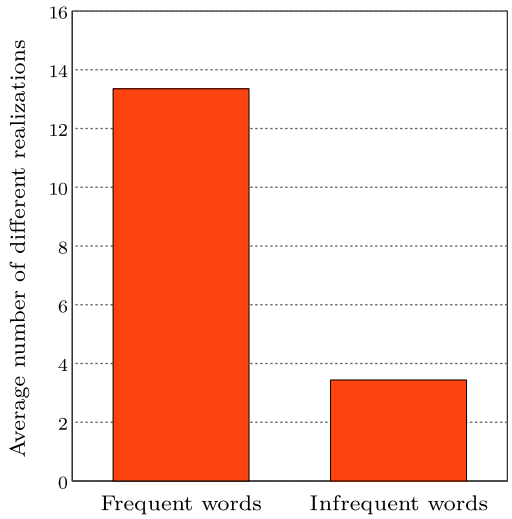


Figure 3.1: Average number of different realizations per word in frequent and infrequent words.

infrequent words. This difference probably occurs because frequent words are usually pronounced faster since they can be easily inferred by speakers. This phenomenon has also been studied in the literature. Fosler-Lussier and Morgan [1998] argue that the probability of a word being spoken in a canonical fashion decreases as the speaking rate increases. The speaking rate feature of words in the Buckeye corpus, as shown in Figure 3.2, seems to confirm this argument as words with an overall faster speaking rate are more variable. Next, the effect of word position in utterance in relation to the amount of observed variations is analyzed. To make the analysis easier to follow, word positions are given relatively to the length of the utterance and discretized in slots of 10%. This is to reduce the difference between short and long utterances. Based on this analysis, Figure 3.3 shows that the PER is rather stable from the beginning and gradually increases towards the ends. Interestingly, at the very end of utterances, PER seems to decline and the lowest PER is observed. These results seem to match with what has been observed in the literature, as Bell et al. [2003] showed that words are more likely to be pronounced with the canonical form in utterance initial or utterance final positions, while more likely to have less canonical forms in mid-utterance positions.

3.2.2 Syllable-level pronunciation variations

Syllables also provide useful information about pronunciation variation [Vazirnezhad et al., 2009, Adda-Decker et al., 2005]. For this purpose, we first analyze syllable position inside a word. To perform this analysis, syllable positions are categorized into *initial*, *middle* and *final*. The analysis does not include monosyllabic words since the

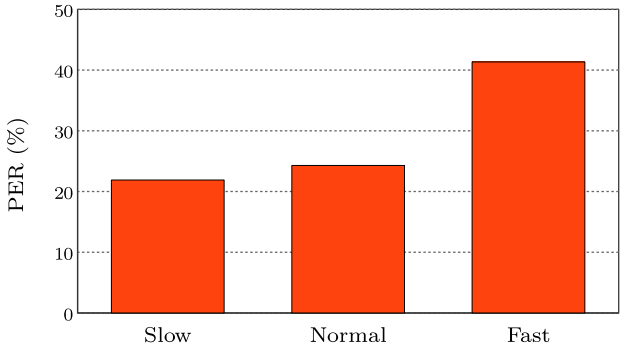


Figure 3.2: PER (%) between canonical and realized phonemes in words with fast, normal and slow speaking rates.

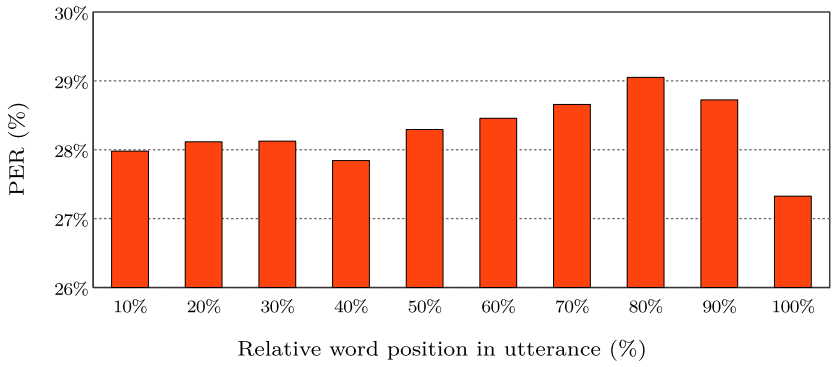


Figure 3.3: PER (%) in words with respect to their relative position in utterance.

sole syllable they contain can be treated as an initial, middle or final syllable at the same time. Therefore, only words with at least two syllables are analyzed. In disyllabic words, first syllables are considered as initial syllables and second syllables as finals, whereas in those with a higher number of syllables, all syllables between the initial and final ones are simply considered as middle syllables. Figure 3.4 shows PER based on these three positions. We can clearly see that the PER is slightly higher in middle and final syllables than in initial ones.

Syllable lexical stress is another factor which is known to impact pronunciation variation [Vazirnezhad et al., 2009, Greenberg, 1999]. In the Buckeye corpus, as shown in Figure 3.5, unstressed syllables have the highest pronunciation variation ratio while syllables with primary stress have the highest matching ratio. This shows that the syllable stress has a significant effect on pronunciation. Finally, variations can also be analyzed according to the parts in a syllable. Figure 3.6 shows that in the Buckeye corpus, onsets have the lowest PER while nucleus and codas have the highest.

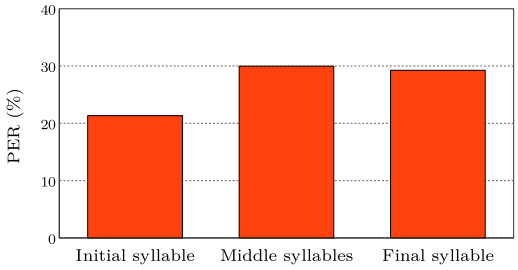


Figure 3.4: PER (%) based on syllable position.

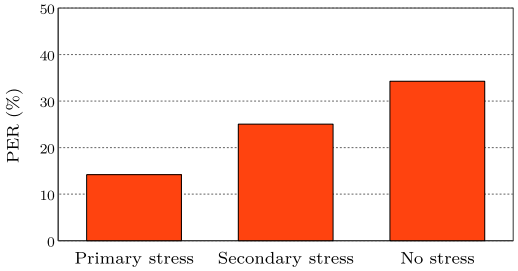


Figure 3.5: PER (%) based on syllable lexical stress.

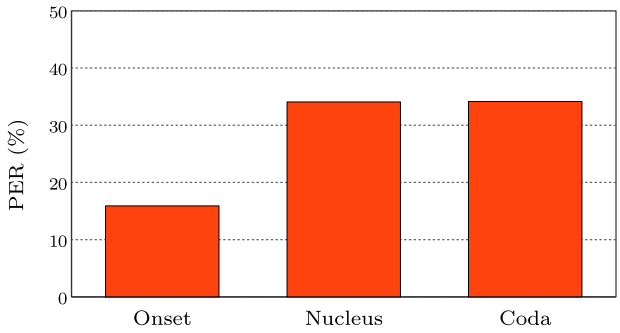


Figure 3.6: PER (%) based on syllable part.

3.2.3 Phoneme-level pronunciation variations

Another type of information that might affect pronunciation variation is the phoneme-level information. In the Buckeye corpus, around 25% of the phonemes have different realizations than their canonical phonemes. Overall, vowels have a higher PER than consonants with 35% and 25% respectively, as shown in Figure 3.7. A phenomenon like vowel reduction has probably a big role in this difference. Among vowels in the Buckeye corpus, /ʌ/, /æ/, and /aɪ/ have the highest number of different realized pronunciations.

One of the factors that lead to pronunciation variation is the position of the phoneme in the syllable. In Figure 3.8, it can be clearly seen that the phonemes at the end of syllables are much more variable than at initial positions. Thus, the results also suggest that PER increases with the size of the syllable.

Lastly, among articulatory features, place and manner of articulation seem to be the most interesting features. Concerning the place of articulation, alveolar and dental phonemes have the highest percentage of variations, while plosives and nasals are the most varied phoneme types when it comes to the manner.

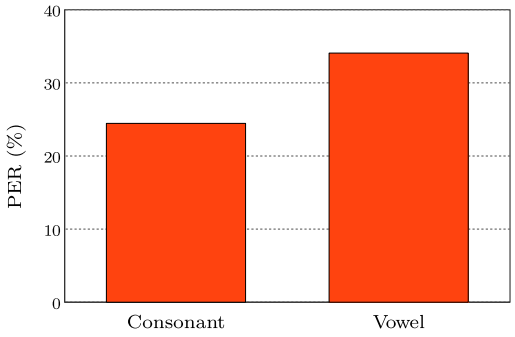


Figure 3.7: PER (%) in consonant and vowels.

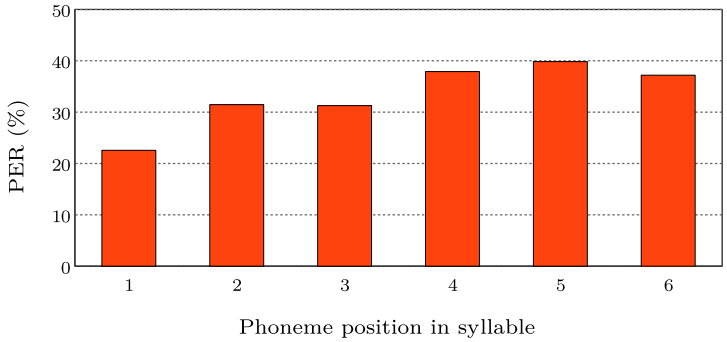


Figure 3.8: PER (%) with respect to position of phonemes in syllable.

3.2.4 Speech disfluencies

In this thesis, three types of disfluencies are recognized: pauses, repetitions, and revisions (*cf.* Section 1.4.2). Detecting and analyzing pauses and repetitions is fairly straightforward, whereas revisions need to be manually labeled by an annotator. Due to the lack of time and human resources, only data from 12 speakers (out of 20) has been annotated for revisions. This section presents the analysis of pauses, repetitions and annotated revisions.

First, in Figure 3.9 the frequency of different pause types is provided. Clearly, silences are dominating other pause types. This is most probably due to the fact that silences are not only used independently but also after other pauses, repetitions and revisions. It is worth mentioning that "you know", "I mean" and "well" can be used both as disfluencies or as normal words. In order to find which occurrences are disfluencies, they have been manually checked. In Chapter 5 details of the process of cleaning the corpus which also included disambiguating such types of disfluencies are given.

Concerning repetitions, they can be analyzed in terms of the number of repeated words. For this analysis, one-word repetitions such as "*I I* will go now", two-word

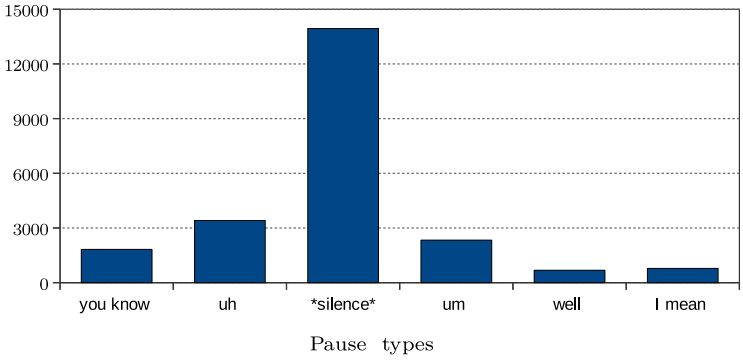


Figure 3.9: Histogram of pauses.

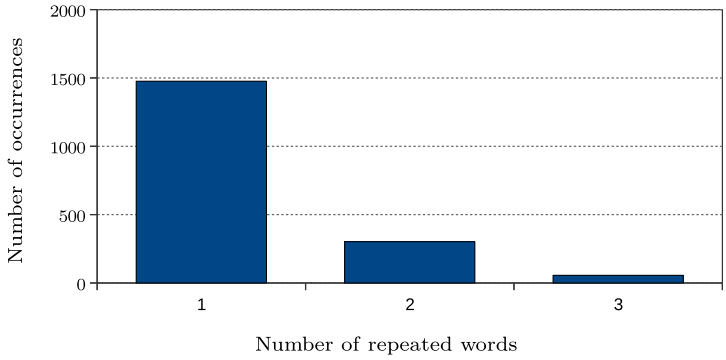


Figure 3.10: Histogram of repetitions according to the number of repeated words.

repetitions like "*I will I will* go now", and three-word repetitions as "*I will go I will go* now" are considered. As shown in Figure 3.10, one-word repetitions outnumber the two other types.

Next, we analyzed the proportion of annotated revisions based on the number of words in their reparandum and repair regions. Results provided in Figure 3.11 show that in the majority of the times, revisions have one or two words in their repair and reparandum regions.

Lastly, we analyzed the relative position of disfluencies in an utterance. As illustrated in Figure 3.12, the majority of the pauses seems to be located at the very beginning of the utterance and the rest equally spread out to the further positions. Repetitions and revisions on the other hand have a similar trend. Their number seems to gradually decrease with respect to their position.

From what can be observed in the above analysis, information from different levels can affect pronunciation variation and disfluencies. Hence, one has to take them into

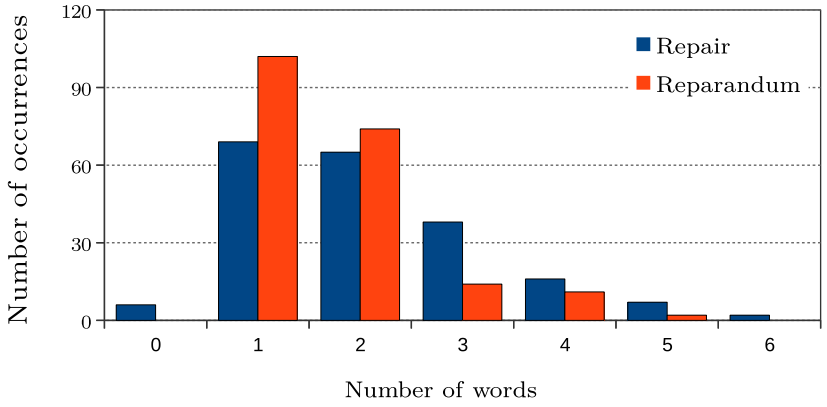


Figure 3.11: Histogram of number of words in repair and revision regions of revisions.

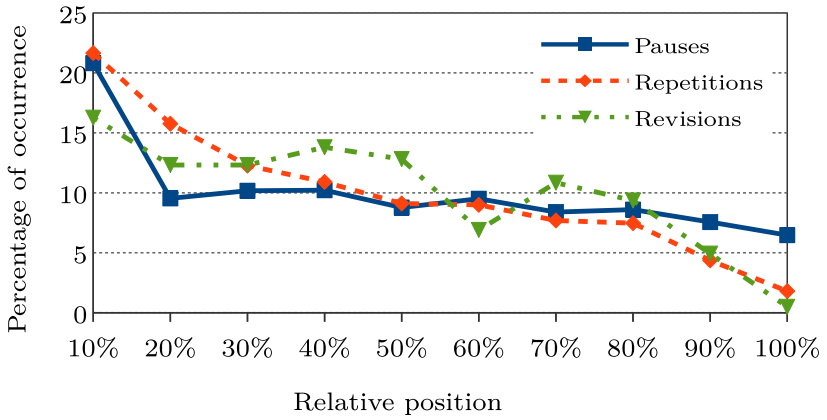


Figure 3.12: Position of disfluencies in the utterance.

account when building pronunciation or disfluency models. In the next section, the list of features that are considered in this thesis is discussed.

3.3 Derived features

Based on the analysis in the last section, we decided that the corpus has to be enriched with a larger number of features in addition to the ones originally provided with the corpus. The set of newly added features is mostly determined based on the statistical analysis of the corpus and also on choices in the literature. The features presented here are mostly related to our pronunciation variants work, while the ones for the disfluency work are introduced in Chapter5. The presented features are grouped into three categories: linguistic-phonological (shortened to *linguistic* in the rest of the document),

Table 3.1: List of linguistic features added to the Buckeye corpus.

Feature	Values
canonical phoneme	40 possible phonemes
phoneme position in syllable	integer
reverse phoneme position in syllable	integer
syllable lexical stress	no stress, primary, secondary
syllable part	onset, nucleus, coda
syllable location	(initial, middle, final) syllable
word	word
word frequency in English	high, medium, low
stem frequency in English	high, medium, low
stop word	true, false
word boundary	beginning, middle, end
grapheme	grapheme
syllable type	open, closed
number of syllables of the word	integer
word frequency in the interview	high, medium, low
stem frequency in the interview	high, medium, low
word count in interview	integer
word position	integer
reverse word position	integer
word length	integer
POS	noun, verb, adjective, etc.
utterance position	integer
reverse utterance position	integer

articulatory, and acoustic-prosodic (*prosodic* in the rest of the document) features. The complete list of features is detailed in this section.

3.3.1 Linguistic features

As we discussed briefly in Chapter 2, the influence of linguistic information on pronunciation variation is well known and has been investigated extensively before. In this work, as shown in Table 3.1, 23 linguistic features have been added to the corpus. The POS tags have been extracted from the corpus itself, while stop words have been identified using a list of 500 words in English, and the word frequencies have been retrieved using Google n-grams. All frequency based features have been binned into three categories with equal probability masses (low/medium/high).

Table 3.2: List of articulatory features added to the Buckeye corpus.

Feature	Values
phoneme type	vowel, consonant
manner	nasal, plosive, fricative, etc.
place	bilabial, labiodental, etc.
shape	front, near front, etc.
aperture	close, near close, etc.
voiced	true, false
rounded	true, false
affricate	true, false
doubled	true, false

3.3.2 Articulatory features

Articulatory features have already been studied and used successfully in the context of ASR [Ghosh and Narayanan, 2011, Kirchhoff, 1999]. The idea of including these features in this study is to see if they have the same positive impact on TTS. Hence, 9 articulatory features have been derived for each phoneme as shown in Table 3.2. These features are either categorical (e.g., manner and place of articulation, shape, etc.) or boolean (e.g., voiced, rounded, etc.). It is important to mention that these articulatory features all together determine the canonical phoneme itself [International Phonetic Association, 1999]. Thus, it might be considered as a redundant information. However, the aim here is to more precisely define the actual canonical phoneme and investigate which of its articulatory features lead to variation during spontaneous speech.

3.3.3 Prosodic features

Several acoustic and prosodic features like F0, energy, tone, speech rate, etc. have been considered in this work. The complete list is presented in Table 3.3. In TTS, these features have to be predicted from the textual input as there is no signal out of which they could be extracted. However, this task is still a research problem and is out of our scope. As a consequence, the acoustic and prosodic features have been directly extracted from the signals uttered by each speaker. This strategy simulates a perfect prosody modeling, leading to optimistic adaptation results. However, the idea here is to test how the existence of a perfect prosody predictor can help in tackling the problem of pronunciation adaptation. To remain realistic, all extracted prosodic features have been simplified and coarsely approximated. For instance, syllable F0 shapes have been discretized into the categories *increasing*, *flat*, *decreasing*, syllable energy into *low*, *medium*, *high*, etc. Thus, the difference between using this approach and a prosody

Table 3.3: List of prosodic features added to the Buckeye corpus.

Feature	Values
syllable energy	high, medium, low
syllable F0 shape	increasing, flat, decreasing
pause per syllable	high, medium, low
phone tone	1 .. 5
distance to next and previous pause	close, mid-far, far
distance to next and previous hesitation	close, mid-far, far
syllable tone	1 .. 5
speech rate	high, medium, low

predictor has been minimized.

3.4 Evaluation methodology

There are mainly two types of evaluation metrics used in speech synthesis: objective and subjective. In objective metrics, usually automatic tools are used to measure the output of a specific TTS component, for instance, pronunciations resulting from a pronunciation model. In subjective metrics, generally, human subjects directly evaluate the achieved results, usually by scoring them, for example, for measuring the generated synthetic speech in terms of naturalness and intelligibility. Each of the two kinds of metrics has its own pros and cons. In the following sections, details on each one is given in the context of pronunciation modeling and speech disfluencies.

3.4.1 Objective evaluations

Several types of objective evaluation metrics can be used to evaluate pronunciation and disfluency modeling [Adda-Decker et al., 1999, Höge et al., 2008, Adell et al., 2007]. The most common ones are: Phoneme Error Rate (PER), perplexity, recall, precision and F-measure.

In the context of pronunciation modeling, PER relates to the minimum number of edits needed to transform the hypothesized sequence of phonemes under examination to the realized reference sequence of phonemes. Considering a set of U utterances, a hypothesis is defined as a set of predicted phoneme sequences $\mathcal{H} = \{\mathbf{h}_i \mid 1 \leq i \leq U\}$, and the reference as a set of realized phoneme sequences $\mathcal{R} = \{\mathbf{r}_i \mid 1 \leq i \leq U\}$, i.e., one $\mathbf{h}_i$ and $\mathbf{r}_i$ per utterance. PER of $\mathcal{H}$ over $\mathcal{R}$ can then be computed by, first, aligning the two sequences ($\mathbf{h}_i$ and $\mathbf{r}_i$) and dividing the sum of the S substitutions, D deletions, and

I insertions in $\mathcal{H}$ by the total number N of phonemes in $\mathcal{R}$:

$$PER(\mathcal{H}, \mathcal{R}) = \frac{S + D + I}{N} \ . \tag{3.1}$$

Hence, PER is given as a percentage, and the lower, the better. For example, for a given reference (/ae/, /n/, /d/), and an hypothesis (/ae/, /n/), the PER of $\mathcal{H}$ over $\mathcal{R}$ is 33.3%. PER is a standard measure as it is straightforward to compute and it enables comparisons between heterogeneous methods. Yet, PERs have a major drawback since they ignore the confidence of the model.

Perplexity is computed based on the average uncertainty assigned by a probability distribution (e.g., a trained CRF model) to each phoneme sequence in the reference $\mathcal{R}$. Consequently, a low perplexity means that the model is a good predictor for the sequence. Mathematically, this is formulated as follows:

$$Perplexity(\mathcal{R}) = 2^{-\frac{1}{N}\sum_{i=1}^{U} \log_2 \Pr(\mathbf{r}_i)} \ , \tag{3.2}$$

where $\Pr(\mathbf{r}_i)$ is the probability given by the trained model to the phoneme sequence $\mathbf{r}_i$. According to Equation 3.2, the best possible perplexity value is 1 and the lower the better.

As opposed to PER, perplexity enables to study the quality of a model beyond the sole best hypothesis it returns. It tells one how far on average a model is from finding the correct phoneme even if it cannot always predict it. Moreover, perplexity can be easily adapted to other tasks, e.g., disfluency generation.

As for recall, precision, and F-measure, they are mostly used in information retrieval tasks and can also be adapted to other tasks like pronunciation variants and disfluency generation. For example, in the frame of disfluency generation, considering that $\mathcal{H}$ as a set of predicted interruption points (IPs), and $\mathcal{R}$ as the set of reference IPs, each at a specific position in the utterance, recall measures the proportion of correctly predicted IPs, i.e., in their exact position, divided by the total number of IPs in the reference as formulated in Equation 3.3, while precision, as shown in Equation 3.4, measures the proportion of correctly predicted IPs divided by the total number of correctly and incorrectly predicted IPs. Lastly, F-measure considers both recall and precision in one equation as given in Equation 3.5. The highest score (also the best) that these three metrics can reach is 1 and the lowest is 0.

$$recall(\mathcal{H}, \mathcal{R}) = \frac{|\mathcal{H} \cap \mathcal{R}|}{|\mathcal{R}|} \tag{3.3}$$

$$precision(\mathcal{H}, \mathcal{R}) = \frac{|\mathcal{H} \cap \mathcal{R}|}{|\mathcal{H}|} \tag{3.4}$$

$$F{-}measure(\mathcal{H}, \mathcal{R}) = 2 \times \frac{precision \times recall}{precision + recall} \tag{3.5}$$

In this thesis, PER and perplexity metrics are adopted as performance measures to evaluate generated pronunciation variants, whereas for evaluating generated disfluencies, perplexity, recall, precision and F-measure are used.

3.4.2 Subjective evaluations

There are several subjective measures for evaluating speech including AB preference test, Mean Opinion Score (MOS), MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA). Each of these methods uses a different approach for evaluating speech samples and ranking the systems that are used to generate the speech samples.

In AB tests, two speech samples coming from two different systems are presented to the subjects and they are asked to evaluate them based on some criterion such as naturalness or intelligibility. At each step, usually two or three choices are provided which include *system A*, *system B*, or *indifferent* and the subject has to state his preference on one of the choices. The advantage of AB tests lies in its simplicity and the fact that it requires minimal effort for the subject to complete the test. A slightly different version of this test is called ABX. In ABX, two speech samples namely A and B and a reference speech sample called X are presented to the subject. The subject is then asked to decide which one of A or B is the closest to X [You, 2010]. In both versions, the results can be aggregated in terms of percentage by counting the choices for each system from all the subjects.

Secondly, in the MOS test, listeners are asked to rate the quality of several speech samples coming from different tested systems one at a time and on a five-point scale ranging from 1 (unsatisfactory) to 5 (excellent). The score of each system is then computed by averaging opinion scores for each speech sample from the subjects. Another method which is very similar to MOS is Degradation Mean Opinion Score (DMOS). Contrary to MOS, DMOS tries to measure the degradation in the quality of speech. This is done by presenting the subjects with a reference speech and the tested speech and asking them to rate the degradation.

Finally, in the MUSHRA test, the subjects are presented with a reference, several tested speech samples from different systems, a hidden unmodified reference and several anchors which are basically modified versions of the original sample passed through low pass filters [Pulkki and Karjalainen, 2015]. The anchors are useful since they help the subjects in not rating samples which have minor artefacts with a very low score. As the subjects are evaluating several systems at a time, MUSHRA has the advantage of requiring less participants than MOS to obtain statistically significant results. However,

MUSHRA is considered to be difficult for the subjects since they have to listen and score several speech samples simultaneously.

In general, subjective measures offer much more reliable results over objective ones since they directly evaluate the quality of the synthetic speech which is not possible with objective measures. However, for a reliable evaluation, large numbers of subjects and speech samples are needed, which is expensive and time consuming. Therefore, when the number of systems to be tested is large and the number of subjects is low, it is a good idea to choose a simple test like AB preference test.

Another issue with perceptual tests is that the samples are usually chosen randomly, which leads to the selection of very similar ones, making it difficult for subjects to identify the difference between them. In order to overcome this problem, it has been suggested that synthesizing several thousand utterances from different systems and choosing the most different samples yields better results than randomly choosing them [Chevelu et al., 2015].

3.5 Conclusion

In the beginning of the chapter, we gave a description of the Buckeye corpus and then a detailed statistical analysis was provided with respect to pronunciation variation and disfluencies. Next, the features that were added to the Buckeye corpus were described including linguistic, articulatory and acoustic-prosodic features. Finally, we went through some of the common objective measures used in speech and machine learning tasks as well as the subjective ones. In the remainder of this thesis, our contributions on generating pronunciation variants and speech disfluencies are presented in Chapter 4 and Chapter 5 respectively.

Chapter 4

Generation of pronunciation variants

The two main objectives of this thesis are to generate pronunciation variants and speech disfluencies in the frame of speech synthesis. In this chapter, we present our contributions on the pronunciation variants side which play a critical role in making synthetic speech more expressive. Our aim is to provide a method which is able to automatically learn such variants. A possible way to do this is to adapt standard pro-

nunciations to a speaking style with much variabilities, for instance, spontaneous speech. Therefore, we propose a method which automatically learns phonemic variants of spontaneous speech from the Buckeye corpus, and applies them on standard pronunciations to generate alternative ones. Developing such a method requires several important aspects to be considered. For instance, which machine learning approach should be used and what types of features are useful for this task? Then, one should also know if working on the pronunciation side is enough to generate expressive speech without any prosodic or linguistic changes. Lastly, we have to decide on how to properly evaluate the results. In the rest of this chapter, alongside the description of the proposed method in Section 4.1, we will try to answer these points in Sections 4.2 to 4.5.

In the last part of this chapter (Section 4.6), we also show that the proposed method can be extended to other similar tasks. Precisely we show that the method can be used to solve the problem of inconsistency between the phoneme sequences generated by G2P converters during synthesis and those from the system's speech corpus. This inconsistency usually leads to poor quality synthetic speech signals. To solve this problem, we use the same adaptation approach to adapt automatically generated pronunciations to the style of the corpus, which should eventually improve the quality of the synthesized speech. This latter work is referred to as corpus-specific adaptation and is conducted on a French corpus.

In the next section, the overall methodology of the proposed approach for pronunciation adaptation is provided.

4.1 Overall methodology

The fundamental idea behind pronunciation adaptation is to predict the sequence of realized spontaneous phonemes from an input sequence of canonical phonemes. We choose to model this task as a labeling problem, and since CRFs have been previously shown to perform well on sequential labeling tasks, the method that we propose more precisely relies on phoneme-to-phoneme CRFs. The overall methodology to develop this solution has been to study all the factors that may influence the performance of the adaptation. The objective of this section is to describe the phoneme-to-phoneme labeling task and to identify these influential factors and their position within the task.

As presented in Figure 4.1, the labeling task consists in mapping canonical phonemes c_i to realized phonemes p_i. As we are in the domain of speech synthesis, these phonemes represent a whole utterance. To make this task easier, canonical phonemes come along features $\{f_i^1, \ldots, f_i^n\}$ which may represent various aspects of the phoneme such as their position in the utterance, the word they are included in, etc. as already described in Section 3.3. Based on these descriptions, the idea is to train a CRF model and to

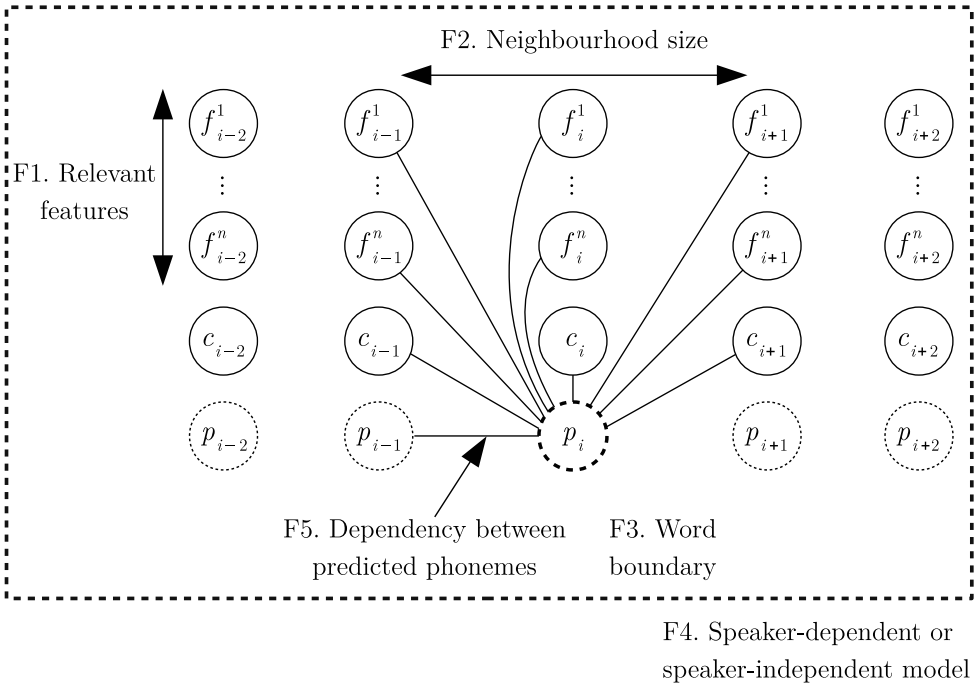


Figure 4.1: Overview of the CRF training and influential factors.

apply it on new data at runtime, i.e., during synthesis.

In the experiments, the goal is to build a CRF model with the best possible labeling performance. In practice, this performance is mainly measured by PERs of the generated phoneme sequences w.r.t. the ground truth, i.e., the sequence realized by the speaker. Thus, the lower the PER the better, and the baseline is the PER of the canonical pronunciation, that is before adaptation. Since PER does not perfectly reflect the quality of a pronunciation[1], perceptual tests on synthesized speech samples have also been conducted in the final experiments to fully validate the proposed method. To conduct these evaluations, the data of each speaker in the Buckeye corpus were randomly partitioned into a training set (60% of the utterances), a development set (20%), and a test set (20%). The training set has been used to train the models, while the development set is used to optimize the method, and the test set for final evaluations.

In our work, following this evaluation scheme, we have studied and optimized our pronunciation adaptation method with respect to various factors which appear on Figure 4.1 and are listed below:

(F1) One main challenge is to identify the optimal subset of features, i.e., the subset which leads to the minimum PER. Finding this optimal subset is of interest to

[1]Especially because all errors do not have the same importance and considering a unique pronunciation ground truth for the spontaneous style is too restrictive to very finely assess the quality of candidate pronunciations.

prevent training from being too long, and overfitting the data, as well as to provide knowledge about useful information for pronunciation modeling in general. To solve this problem, we propose a selection process that we apply on linguistic, articulatory, and prosodic features.

(F2) Next, information about the phoneme c_i may not be enough to predict the phoneme p_i, and it may be useful to consider the neighboring canonical phonemes and their associated features. Thus, the benefits of adjusting this neighborhood, i.e., adjusting the size of a window around c_i, have been examined.

(F3) It is not clear whether phonemes of an utterance should be processed word by word, or all at once. In other terms, the question is to wonder if information should be propagated across word boundaries. In our work, we have studied pronunciation adaptation on isolated words and on utterances, i.e., connected words.

(F4) Like in most machine learning approaches, it may be assumed that the more data the better. Nonetheless, pronunciation styles can greatly differ across speakers and this assumption might be uncertain in our case. To validate or invalidate this point, we have examined the effects of training CRF models on the data of each speaker from our corpus separately as well as by combining their data. These experimental conditions are referred to as speaker-dependent and independent adaptation respectively in the remainder.

(F5) Finally, we studied the question whether predicting the phoneme p_i depends on the preceding predicted phoneme p_{i-1} or not. Thus, we have considered different ways of integrating dependencies between predicted phonemes: either directly within the phoneme-to-phoneme CRF, or as a post-processing of adapted pronunciation hypotheses. Precisely, we propose to perform this post-processing through a rescoring mechanism based on a phonological n-gram model.

As a summary of all these studies, we end up with an effective adaptation method which is illustrated in Figure 4.2. The main conclusions are that feature selection and integration of the neighborhood context are useful. Likewise, including dependencies between predicted phonemes brings improvements but this requires to be done through a post-processing step rather than directly within the CRF. At the opposite, our studies shows that the difference between considering isolated words or continuous utterances is not clear neither. Lastly, the benefits of training CRFs on a large speaker-independent corpus are not very clear, with respect to the smaller speaker-dependent data. The following sections provide details about these conclusions. Specifically, factors F1-3 are

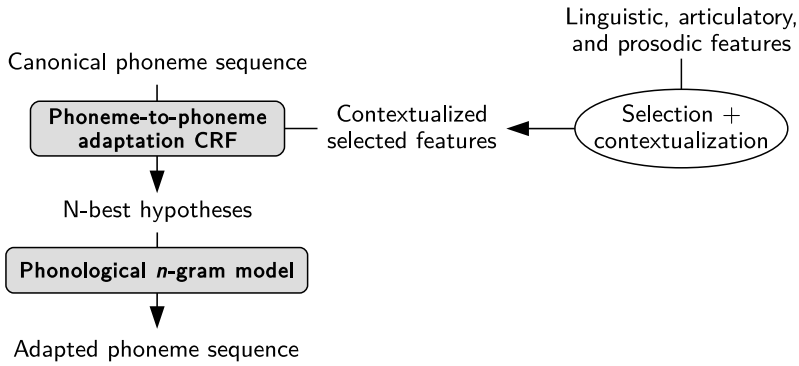


Figure 4.2: Overview of the proposed pronunciation adaptation method.

addressed as part of the phoneme-to-phoneme CRF training, presented in Section 4.2. Speaker-dependent and independent studies (F4) are detailed in Section 4.3. Finally, questions related to F5 are examined in Section 4.4.

4.2 Phoneme-to-phoneme spontaneous pronunciation adaptation using CRFs

In this section, details of the proposed CRF-based spontaneous pronunciation adaptation are given alongside individual studies on feature selection, phoneme neighborhoods, and cross-word information. These studies have been carried out on the development set to determine the best CRF configuration. Additionally, we only consider speaker-dependent adaptation in this section, i.e., one CRF is trained for each speaker separately. This choice has been made (i) to get rid of variabilities across speakers, thus, making training more accurate and tunning easier, (ii) to make the feature selection more robust, and (iii) to reduce training times.

In this section, we separately present the outcomes of tunning feature sets, phoneme neighborhoods, and cross-word information, while their combination is left for Section 4.3.

4.2.1 Feature selection

A crucial step in any machine learning task is to identify the set of features that best represent the given data using feature selection techniques. Selection provides many advantages. First, it identifies the most relevant features by removing redundant or less useful ones. Second, it reduces the time and memory needed for the training process as the models are trained using less features. For this purpose, a selection process is applied on the development set for all the three feature groups already presented

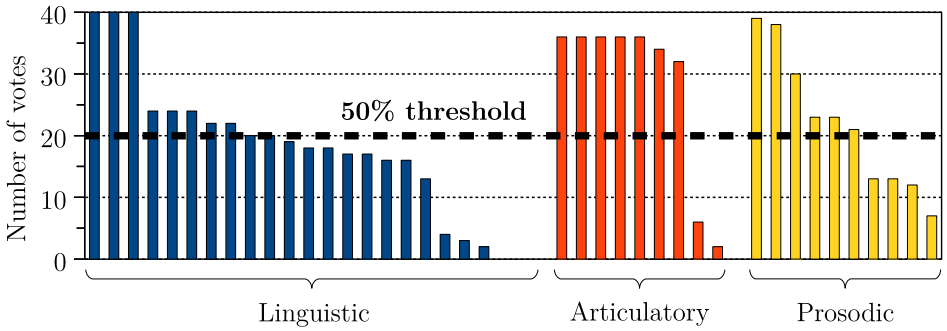


Figure 4.3: Number of votes for each linguistic, articulatory, and prosodic features.

in Section 3.3: linguistic, articulatory, and prosodic features.

The basic idea of our process is to run an election over the features by searching for the best feature set for each speaker, i.e., the set leading to a minimal PER by comparing canonical and realized phonemes. Features receive a vote each time they appear in the best set of some speaker. To make the selection process more robust, two selection schemes are combined. First, a greedy backward elimination was conducted where all features are considered at the beginning and features are eliminated one at a time until the best set is found. Second, a greedy forward selection was applied, i.e., the process starts with canonical phonemes as a unique feature and other features are added one at a time until the optimal set is found. Results of both methods were then summed to provide an overall ranking of the features. Figure 4.3 shows the number of votes for each feature when running feature selection in each feature group separately. Since there are 20 speakers and since each one provides 2 votes, a feature can get a maximum of 40 votes. Based on these votes, features which received less than 50% of the maximum number of votes, i.e., less than 20 votes, were discarded. This threshold has been empirically set on linguistic features and propagated to articulatory and prosodic ones[2]. No further tuning has been performed.

As a result of this selection process, in addition to the canonical phoneme which is always included, the following features were selected for each feature group:

- Linguistic features: phoneme position and reverse phoneme position in syllable, syllable lexical stress, syllable part, syllable location, word, word frequency in English, stop word, word boundary.

- Articulatory features: phoneme type, manner, place, shape, aperture, voiced, rounded.

[2]Various thresholds have been tested on linguistic features but no significant PER difference has been observed.

Table 4.1: PERs (%) for selected features versus all features. These tests are conducted on the development set and on isolated words. Absolute variations with the baseline are reported between square brackets.

Baseline (no adaptation)		28.3
Canonical phoneme only (with adaptation)		30.7 [+2.4]
+ Linguistic	Selected features (9)	25.1 [−3.2]
	All features (23)	26.6 [−1.7]
+ Articulatory	Selected features (7)	30.8 [+2.5]
	All features (9)	30.9 [+2.6]
+ Prosodic	Selected features (6)	26.7 [−1.6]
	All features (10)	27.1 [−1.2]

- Prosodic features: syllable energy, syllable F0 shape, pause per syllable, phone tone, distance to next and previous pause.

After the selection process was completed, we investigated the change in PERs of adapted pronunciations before and after selection for each group of features with comparison to the baseline, i.e., canonical phonemes. To make the process faster, all the experiments are conducted on isolated words and without considering phoneme neighborhoods. The results are presented in Table 4.1. First, results show that the sole 9 selected linguistic features lead to a significantly lower PER compared to the complete 23 features. Most selected features in this group are syllable-based, which makes the result consistent with previous studies [Vazirnezhad et al., 2009, Bell et al., 2009]. Concerning articulatory features, the impact of the selection on this feature type is limited, since out of 9 features, only two have been removed, i.e., affricate and doubled. These two features have received extremely low votes in comparison to other features, meaning that they might be completely irrelevant. Lastly, the feature selection brings a PER reduction for prosodic features even though it is small. The removed features in this group are speech rate, distance to next/previous hesitation, and syllable tone. As a conclusion, the effect of the feature selection is positive for each feature group, leading to a final set of 22 (excluding the canonical phoneme) remaining features for the experiments.

4.2.2 Window size tuning

One important step apart from feature selection is to decide on the neighborhood scope around each canonical phoneme, that is determining the best suited size of canonical phoneme windows. These windows are centred on the canonical phoneme to be adapted. They are symmetrically[3] defined by the number W of the left and right hand

[3] Asymmetric windows were also tested but they led to worse results.

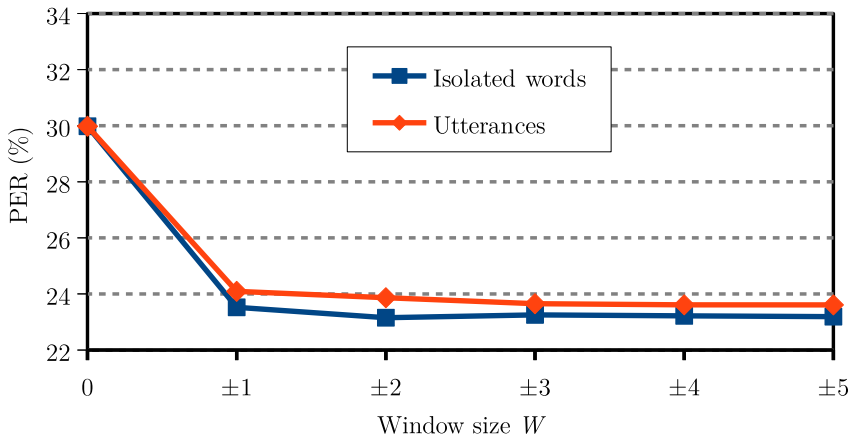


Figure 4.4: PERs (%) on the development set according to the window size, for isolated words and utterances.

surrounding phonemes. For instance, W=±2 means that 2 neighbors from each side are considered along with the current canonical phoneme, hence considering 5 phonemes in total. The investigated values of W range from 0 to 5.

Figure 4.4 presents PERs obtained without windows (W=0) or with different window sizes, for both isolated words and utterances. CRFs were trained without any other feature than canonical phonemes. Results show that phoneme neighborhoods bring significant improvements. For both isolated words and utterances, results seem to converge when W reaches ±2. To have consistency between isolated words and utterances, W is fixed at this value in the experiments.

4.2.3 Cross-word information

In addition to within-word pronunciation variations, there also exists variations across words in spontaneous speech. The latter is usually observed when a word is surrounded by some specific words. For instance, the phoneme /t/ in the word "*what*" (/wʌt/) is sometimes pronounced as a glottal stop when it is followed by the word "*I*" (/aɪ/) like in "*what I mean*", /wʌʔ aɪ miːn/. To verify that including cross-word information is useful, some tests on the development set using only the canonical phoneme were conducted. The results show that an improvement of 0.3 percentage points (pp) can be obtained when cross-word information is included with 30.7% and 30.4% of PER for isolated words and utterances respectively.

By now, we have determined the most useful features for spontaneous pronunciation adaptation, and the best size of windows to be considered. We have also reached

Table 4.2: Speaker-dependent PERs (%) on the test set for all possible combinations of feature groups with $W=\pm 2$ on isolated words and utterances. Absolute variations with the baseline are reported between square brackets.

					Isolated words	Utterances
Baseline (no adaptation)					28.3	28.0
	Can. ph.	Ling.	Artic.	Pros.		
1	✓				24.2 [−4.1]	25.2 [−2.8]
2	✓	✓			24.0 [−4.3]	23.7 [−4.3]
3	✓		✓		24.4 [−3.9]	25.2 [−2.8]
4	✓			✓	21.5 [−6.8]	22.0 [−6.0]
5	✓	✓	✓		24.0 [−4.3]	24.1 [−3.9]
6	✓	✓		✓	**21.1** [−7.2]	**20.8** [−7.2]
7	✓		✓	✓	21.4 [−6.9]	22.0 [−6.0]
8	✓	✓	✓	✓	21.2 [−7.1]	21.1 [−6.9]

a conclusion that including cross-word information might be useful for pronunciation adaptation. In the next two sections, the results of these three tests are combined to conduct speaker-dependent and independent experiments.

4.3 Speaker-dependent and independent adaptation

As we argued earlier, due to the possible pronunciation style differences across speakers in our corpus, it might not be a good idea to train CRF models on the combined data of all speakers. In this section, we will examine this issue through speaker-dependent and independent adaptation CRFs.

4.3.1 Speaker-dependent spontaneous adaptation

Speaker-dependent adaptation CRFs are basically trained and evaluated *independently for each speaker*. The PER is computed by comparing the realized phonemes with either the canonical phonemes, i.e., before adaptation, or those resulting from an adaptation for each speaker separately. Mean error rates are then reported by averaging PERs over all the speakers. PER before adaptation on the development and test sets is 28.3%. Individual error rates differ significantly across speakers, ranging from 22.0% to 39.8%. This disparity between the canonical and realized phonemes is a strong argument to perform pronunciation adaptation on a speaker basis rather than on all the speakers together, as capturing variations may be very difficult in the latter case.

Experiments were carried out on the test set of each speaker using canonical phonemes plus the three groups of selected features resulting from Section 4.2.1. The objectives of these experiments are to (1) determine which feature groups are most useful for pro-

nunciation adaptation; (2) which combination of features leads to the best results; and lastly (3) investigate if considering cross-word information is useful or not. This section presents the raw results before developing a deeper analysis.

Two series of experiments have been carried out for both isolated words and utterances. First, each group of selected features has been evaluated separately. Second, the groups have been combined in all possible ways. Table 4.2 reports PERs for isolated words and utterance experiments respectively compared with the baseline[4], i.e., without adaptation.

Firstly, it can be observed that using only linguistic features provides a small improvement over *canonical phoneme only* pronunciations in both isolated words and utterance experiments (*line 2, Table 4.2*). This difference in utterance experiments is larger due to the absence of word boundary information in (*line 1*). Concerning articulatory features, it can be noticed that adding them (*line 3*) brings worse results in the case of isolated words and does not provide any improvement on utterances over the *canonical phoneme only* configuration. Finally, prosodic features (*line 4*) lead to a clear improvement with a reduction of 3.1 pp for isolated words and 4.4 pp for utterances compared to the *canonical phoneme only* configuration. Although extracting the features directly from the signal instead of predicting them might have a big role in this result, this shows how important prosodic features are for pronunciation adaptation.

In the second series of experiments where feature types are combined, we can see that combining articulatory features with any other feature group (*line 5, 7*) brings worse results. In contrary, linguistic and prosodic features (*line 5, 6, 7*) always improve the results. In both isolated words and utterance experiments, combination of linguistic and prosodic features (*line 6*) brings the best results with 21.1% and 20.8% on isolated words and utterances respectively.

Overall results demonstrate that (i) prosodic features have the strongest influence in pronunciation adaptation, (ii) articulatory features lead to worse results in most experiments which clearly shows that they carry no additional information over the canonical phoneme, (iii) although linguistic features alone have minimal effect, when combined with other features they bring extra improvements, and (iv) considering cross-word information brings a small improvement (particularly when linguistic and prosodic features are combined (*line 6*)), however this improvement is statistically significant[5].

Next, we compared adaptation models by measuring how well they predict the realized pronunciations, that is, how high their probability is given the reference. This can be achieved by computing the perplexities of the test set according to the different

[4]Baseline numbers between isolated words and utterances have different values. This is because the alignment might slightly change in case of utterances.

[5]The p-values are 6.889×10^{-4} and 8.005×10^{-4} using a paired t-test and a paired Wilcoxon test, respectively, with a confidence level $\alpha = 0.05$.

Table 4.3: Speaker-dependent perplexity on the test set for all possible combinations of feature groups with $W=\pm 2$ on isolated words and utterances. Relative variations with the canonical phoneme only configuration are reported between square brackets.

	Can. ph.	Ling.	Artic.	Pros.	Isolated words	Utterances
1	✓				2.71	2.68
2	✓	✓			2.69 [−0.7 %]	2.57 [−4.1 %]
3	✓		✓		2.65 [−2.2 %]	2.62 [−2.2 %]
4	✓			✓	2.49 [−8.1 %]	2.51 [−6.3 %]
5	✓	✓	✓		2.66 [−1.9 %]	2.58 [−3.7 %]
6	✓	✓		✓	2.45 [−9.6 %]	2.4 [−10.5 %]
7	✓		✓	✓	2.42 [−10.7 %]	2.44 [−9 %]
8	✓	✓	✓	✓	**2.41** [−11.1 %]	**2.39** [−10.8 %]

models. Table 4.3 presents perplexity results over phonemes for all the adapted pronunciation models. The results partially confirm what is observed when computing PERs. On the one hand, separately tested features (*line 1-4*) have consistent results with those of PERs. The canonical phoneme only configuration has the highest perplexity, while prosodic features lead to the lowest one. Similarly to PER results, linguistic and articulatory features have less impact than prosodic features. On the other hand, when features are combined, articulatory features do not always bring worse results (*line 7, 8*). This can be again observed when combining all the features, which leads to the lowest perplexity.

Results are thus consistent to what was achieved with PERs, although minor differences exist particularly in case of linguistic + prosodic and linguistic + articulatory + prosodic features. Finally, these results confirm the benefits of combining different features since the best values are always achieved when different feature groups are combined.

4.3.2 Speaker-independent spontaneous adaptation

In speaker-independent adaptation experiments, the training data of all speakers are combined together and then a phoneme-to-phoneme CRF model is trained and validated on the combined test sets. The main objective here is to know if increasing the amount of training data will compensate for the disparity of the data and improve the results. We would also like to know which feature groups are mostly affected by this change in the amount of data.

Speaker-independent experiments similar to dependent ones were carried out using the canonical phonemes plus the same three other groups of selected features for both isolated words and utterances separately. Table 4.4 reports PERs for isolated words

Table 4.4: Speaker-independent PERs (%) on the test set for all possible combinations of feature groups with $W=\pm 2$ on isolated words and utterances. Absolute variations with the baseline are reported between square brackets.

					Isolated words	Utterances
Baseline (no adaptation)					28.3	28.0
	Can. ph.	Ling.	Artic.	Pros.		
1	✓				24.8 [−3.5]	25.5 [−2.5]
2	✓	✓			24.8 [−3.5]	25.5 [−2.5]
3	✓		✓		24.9 [−3.4]	25.5 [−2.5]
4	✓			✓	20.6 [−7.7]	21.5 [−6.5]
5	✓	✓	✓		24.8 [−3.5]	25.3 [−2.7]
6	✓	✓		✓	**20.3** [−8.0]	20.9 [−7.1]
7	✓		✓	✓	20.6 [−7.7]	21.4 [−6.9]
8	✓	✓	✓	✓	**20.3** [−8.0]	**20.8** [−7.2]

and utterances respectively compared with the baseline. First of all, we clearly see that compared to speaker-dependent experiments, the first three configurations achieve worse results. Moreover, linguistic features no longer add any improvement over the canonical feature only configurations (*line 1 and 2*, Table 4.4). On the contrary, prosodic features lead to a clear improvement with a reduction of 0.9 and 0.5 pp for isolated words and utterances respectively compared to the speaker-dependent configurations (*line 4*). This shows that the prosodic feature patterns might be more similar across different speakers in the corpus.

Then in the second series of experiments where feature types are combined, we can see that the configurations that contain prosodic features mostly achieve better results than their corresponding speaker-dependent experiments. The configurations which include all features bring the best results with 20.3% on isolated words and 20.8% on utterances. It can be noticed that no additional improvements is achieved for utterances when the amount of data is increased (*line 6* in Table 4.2 and *line 8* in Table 4.4). As for isolated words, the improvement is much clearer, as a reduction of 0.9 pp is achieved.

In order to validate the improvements—particularly in the configurations that include prosodic features—we compared speaker-independent CRF models by computing their perplexities. The results are presented in Table 4.5. In general, perplexity is consistent with PER and shows that speaker-independent experiments lead to better results particularly when including prosodic features. Among separately tested features, the configuration with prosodic features (*line 4*) achieve a relative reduction of 16.3% and 10.9% on both isolated words and utterances, whereas on combined feature experiments, the configuration which include all the features lead to the lowest perplexity with a higher relative reduction compared to speaker-dependent experiments.

Table 4.5: Speaker-independent perplexity on the test set for all possible combinations of feature groups with $W=\pm 2$ on isolated words and utterances. Relative variations with the canonical phoneme only configuration are reported between square brackets.

	Can. ph.	Ling.	Artic.	Pros.	Isolated words	Utterances
1	✓				2.61	2.51
2	✓	✓			2.54 [−2.8 %]	2.47 [−1.5 %]
3	✓		✓		2.57 [−1.8 %]	2.48 [−1.3 %]
4	✓			✓	2.19 [−16.3 %]	2.24 [−10.9 %]
5	✓	✓	✓		2.58 [−1.3 %]	2.44 [−2.8 %]
6	✓	✓		✓	2.17 [−17.1 %]	2.18 [−13.0 %]
7	✓		✓	✓	2.16 [−17.2 %]	2.20 [−12.2 %]
8	✓	✓	✓	✓	**2.15** [−17.6 %]	**2.17** [−13.5 %]

In conclusion, based on both PER and perplexity results, it can be confirmed that the increase in the amount of training data does not bring improvements in most cases for linguistic and articulatory features. Again this is probably due to the disparity of the data across different speakers. As for prosodic features, it was shown that results vary positively and significant improvements are achieved. This proves that prosodic features are more similar across the different speakers in the corpus.

4.4 Phonological reranking

Adapted pronunciations resulting from CRFs can sometimes show undesired behaviors, for example prediction of /d t/ successively which is unlikely in English. These kinds of predictions occur because CRFs ignore phoneme dependencies when predicting a new phoneme, i.e., they do not take into account the previously predicted phoneme. Two separate approaches were followed in order to introduce predicted phoneme dependencies. First, CRF models were configured to directly take into account these dependencies. Alternatively, pronunciation generated by CRFs were rescored and reranked using a phonological n-gram model trained on realized phonemes. Both approaches were first tested on the configuration including linguistic + prosodic features. Then for comparison purposes, canonical phoneme only and linguistic features configurations were considered as well.

4.4.1 Phoneme dependencies using CRFs

CRFs can take into account phoneme dependencies using bigram and uni+bigram configurations (see Section 2.2.3). Table 4.6 presents the comparison of these two configurations against unigrams for isolated words. It can be noticed from the results that

Table 4.6: PERs (%) on the development set for unigram, bigram, and uni+bigram configurations on canonical phoneme, linguistic and linguistic + prosodic feature configurations. Absolute variations with the baseline are reported between square brackets.

Baseline (no adaptation)			28.3
	Canonical phoneme	Linguistic	Linguistic + prosodic
Unigram	24.5 [−3.8]	24.0 [−4.3]	21.5 [−6.8]
Bigram	30.9 [2.6]	32.3 [4.0]	32.6 [4.3]
Uni+bigram	24.8 [−3.5]	24.6 [−3.7]	22.8 [−5.5]

bigrams lead to increase in PER for all the experiments. Particularly, in the case of linguistic + prosodic features, compared to the unigram configuration, the PER is increased by 11.1 pp for bigram and 1.3 pp for uni+bigram. This behavior is probably due to the sparsity of the data in the training set where only a limited number of realized phoneme bigrams can be observed. These results show that phoneme dependencies using CRFs should be avoided in our case.

4.4.2 Phoneme dependencies using a phonological n-gram model

In order to introduce predicted phoneme dependencies in another way, a phonological model was trained and used to rescore and rerank N-best hypotheses predicted by CRF models. Precisely, each hypothesis $\mathbf{h} = (p_1, \ldots, p_n)$ of n phonemes p_i is assigned a score $s(\mathbf{h})$ mixing the CRF and phonological model (PM) probabilities. This mixture is computed by a log-linear interpolation—which has been successfully used for N-best list reranking in various domains [Rosti and Matsoukas, 2007, Huet et al., 2010]—, and is formulated as follows:

$$s(\mathbf{h}) = \Pr_{\text{CRF}}(\mathbf{h}) \times \Pr_{\text{PM}}(\mathbf{h})^{\alpha} \times \beta^{n} \ , \tag{4.1}$$

where α and β are two parameters to be optimized. The parameter β is used to prevent the phonological model from favoring short hypotheses. Finally, the hypothesis with the highest score is selected as the adapted pronunciation.

In our experiments, the phonological model is a phoneme-based n-gram model estimated on the training set using a Witten-Bell smoothing. The order n of the model as well as α and β have been optimized such that they minimize PER on the development set, and consequently set to 5, 0.48 and 0.024, respectively. Training, optimization and reranking have all been conducted using SRILM [Stolcke et al., 2011]. Reranking is performed on the 10 best hypotheses predicted by the adaptation CRF, as empirically tuned on the development set.

As shown in Table 4.7, our reranking technique on speaker-dependent experiments

Table 4.7: Speaker-dependent PERs (%) for canonical phoneme, linguistic, and linguistic + prosodic feature configurations with $W=\pm 2$ before and after reranking using the 10 best hypotheses on the test set. Absolute variations with the baseline are reported between square brackets.

Isolated words

Baseline (no adaptation)	28.3	
	Before reranking	After reranking
Canonical phoneme only	24.2 [−4.1]	23.7 [−4.6]
+ Linguistic features	24.0 [−4.3]	23.7 [−4.6]
+ Linguistic + prosodic features	21.1 [−7.2]	**20.6** [−7.7]

Utterances

Baseline (no adaptation)	28.0	
Canonical phoneme only	25.2 [−2.8]	25.0 [−3.0]
+ Linguistic features	23.7 [−4.3]	23.5 [−4.5]
+ Linguistic + prosodic features	20.8 [−7.2]	**20.7** [−7.3]

always reduces PERs. The highest reduction (0.5 pp) is achieved on isolated words for canonical phoneme and linguistic + prosodic configurations. Concerning utterances, the impact of reranking appears to be more limited, as it only reduces the results by 0.1 - 0.2 pp. This behavior can be explained by the fact that, for utterances, some hypotheses are too long, and narrowing down the number of hypotheses to only 10 does not offer enough diversity to find a better hypothesis. To check if including more hypotheses can further reduce PERs on utterances, the number of hypotheses was increased to 100. Still the results were not up to expectations as the PER on linguistic + prosodic features was reduced by only 0.1 pp to 20.6%. We strongly believe that this is due to the variable length of utterances in our corpus, since the number of words in an utterance can be as low as few words or as high as hundreds of words. Adaptation results when considering few hypotheses for long utterances is limited, while, having many hypotheses worsen the results in the case of short utterances. A reasonable compromise is to consider a variable number of hypotheses based on the length of the utterance.

Next, PER results for speaker-independent experiments are shown in Table 4.8. As it can be seen, on isolated words, the reranking is not very effective on the canonical phoneme and linguistic + prosodic feature configurations, whereas for the configuration with only linguistic features, our reranking seems to provide a very significant improvement of 0.7 pp. On the side of utterances, similarly to the speaker-dependent experiments and probably for very similar reasons, the reranking process does not provide much improvement, since, in two out of the three experiments, PERs remain completely unchanged.

The outcome here is that the reranking process has a positive impact on adapted

Table 4.8: Speaker-independent PERs (%) for canonical phoneme, linguistic, and linguistic + prosodic feature configurations with $W=\pm 2$ before and after reranking using the 10 best hypotheses on the test set. Absolute variations with the baseline are reported between square brackets.

Isolated words

Baseline (no adaptation)	28.3	
	Before reranking	After reranking
Canonical phoneme only	24.8 [−3.5]	25.1 [−3.2]
+ Linguistic features	24.8 [−3.5]	24.1 [−4.2]
+ Linguistic + prosodic features	20.3 [−8.0]	**20.1** [−8.2]

Utterances

Baseline (no adaptation)	28.0	
Canonical phoneme only	25.5 [−2.5]	25.5 [−2.5]
+ Linguistic features	25.5 [−2.5]	25.4 [−2.6]
+ Linguistic + prosodic features	20.9 [−7.1]	**20.9** [−7.1]

pronunciations particularly on isolated words. Overall results show that the proposed approach reduces the PERs to a great extent: with a baseline score of 28.3% on isolated words, a significant improvement of 7.7 pp is achieved using linguistic + prosodic features after phonological reranking.

4.5 Perceptual tests

To assess the impact of our approach on synthesized speech samples, AB tests on 40 synthesized speech samples have been conducted with 10 native English speakers. Listeners were asked to answer two questions: *Between A and B, which sample is pronounced in the most spontaneous way?* and *Which sample is pronounced in the most intelligible way?*. For both questions, listeners can also indicate that they do not hear any difference. Orthographic transcripts were given along with the samples to help listeners to focus on pronunciations. Tests were set up to compare canonical and realized pronunciations to those generated using our speaker-dependent adaptation method with various configurations: either based on the sole canonical phonemes (C), additionally with linguistic features (C + L), or linguistic and prosodic features together (C + L + P), all including phonological reranking.

Utterances have been selected among the 2,000 available utterances in the test set such that their PER between the canonical and realized pronunciations is high. This strategy has been designed to ensure that selected utterances reflect the difficulty of the task. Utterances were synthesized using HTS v2.2 trained with standard features [Zen et al., 2007] and on the Blizzard Challenge 2012 data [King and Karaiskos, 2012],

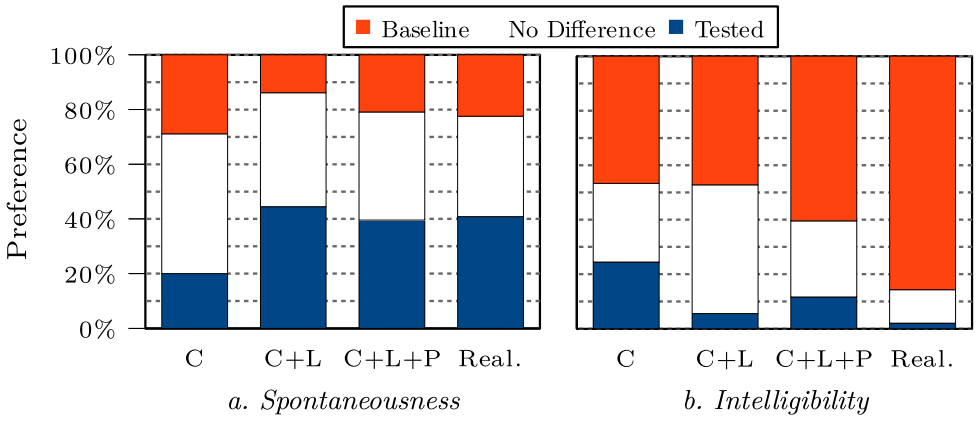


Figure 4.5: Preference on spontaneousness and intelligibility by comparing realized and adapted pronunciations to the baseline. Adaptations were performed using canonical phonemes (C), linguistic features (L), and prosodic features (P).

i.e., audiobooks with mixed speech styles and uttered by a US male speaker. Hence, no bias toward standard or spontaneous speech can be observed. Let us precise that unit selection has voluntarily been excluded here since this type of system is usually very sensitive to pronunciation variants, producing disturbing artefacts.

Figure 4.5 shows the comparison of speech samples generated using the baseline pronunciations against adapted or realized ones in terms of (a) spontaneousness and (b) intelligibility. Preference percentages are given as bar segments on the y-axis. Statistical significances of these ratios have been computed for all the tests[6]. First, we can notice that realized pronunciations are logically judged as more spontaneous than the baseline, while being much less intelligible. Regarding adapted pronunciations, the configuration C performs poorly. Conversely, the two other adapted configurations are judged as much more spontaneous than the baseline, but again leading to intelligibility degradations. Finally, adaptation performs equally or even slightly better when using linguistic features alone, i.e., without prosodic ones. This is interesting since predicting prosodic features is difficult in TTS.

To complete these results, Figure 4.6 compares realized pronunciations against adapted ones. Results against the baseline are also reported from Figures 4.5. Surprisingly it appears that C + L and C + L + P configurations are preferred over the realized pronunciations for spontaneousness. This importantly proves that pronunciations adapted using our method strongly reflect a spontaneous style. Then, samples resulting from realized pronunciations are always considered to be less intelligible. Lastly, it can again be noticed that the sole use of linguistic features performs slightly better than

[6]Binomial test with $\alpha = 0.1$ and votes for "No Difference" equally spread over A and B, following the methodology proposed in [Karhila et al., 2014].

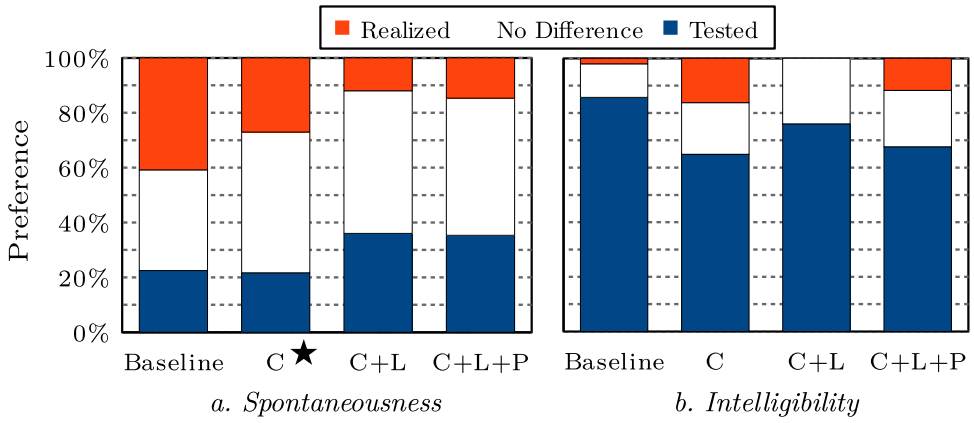


Figure 4.6: Preference on spontaneousness and intelligibility by comparing baseline and adapted pronunciations to realized ones. Adaptations were performed using canonical phonemes (C), linguistic features (L), and prosodic features (P). ★ stands for "*not statistically significant*[6]".

when also accounting for prosodic features, especially regarding intelligibility. While this counterbalances conclusions of PER in Table 4.7, a qualitative analysis shows that pronunciations produced using prosodic features, as well as the realized ones, are too complex for current TTS systems, especially because of strong coarticulations such as /dn̩/ (like in "didn't") or /fm/ ("familiarity"). This penalizes intelligibility and, as a side effect, spontaneousness.

In conclusion, the conducted tests entirely validate our proposed pronunciation adaptation method. They even further show that prosody is not necessary to produce spontaneous pronunciations.

4.6 Extension to corpus-specific adaptation

In this section the same pronunciation adaptation method that was introduced in Section 4.1 will be applied. The idea is to solve the problem of inconsistency between the phoneme sequences generated by G2P converters during synthesis and those from the TTS speech corpus. Generally, in a TTS system, the waveform is generated from this phoneme sequence by querying a dedicated database of speech segments or generative models, be it a unit selection or an SPSS system. In both cases, the system has been built using a speech corpus in which realized phonemes have been carefully labelled and segmented. Hence, TTS systems highly depend on the consistency between phonemes as labelled in their underlying speech corpus and those generated by the phonetizer during synthesis. Especially, strong differences would lead to a low quality of the synthesized speech signals. In the case of unit selection, inconsistencies would result in

a low number of candidate segments and a high number of concatenations, while, in systems like HTS, they would end up in using poorly trained or non-contextual models. To solve this problem, here, we propose to adapt phonemes generated by the phonetizer to the style of the TTS speech corpus, in order to minimize the difference between the two.

In this section, the corpus used for this specific task and the complete list of features used to train CRF models are first introduced. Then, PER evaluation is provided before conducting a perceptual test.

4.6.1 Corpus

This particular work adapts the method we proposed in Section 4.1 to a French speech corpus dedicated to interactive vocal system TTS. The corpus covers all diphonemes present in French and comprises most used words in the telecommunication field. It is composed of $7,208$ utterances, containing $225,08$ phonemes and $24,160$ non speech sounds, totaling 6h40 of speech. Pronunciations and non speech sounds have been strongly controlled during the recording process. Other information has been automatically added and manually corrected.

The corpus has been randomly split in two parts: a training set (70%) and a test set (30%). The training set has been divided in seven folds, and used to select and combine features in cross-validation conditions. Models are trained on six folds, the remaining fold being used for testing. The test set is used to evaluate the resulting pronunciation models in final experiments in terms of PER and through perceptual tests. This protocol ensures that data used for training the models and data used for validation do not overlap.

4.6.2 Features

For this work, a total of 52 linguistic, phonological, articulatory and prosodic features has been added to the corpus. The features presented in Table 4.9 are inspired by our spontaneous adaptation work, however they have been enriched and adapted to French. Most features have been normalized to corpus or utterance and discretized.

Features are first selected separately for each group of features using a forward selection process. Then groups of selected features are combined to find the optimal configuration. Selected features are reported in bold in Table 4.9 along with their number of votes. First, the feature selection results show that 2 linguistic features were selected for all the folds: the word itself and its stem. Since these features are highly correlated, one would have expected only one feature to be selected. However, as stated in [Guyon and Elissef, 2003], “noise reduction and consequently better class separation

Table 4.9: Groups of features used for corpus-specific adaptation experiments. In bold, features that have been selected. In brackets, the number of votes.

Feature	Value
a. Linguistic features (18)	
Word [7]	word
Stem [7]	word stem
Lemma [0]	word lemma
POS [2]	part of speech tag
Stop word [0]	boolean
Word [0], stem [2], lemma [1] freq. in French	common, normal, rare
Word [1], stem [1], lemma [2] freq. in corpus	common, normal, rare
Word prob. knowing previous word in French [2], in corpus [1]	common, normal, rare
Word prob. knowing next word in French [2] in corpus [3]	common, normal, rare
Number of word occurence in corpus [0]	integer
Word position [3], reverse position [0] in utterance	integer
b. Phonological features (17)	
Canonical syllables [7]	syllable phonemes
Phoneme in syllable position [0]	integer
Phoneme in word position [0]	begin, middle, end
Syllable in word position [6]	integer
Phoneme position [0] and **reverse position** [4] in syllable	integer
Phoneme position [5] and **reverse position** [5] in word	integer
Syllable position [3] and reverse position [1] in word	integer
Word length in phoneme [4]	integer
Word length in syllable [2]	integer
Syllable short [1] and long [0] structure	CVC, CCVCC
Syllable type [1]	open, closed
Phoneme in syllable part [0]	onset, nucleus, coda
Pause per Syllable [4]	low, normal, high
c. Articulatory features (9)	
Phoneme type [2]	vowel, consonant
Phoneme aperture [3], shape [1], place [1] and manner [2]	open, close, front, etc.
Phoneme is affricate [0], rounded [3], doubled [0] or voiced [3] ?	boolean
d. Prosodic features (7)	
Syllable Energy [7]	low, normal, high
Syllable [4] and phoneme [7] tone	from 1 to 5
F_0 **phoneme contour** [7]	decreasing, flat, increasing
Speech rate [7]	low, normal, high
Distance to next [3] and **previous pause** [7]	from 1 to 3

Table 4.10: Average PERs on the training set obtained on 7 folds. In brackets, percentage point w.r.t. the baseline.

Baseline (no adaptation)		11.5 [0.0]
Canonical phoneme only		6.9 [-4.6]
+ Linguistic	All features (18)	4.4 [-7.1]
	Selected features (2)	4.4 [-7.1]
+ Phonological	All features (17)	4.5 [-7.0]
	Selected features (7)	4.6 [-6.9]
+ Articulatory	All features (9)	7.1 [-4.4]
	Selected features (0)	-
+ Prosodic	All features (7)	4.8 [-6.7]
	Selected features (6)	4.8 [-6.7]

may be obtained by adding variables that are presumably redundant". Word frequencies and left/right linguistic context features received only very few votes. Moreover, 7 phonological features were included in the optimal set. Most of the selected features concern phoneme positions in the utterance. None of the characteristics of syllables (such as syllable part, structure or type) have been selected. Surprisingly, it appears that no articulatory features have been selected. Since previous studies have shown the interest of such features for pronunciation variation modeling [Livescu et al., 2016], they were expected to have better votes. Finally, 6 out of 7 prosodic features have been selected. The only feature which was discarded by the selection algorithm is "distance to next pause".

Table 4.10 presents the average PERs obtained on the seven folds before and after selection for each group of features with comparison to the baseline. The baseline PER is obtained by comparing phoneme sequences generated by the phonetizer and realized phoneme sequences (ground truth). An improvement of 4.6 pp is obtained while using a pronunciation model trained with canonical phonemes only, thus showing how pronunciation adaptation can reduce the inconstancy between the phonetizer output and the speech corpus. Separately adding a group of features further improves the PER, except with the articulatory group. Interestingly, the reduction of the number of features in each group does not affect these average PERs. The most significant reduction lies in the linguistic group: with only two apparently redundant features, a drop of 7.1 pp is obtained from the baseline.

Table 4.11: PERs obtained on the test set. In brackets, percentage point w.r.t. the baseline.

Baseline (no adaptation)	11.2
Canonical phoneme only (with adaptation)	4.1 [−7.1]
+ Linguistic + phonological	3.2 [−8.0]
+ Linguistic + phonological + prosodic	2.7 [−8.5]

4.6.3 Evaluation

This section focuses on the objective evaluation of the trained adaptation models. Pronunciation models are now trained on the whole training set and validated on the test set. They are tested with different subsets of features: canonical phonemes only, best selected linguistic and phonological features and best selected linguistic, phonological and prosodic features. All experiments have been conducted on a windows size of ±2. The window size has been determined as a result of several experiments similarly to Section 4.2.2.

PERs provided in Table 4.11 show that an improvement of 7.1 pp is obtained using a pronunciation model trained with canonical phonemes only. Adding linguistic and phonological features brings an additional improvement of 0.9 pp over the canonical phoneme only configuration. Finally, the most significant reduction is achieved with the combination of selected linguistic, phonological and prosodic features which leads to a reduction of 8.5 pp. Based on the PER results, we are expecting that the addition of three groups of features to the canonical phonemes improves the synthesized speech quality.

4.6.4 Perceptual tests

To assess the quality of synthesized speech samples generated with adapted pronunciations, a perceptual test was conducted with 14 French native speakers. Similarly to experiments on spontaneous speech, the evaluation is based on AB tests with 40 utterances. Listeners have to answer the following question: *"Between A and B, which sample reaches the best quality ?"*. Possible answers are: A, B, or no difference. Utterances were randomly selected by subsampling the test set according to the PER distribution between canonical and realized pronunciations. Speech samples were synthesized using the unit selection TTS system described in [Guennec and Lolive, 2014] and also with HTS v2.2 with standard features [Zen et al., 2007]. Five pronunciations are evaluated: canonical phonemes without adaptation (baseline), adapted phonemes based on canonical phonemes (C), selected linguistic and phonological features (C + L + Ph), selected

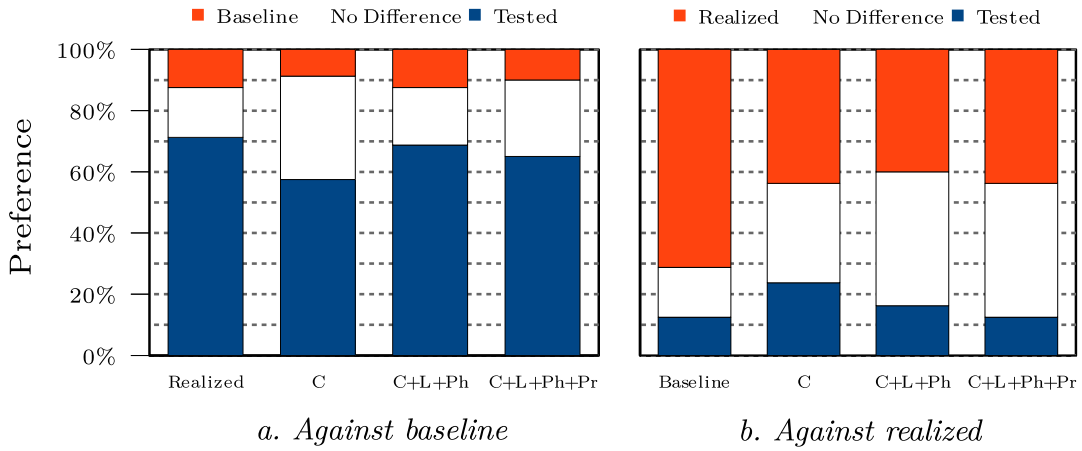
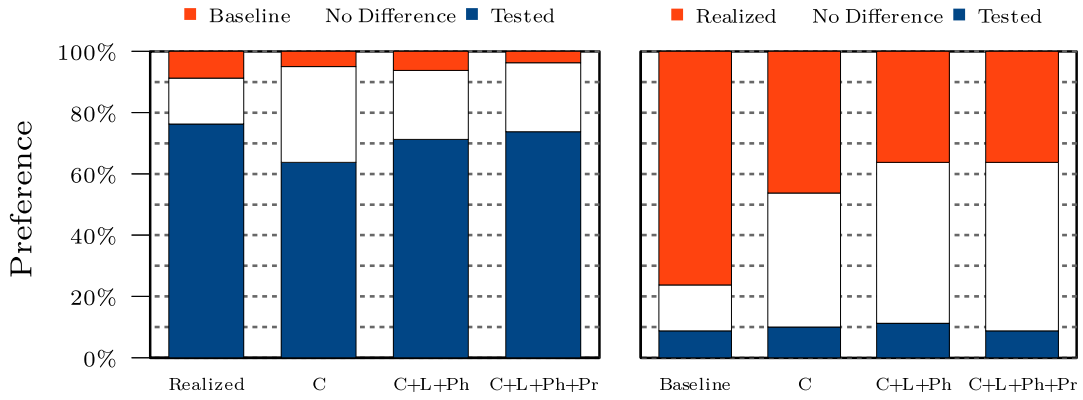


Figure 4.7: AB test results with unit selection, (a): realized, C, C + L + Ph and C + L + Ph + Pr compared against baseline, (b): baseline, C, C + L + Ph and C + L + Ph + Pr compared against realized.

linguistic, phonological and prosodic features (C + L + Ph + Pr) and realized phonemes as they are annotated in the speech corpus.

Figures 4.7 and 4.8 show the comparison of speech samples using adaptation against the baseline (a) and the realized (b) pronunciations with the two synthesis systems. Tested systems are expected to be mostly preferred against the baseline, that is the larger the blue bar, the better. At the opposite, the tested system is considered as correct when its signals are preferred or judged as similar against the realized signals, that is the smaller the red bar, the better. With both synthesis systems, adapted pronunciations resulting from the presented approach outweigh the baseline pronunciations in terms of quality. The addition of linguistic and phonological features increases the ratio of preferred adapted pronunciations. However, again, prosodic features do not seem to improve TTS quality, what is of interest since these features are not easy to obtain from text.

Adapted pronunciations can be considered as correct in comparison to realized pronunciations because the synthesized adapted pronunciations are mainly judged as similar to or even better than the realized pronunciations (in more than 50% of the samples). Interestingly, the C+L+Ph configuration is even more preferred than the configuration with prosodic features. This confirms that linguistic and phonological features are more robust than prosodic features. Based on this perceptual evaluation, it seems that pronunciation adaptation using linguistic and phonological features is our best model.



a. Against baseline *b. Against realized*

Figure 4.8: AB test results with unit HTS, (a): realized, C, C + L + Ph and C + L + Ph + Pr compared against baseline, (b): baseline, C, C + L + Ph and C + L + Ph + Pr compared against realized.

4.7 Discussion

In this section, a qualitative analysis of both spontaneous and corpus-specific adaptations is provided. This analysis consists in observing the most common errors in the test set through the comparison of adapted pronunciations with the realized ones. The purpose here is to better understand the results by identifying the most common situations or patterns that lead to errors.

First, concerning spontaneous adaptation results, rather obviously the most common errors are substitutions, i.e., the predicted phoneme does not match the realized phoneme. In Table 4.12, some of the most frequently substituted phonemes are given along with: their occurrence frequency in the test set, most common phonemes they are substituted with, and their position in the word when a substitution occurs. A possible explanation of why these phonemes are mostly substituted is the high frequency of the words that these phonemes are located in, such as *a*, *and*, *the*, *went*, *yknow*, etc. Some of these words tend to have frequent variants such as the word "the" usually being realized as either /ðʌ/ or /ðɪ/. Therefore, CRF models might easily mispredict them. This is also the reason why the phoneme /ʌ/ is mainly substituted with /ɪ/. In other cases, some phonemes are substituted with other phonetically close phonemes such as a substitution of /n/ with /ŋ/. This kind of substitution is considered as an error when computing PERs; however, in the perceptual tests, some of these substitutions are obviously not noticed by testers. Thus, better results are achieved in the perceptual tests than in the objective evaluations. Classifying the substitutions based on their position in a word is an important factor to identify at which places these substitutions mostly

Table 4.12: Most frequently substituted phonemes along with their frequency, the most common phonemes they are substituted with and their position inside word when they are substituted.

			Position in word		
Phoneme	Freq.	Substitutions	Initial	Middle	Final
/ʌ/	8.9%	/ɪ, o, ɛ/	35.5%	**43.1%**	21.5%
/ɪ/	7.1%	/ʌ, i, ɛ/	27.9%	**43.0%**	29.0%
/n/	6.0%	/ð, n̩, ŋ/	**41.8%**	18.8%	**39.4%**
/s/	4.9%	/z, ð, ʃ/	17.9%	9.1%	**73.0%**

Table 4.13: Pronunciations for the word "community" before and after reranking on isolated words using linguistic+acoustic features.

Realized phonemes	/k ɪ m j _ ɪ n ɪ d i/
Canonical phonemes	/k **ʌ** m j **u** _ n **ʌ** **t** i/
Adapted phonemes before reranking	/k _ m j **u** **ɪ** **ɾ** _ **ɾ** i/
Adapted phonemes after reranking	/k **ʌ** m j **u** **ɪ** _ _ **ɾ** i/

occur. For example, consonants like /n/ and /s/ are usually substituted when they are at the word-initial or word-final positions, while vowels like /ʌ/ and /ɪ/ are mostly substituted at the middle of the word. Taking further measures, such as adding specific features in CRFs to handle cases where these common substitutions occur, could help improving the results even more.

Second, in order to understand the results after phonological reranking, in Table 4.13 an adapted pronunciation example before and after reranking along with the canonical and realized pronunciations are provided. In this example, it can be seen that the canonical phonemes are rather far from the realized ones. After adaptation and before reranking, the pronunciations seem to be more spontaneous. Yet, it can be noticed that some portions of the pronunciation, such as having two successive /ɾ/, are phonetically not possible. After reranking, as it can be observed, this error has been successfully fixed. Despite this, many other errors exist even after reranking such as /ɪ/ still being predicted as /ʌ/.

Finally, regarding corpus-specific adaptation results, our analysis shows that the most frequent confusions between canonical and realized phonemes concern allophones: /o/ ≒ /ɔ/, /e/ ≒ /ɛ/ and /ɛ̃/ ≒ /œ̃/. Such confusions cannot be considered as errors in French but rather related to different speaking styles. Similarly, some of the frequent observed insertions include the /ə/ which is known to be optionally elided in French. Other substitutions concern labeling strategies and alphabet choices, for instance /ɲ/ ≒ /nj/, /ə/ ≒ /ø/. Deletions mainly concern liaisons between words, such as /t/ and /z/ which are not generated by the phonetizer whereas systematically

pronounced in the speech corpus. Pronunciation models contribute to minimize all these confusions.

From this analysis, we conclude that the pronunciations resulting from our adaptation models are reasonably good in reflecting a spontaneous style, and there is still room for further improvements. As for corpus-specific-adaptations, we showed that most of the confusions in predicted phonemes are mostly related to different speaking styles, thus, cannot be considered as errors in French.

4.8 Conclusion

In this chapter, a CRF-based pronunciation adaptation method was proposed for the purpose of speech synthesis. The proposed method which relies on using a wide range of features was validated in the context of spontaneous and corpus-specific adaptation.

Using objective measures and perceptual tests, backend experiments on the Buckeye corpus showed that adapted spontaneous pronunciations using a combination of the linguistic and prosodic features significantly better reflect spontaneous speech than standard ones. Listening test results even suggest that speech samples synthesized using adapted pronunciations are perceived as more intelligible than those using pronunciations realized by real speakers. Moreover, it was verified that linguistic features alone perform well for the task of pronunciation adaptation, since the spontaneousness and intelligibility of the speech samples generated using this group of features are comparable or even better than those of prosodic features.

Concerning corpus-specific adaptations, it was shown that the proposed method brings an improvement of 7.9 pp in terms of PER. Perceptual tests also showed an improvement in the quality of speech synthesis when pronunciation models are included in the phonetization process. Hence, we have shown that pronunciation adaptation helps to reduce inconsistencies between phonemes as labelled by their underlying speech corpus and those generated by the phonetizer during synthesis.

Overall, we can say that although the technique was applied on two different corpora for two different tasks, the results have very similar aspects. Thus, it can be concluded that the proposed method is effective when applied on the task of pronunciation adaptation.

Chapter 5

Generation of speech disfluencies

Another aspect that makes speech, expressive, is that it contains disfluencies. Speech disfluencies have been shown to make speech richer by serving a wide range of functions, such as aiding in language production and comprehension, helping in speaking turn management, and aiding in creating a friendly atmosphere [Fox Tree and Schrock, 2002]. Therefore, we believe that being able to automatically generate them is crucial to have more expressive synthetic speech. The problem is that usually in TTS, the input to the system is a text with a written style without any sort of disfluencies. So the main

question here is how to make the written text disfluent, i.e., how to insert disfluencies. Moreover, depending on the speaking style and context the required degree of disfluency might vary. For instance, speech of stressed speakers might be more disfluent than the one of relaxed speakers. So a second question is how to control the number of inserted disfluencies. Hence, in this chapter, we propose a new disfluency generation method that is able to insert several types of disfluencies.

To achieve this, we first formalize the problem as a theoretical process, where the initial text is iteratively transformed until we reach the desired level of disfluency. More precisely, the process is decomposed into a labeling problem in which sections to be edited have to be identified, and a natural language generation task in order to insert the disfluent words. This is a novel contribution since most of the previous work concentrates only on generating one type of disfluency, whereas our method is generic enough to model and generate several types of disfluencies. Similarly to our pronunciation variants work, we studied which linguistic features are useful for generating disfluencies as well. We also studied how to control the degree of disfluencies and the way it is perceived by listeners.

It is worth noting that the current work is exploratory since the formalization of the problem attempts to covers all types of disfluencies, however, the experiments have been carried out only on few types. This is mainly because corpora about disfluencies are still rare and there are no clearly defined benchmarks in the community, particularly evaluation methods are not clearly defined.

In the rest of this chapter, first, the formalization of the process is presented in Section 5.1. Then, one way of implementing the proposed method including the algorithm and machine learning techniques is presented in Section 5.2. Section 5.3 presents details of the corpus, the annotation process which is developed for our purpose, and the evaluation metrics. In Section 5.4, the study of finding which linguistic information is useful to find the position of disfluencies is given alongside its validation with the first perceptual test. Finally, Section 5.5 presents the study of controlling the degree of disfluencies, i.e., controlling the iterative process of disfluency generation, and its validation with a final perceptual test.

5.1 Formalization

The goal of this work is to define a complete process that is able to generate disfluencies for any given utterance. This underlying process has to be a deterministic one, therefore, it has to be unambiguous and clearly defined. To do so, we initially adopted Shirberg's disfluency schema as a starting point [Shriberg, 1994]. However, as described in the next subsection, this schema has some weaknesses and does not exactly fit our proposed

process. Therefore, we propose a process, that build upon this schema, with some adjustments such that it fits our goal.

In the rest of this section, firstly, Shriberg's disfluency schema is reviewed, and then we will present the proposed disfluency generation process.

5.1.1 Shriberg's schema

The schema suggested by Shriberg considers a disfluent utterance as a sequence of words in which certain words play a specific role. Thus, a disfluent utterance can be represented as ⟨A, RM, IM, RR, B⟩ where:

- A and B are two sequences of words surrounding the disfluent part of the utterance, they might also contain other disfluencies. These two regions do not exist in Shriberg's schema, however, to make the explanation process easier, they have been added here.
- RM is the sequence of erroneous words, called as the reparandum region.
- RR is the sequence of corrected words for the RM region, referred to as the repair region.
- IM is an indication of the start of an editing phase or correction phase and known as the interregnum.

In this schema, the point between the reparandum and the interregnum is called the interruption point (IP). For example, the phrase " *Show me flights from Boston on uh from Denver on Monday.*", can be split into the following regions based on the schema:

$$\overbrace{Show\ me\ flights}^{A}\ \overbrace{from\ Boston\ on}^{RM}\ \underset{IP}{\uparrow}\ \overbrace{uh}^{IM}\ \overbrace{from\ Denver\ on}^{RR}\ \overbrace{Monday}^{B}$$

Thus, based on this schema, generating a disfluent utterance from a fluent one consists in the identification of the interruption point and then the repair, reparandum and interregnum regions. This schema is supposed to be generic enough to encompass all the disfluency types including pauses, repetitions, and revisions. However, the problem arises when disfluencies are intertwined or several ones appear successively. For instance, the phrase "*I want to to uh I mean I have to go*", can be analyzed as having a repetition, two pauses, and a revision, or as having a repetition and a revision, or even as only having a revision. To solve this problem, we propose a schema inspired by this one, but adapted such that it can be used to generate disfluencies in a deterministic way.

5.1.2 The proposed disfluency generation process

In this work, we propose a complete process for disfluency generation. This process is based on the principle of composition of disfluencies on the definition of each disfluency type. In this section, we provide the general principles of this process, then we detail each disfluency type, and finally, we present details of the composition approach.

5.1.2.1 Main principles

In the process that we propose, a disfluent sentence can be seen as a result of applying a transformation function on a fluent utterance. Hence, an utterance with multiple disfluencies results from an iterative application of these transformation functions. In order to avoid ambiguity, we propose to break Shriberg's generic schema into sub-schemas, each dedicated to a disfluency type in accordance with their specific structures. We propose to consider one transformation function per disfluency type. In a general form, given a disfluency type T, we define a function f_T which takes a sequence of n words $u \in V^n$, where V denotes the vocabulary and outputs a sequence of m words as given below:

$$\begin{aligned} f_T : & V^n \longrightarrow V^m \\ & \mathbf{u} \longrightarrow f_T(\mathbf{u}) \end{aligned} \tag{5.1}$$

Hence, multiple disfluencies can be generated by composition of their functions, for example, to generate an utterance with a pause and a repetition, the repetition functions have to be combined with the pause function. Furthermore, each of these transformation functions consists of two sub-functions: σ_T, which determines the position of the interruption point, and ω_T which inserts the actual disfluent words using the result of ω_T. Mathematically, these two functions can be defined as below:

$$\begin{aligned} & \sigma_T : V^n \longrightarrow [\![0, n]\!] \\ & \omega_T : V^n \times [\![0, n]\!] \longrightarrow V^m. \end{aligned} \tag{5.2}$$

In our work, each sub-function is specific to a disfluency type. This choice has been made because interruption points may not appear in the same context across disfluency types, and we consider that each disfluency type has a specific structure. These peculiarities are described in the next sections.

5.1.2.2 Pauses

Syntactically pauses can be seen as a simple interruption in an utterance. They do not contain any reparandum and thus no repair. They can be reduced to a sole interregnum region that can be instantiated by different pause types. In our work, following the state of the art (see Chapter 1), we consider three types:

- Silent pauses – We represent them by a token **silence**. This token represents silences of any duration since we consider predicting their duration is a prosody modeling problem.
- Filled pauses – We consider short fillers represented by *uh*, and long fillers represented by *um*, as commonly admitted in English.
- Discourse markers – They are phrases that help the listener understanding the interruption in the speech. We decided to consider the phrases *you know*, *I mean* and *well* as they are the most frequent ones. Since these phrases can also be fluent words, intonation plays an important role to indicate when they are used as discourse markers. Again, this problem has not been studied as prosody is out of our scope.

To make the connection with Shriberg's schema, pauses can be summarized as follows:

RM = ∅
IM = {∗silence∗, uh, um, you know, I mean, well}
RP = ∅

In the remainder, we denote the transformation function for pauses as f_{pause} and its IP prediction and word insertion functions as σ_{pause} and ω_{pause}, respectively. Below is an illustration of application of f_{pause} on a sample utterance:

Example 1:

u : once you get to a certain degree of frustration you need to relieve

$f_{pause}(\mathbf{u})$: once you get to a certain degree of $\underset{IP}{\uparrow}\overbrace{uh}^{IM}$ frustration you need to relieve

5.1.2.3 Repetitions

Repetitions are duplications of one or few words, hence, their reparandum and repair regions are identical. Due to the proposed composition mechanism, we decided that these two regions are not separated by any interregnum region, in other words, it means that repetitions have the following structure:

$$\text{RM} = \text{RP}$$
$$\text{RM} \neq \varnothing\,,\ \text{RP} \neq \varnothing$$
$$\text{IM} = \varnothing$$

Similarly to pauses, functions related to repetitions are denoted as $f_{repetition}$, $\sigma_{repetition}$ and $\omega_{repetition}$, where $\sigma_{repetition}$ determines the position of the IP, and $\omega_{repetition}$ decides on the span of the repeated section and performs the duplication. The performed transformation is demonstrated in the following example:

Example 2:

u : and also I think this happens to a lot of people

$f_{repetition}(\mathbf{u})$: and also $\overbrace{I\ think}^{RM}\ \underset{IP}{\uparrow}\ \overbrace{I\ think}^{RR}$ this happens to a lot of people

5.1.2.4 Revisions

In a similar fashion, the revision function $f_{revision}$, uses $\sigma_{revision}$ to determine the IP, then the $\omega_{revision}$ locates the repair region and generates the reparandum region. Likewise, the interrugnum is always empty but as opposed to repetitions, the predicted reparandum is different from the repair region, i.e., :

$$\text{RM} \neq \text{RR} \ \text{ or } \ \text{RM} \approx \text{RR}$$
$$\text{RM} \neq \varnothing\,,\ \text{RR} \neq \varnothing$$
$$\text{ET} = \varnothing$$

In this work, there is no difference between false starts and other types of revisions, but this formalization could be easily extended to distinguish them. In the former case, RM and RP would have to be completely different, whereas, in the latter, they would have to be almost the same, i.e., only few words are different. An example of each case is given below:

Example 3:

u : oh god that so that if whoever won would get

$f_{revision}(\mathbf{u})$: oh god that so that $\overbrace{if\ you}^{RM}\ \underset{IP}{\uparrow}\ \overbrace{if\ whoever}^{RR}$ won would get

5.1.2.5 Composition of disfluency functions

As we stated earlier, each disfluency function is responsible for generating certain disfluency regions, therefore, the only way to generate all disfluency regions at once, or in

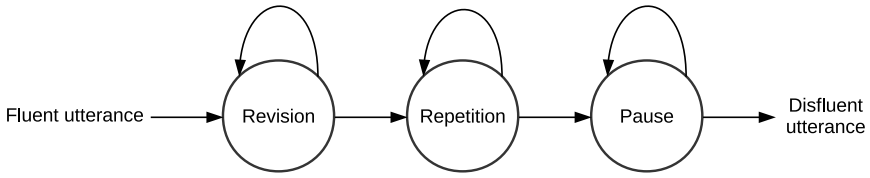


Figure 5.1: Precedence order of disfluency functions.

I have to go

[I want to I have to]$_{revision}$ go

[I want [to to]$_{repetition}$ I have to]$_{revision}$ go

[I want [to to]$_{repetition}$ [uh]$_{pause}$ I have to]$_{revision}$ go

[I want [to to]$_{repetition}$ [uh]$_{pause}$ [I mean]$_{pause}$ I have to]$_{revision}$ go

Figure 5.2: Example of composition of revision, repetition and pause functions for the utterance "*I have to go*" resulting in the disfluent utterance "*I want to to uh I mean I have to go*".

order to generate several disfluencies successively, is to combine multiple functions. In the rest of this section, we give details of how the composition of disfluency functions works.

Disfluency functions can be combined by feeding the output of one function as an input into another one. For instance, an utterance containing a revision and a pause can be seen as the result of the composed function $f_{revision} \circ f_{pause}$. However, this may be also the result of $f_{pause} \circ f_{revision}$. To minimize such ambiguities and to make the whole process deterministic, we specify the following precedence order between the disfluency types:

$$Revision \prec Repetition \prec Pause. \tag{5.3}$$

Fundamentally, this order is justified by the fact that knowing where revisions and repetitions are, can be useful to determine where to insert pauses. In practice, let us also note that it is technically easier to insert a pause in between repeated words than inserting repeated words around one or several pause tokens. Then, predicting repetitions before revisions may break the repetition phenomenon, while the contrary is not true. Even, repetitions applied on top of revisions may strengthen revisions.

Based on this order, our disfluency generation process is as illustrated in Figure 5.1. First, the revision function takes the input fluent utterance and generates revisions iteratively, then the output is given to the repetition function and to the pause function,

which respectively inserts repetitions and pauses in the same way. In order to see how this precedence works, let us demonstrate it with an example. Given the fluent utterance $\mathbf{u}$= "*I have to go*", the disfluent utterance "*I want to to uh I mean I have to go*" can be obtained by combining a revision, a repetition and two pauses, i.e., formally by computing:

$$f_{revision} \circ f_{repetition} \circ f_{pause} \circ f_{pause}(\mathbf{u}). \tag{5.4}$$

The intermediate steps of this composition are shown in Figure 5.2 where firstly, the phrase "I want to" is inserted as the reparandum region of the revision (second utterance), then the repetition function replicates the word "to" (third utterance) and finally two pauses "uh" and "I mean" are successively inserted after the repeated words (last two utterances). It is clear from this example that our method is not completely deterministic since the same disfluent utterance could have been obtained by inverting the insertion of the two pauses, that is inserting "I mean" at first, and then "uh". Still, the whole process is more compliant with practical usage than Shriberg's original schema. Especially regarding implementation, building the transformation functions and their sub-functions can be seen as machine learning problems, and it is straightforward to transform the automaton of Figure 5.1 into an actual algorithm. This is what we describe in the next section.

5.2 Implementation

In the previous section, we defined a theoretical framework for inserting disfluencies, which can be implemented in different ways. In this section, we propose one possible way to implement it. Since this is an exploratory work, the idea is to validate that the proposed method works, particularly the part about the composition of disfluency functions. Therefore, we only implement two disfluency types, i.e., pauses, and repetitions, and leave aside revisions due to complexities in generating artificial reparandums.

Hence, we propose, for each disfluency type T, to implement an IP prediction function σ_T, and a word insertion function ω_T. To do so, we formalize the IP prediction part as a labeling task achieved using CRFs and the word insertion part as the selection of the best phrase among a set of automatically built candidates, using a language model. Hence, as shown in Figure 5.3, the whole process relies on four models, two CRFs and two language models. In the rest of this section, we describe the main algorithm and details about the IP prediction and word insertion steps.

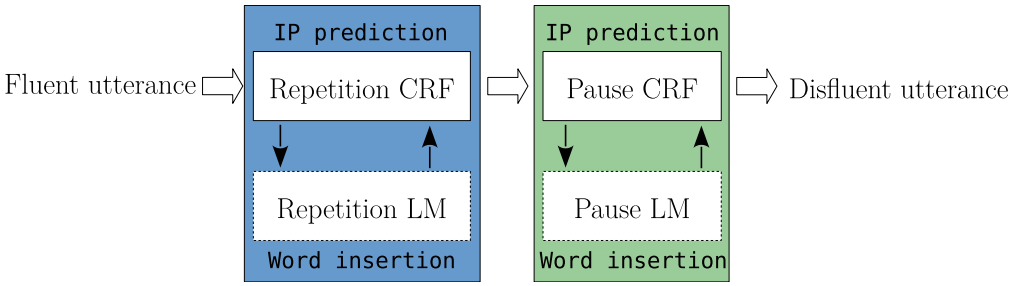


Figure 5.3: Overall methodology of the proposed disfluency generation work.

Algorithm 1: Main algorithm

input : *CleanUtterance*: A fluent utterance
output: The utterance with inserted disfluencies

```
1  type IP defined as
2      Position: integer
3      Probability: [0,1]
4  data:
5  Types: list of considered disfluency types
6  CurrentUtterance: sequence of words
7  i: IP
8  Types ← [Repetition, Pause]
9  CurrentUtterance ← CleanUtterance
10 foreach T ∈ Types do
11     i ← σ_T(CurrentUtterance)
12     while not StoppingCriterion(T, CurrentUtterance, i) do
13         CurrentUtterance ← ω_T(CurrentUtterance, i)
14         i ← σ_T(CurrentUtterance)
15 return CurrentUtterance
```

5.2.1 Main algorithm

Algorithm 1 presents how to transform an input fluent utterance into an output disfluent one. In this algorithm, we consider an IP as its position and its posterior probability as returned by an IP prediction CRF. We consider a list of disfluency types to be inspected for potential insertions. This list respects the precedence order between the disfluency types. Here, it is initialized to repetitions and pauses (*line 8*). For each type T, the algorithm tries to insert instances of that type. To do so, a candidate IP i is determined (*line 11*) using the function σ_T. The validity of this candidate is asserted by a stopping criterion (*line 12*), and, if validated, a new disfluency of type T is inserted in the current utterance at the position i by ω_T (*line 13*). After updating the utterance, a new IP candidate is generated and the process starts again. As soon as a candidate

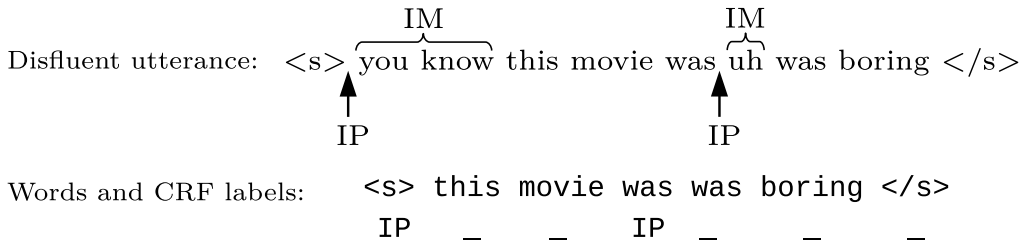


Figure 5.4: An example of assigning correspondence between words and IP labels.

is rejected by the stopping criterion, the algorithm moves to the next disfluency type. This mechanism enables multiple insertions of the same disfluency type, but it may also result in no insertion at all for that type.

The stopping criterion is a key element in the algorithm. Deciding on the insertion of disfluencies depends on the considered type T, the status of the utterance (e.g., its degree of disfluence), and the IP candidate (e.g., its position, its probability). It may even be possible that no IP candidate has been found, which obviously stops insertions. These aspects are the way to control how disfluent the output utterance will be, and we hope that this may also indirectly enable controlling how spontaneous the utterance will be perceived. By default, the algorithm stops when a certain threshold over the proportion of disfluencies is met. This threshold is set to the proportion of disfluencies observed in the Buckeye corpus, i.e., 12% for pauses, and 1% for repetitions. The definition of the stopping criterion is more deeply discussed in the last section of this chapter.

5.2.2 IP prediction

IP prediction as already stated, is achieved using CRF-based labeling. For each word of an input utterance, the CRF seeks to decide whether it should be followed by an interruption or not. In other words, and as exemplified in Figure 5.4, IP labels are reported on the words which just precede interruptions. Moreover, as shown in the figure, disfluency types are considered separately. For instance, the example focuses on pauses, thus, the repetition "was was" is disregarded and no IP is reported for it. In practice, training of these CRFs is performed on data derived from the Buckeye corpus, where words come along with linguistic features. These aspects are addressed in Section 5.3.

Once CRFs are trained, their usage during the generation process is given by Algorithm 2. Given an utterance, the objective is to find a new IP, i.e., an IP whose position has not already been used to insert a disfluency of the same type as the one currently considered (T). This is an important point since CRFs may tend to infinitely predict

Algorithm 2: IP prediction function σ_T

```
1  function σ_T:
       input  : Utterance: An utterance
       output: An IP or NO_IP_FOUND
2      data:
3      CRF_T: CRF model for the type T
4      N: integer     /* number of hypotheses requested from the CRF */
5      H: list of lists of IPs  /* N-best hypotheses returned by the CRF */
6      IpFound: boolean
7      N ← < USER DEFINED VALUE >
8      H ← GenerateHypotheses(CRF_T, N, Utterance)
9      IpFound ← false
10     while not IpFound do
11         h ← Shift(H)          /* removes and returns first item */
12         Sort h according to descending order of probabilities of each IP
13         while h ≠ ∅ and not IpFound do
14             Ip ← Shift(h)
15             if Ip ∉ Utterance then
16                 IpFound ← true
17     if IpFound then
18         return Ip
19     else
20         return NO_IP_FOUND
```

IPs on some likely positions. To maximize the number of IP candidates for a given input utterance, several labeling hypotheses are generated by the CRF of the considered type (*line 8*). Each hypothesis may contain zero, one, or several IP candidates along with their posterior probabilities. The number N of requested hypotheses is defined by the user. A small value will lead to a low number of IP candidates, thus, influencing the stopping criterion of the main algorithm. On the contrary, high values lead to large sets of candidates, but last hypotheses' candidates can be unlikely, leading to wrongly positioned disfluencies. In order to select the returned IP, hypotheses are browsed one at a time. For each hypothesis, IP candidates are sorted according to their probability and iterated until finding a candidate which is not already in the utterance (*line 15*). If ever no new candidate is found, the algorithm propagates this information to the main algorithm

Algorithm 3: Word insertion function ω_T

```
1 function ω_T:
      input  : Utterance: An utterance to be modified
               Ip: A predicted IP
      output: Input utterance with inserted disfluency
2     data:
3     LM_T: language model for type T
4     Candidates: set of utterances
5     Scores: map from utterances to real numbers
6     Candidates ← BuildCandidates(T, Utterance, Ip)
7     foreach C ∈ Candidates do
8         Scores[C] ← ComputeScore(LM_T, C) × length(C)
9     return arg max_{C∈Candidates} Scores[C]
```

5.2.3 Word insertion

Based on the predicted IP position, the word insertion step inserts disfluent words depending on the disfluency type, that is the RM region for repetitions and the IM region for pauses. This is performed in a two fold way (Algorithm 3). First, disfluent utterance candidates are generated for the considered type (*line 6*), and these candidates are scored using a language model (*line 8*). At the end, the candidate with the highest score is returned.

Candidate for repetitions are built by considering the IP, and consider several repair regions to be duplicated. In practice we limit these possible regions to one or two words after the IP[1]. For instance, considering that our input utterance is:

$$\text{I would like } \underset{IP}{\uparrow} \text{ to have a coffee,}$$

the possible repetitions are:

I would like **to to** have a coffee.

I would like **to have to have** a coffee.

For pauses, candidates are built by inserting the six possible tokens already introduced

[1] This choice has been made since repetition of longer words does not happen in our corpus. Obviously, it could easily be extended to enable longer repetitions.

in 5.1.2.2. For the below input utterance:

$$\text{I would like to } \underset{IP}{\uparrow} \text{ to have a coffee,}$$

the resulting candidates are:

I would like to ***silence*** to have a coffee.
I would like to **um** to have a coffee.
I would like to **uh** to have a coffee.
I would like to **you know** to have a coffee.
I would like to **well** to have a coffee.
I would like to **I mean** to have a coffee.

Scoring is achieved by computing a probability of a window of words around the inserted token. Focusing on a window rather than the whole utterance will let us see more easily the impact of the inserted disfluency on its local context. In practice, three context words on the left and three on the right are considered, and the language models used are 3-gram models. To prevent short word sequences from obtaining higher probabilities than longer ones, scores are normalized according to the length of the word sequence (*line 8*).

This implementation has been applied on the Buckeye corpus. The data preparation and the experimental setup are given in the next section, while the results presented in Section 5.4.2.

5.3 Corpus preparation and experimental setup

Our work on disfluencies is conducted on the same 20 speakers of the Buckeye corpus used in Chapter 4. The data is firstly annotated in order to detect and mark disfluency sections. Then it is augmented with additional features before deriving related datasets for each disfluency type. Finally, the obtained results are evaluated using objective and subjective measures. The rest of this section describes these four points.

5.3.1 Annotation

The goal of annotation was to mark disfluency sections in the sense of Shriberg's schema, i.e., reparandum, interregnum, and repair. This annotation has been automated as much as possible with manual checking when required. Repetitions were automatically annotated by spotting successively duplicated word sequences or those separated by one of

the six possible pause tokens. Similarly, pauses where the IM is "uh", "um" or "*silence*" have been automatically marked. Other pauses (you know, I mean, well) required manual disambiguation since these phrases can be used outside of disfluencies in English. Finally, revisions are the most complex disfluencies, annotating manually would have been too expensive. As a consequence, a semi-automatic process was adopted, where all phrases containing a pause were automatically considered as a revision candidate, then candidates were automatically discarded or selected. In the case of selection, the annotator manually marked the RM and RR sections. No specific annotation for IPs has been added since it is straightforward to derive this information from the disfluency sections.

As a result of this process, 20264 pauses and 2714 repetitions were annotated for the 20 speakers. Regarding revisions, only 12 of the speakers were annotated, resulting in 203. Even if revision are disregarded in our current implementation, it is clear from this low number, that training good models would be a difficult task.

5.3.2 Features

Beside disfluency annotations, words come along with linguistic features, specifically we consider the set of features given in Table 5.1. These features are mainly composed of information about the position of the word in utterance, while other information relates to its frequency, nature, structure, etc. Additionally, if the word is part of a disfluency, information is given about the type of the disfluency. Obviously, this information is handled carefully at training time, such that it is not used to predict the IP of that particular disfluency.

5.3.3 Disfluency-specific datasets

Since the proposed method relies on independent processing of disfluency types, specific datasets have been derived for each type from the global Buckeye corpus. Hence, utterances containing repetitions have been selected to train the repetition models. Similarly, utterances containing pauses are used to train the pause models. As a consequence, remaining utterances have been discarded. Furthermore, to be in accordance with the precedence order over disfluency types, utterances of repetition-specific dataset should not contain any pause. Thus, all pauses have been removed from them.

Language models used for the word insertion step are directly trained on these datasets. On the contrary, the training data for the IP prediction CRFs require further processing. The key idea here is to process IPs one at a time. Since CRFs should be able to predict IPs, both from fluent and partially disfluent utterances, several training sequences are derived from each IP as shown in Figure 5.5. This mechanism is carried

Table 5.1: List of features used for disfluency prediction.

Feature	Values
word	any word
POS	part of speech tag
is a stop word?	*true*, *false*
is part of a disfluency?	no, pause, or repetition
word frequency	*high*, *medium*, *low*
word occurrence count	integer
number of time the word is repeated afterwards	integer
number of syllables in the word	integer
word length	integer
absolute position	integer
absolute reverse position	integer
position in three categories	*beginning*, *middle*, *end*
position in five categories	1 .. 5
position in ten categories	1 .. 10
is one of the first three words?	*true*, *false*
is one of the first five words?	*true*, *false*
is one of the first ten words?	*true*, *false*
is one of the last three words?	*true*, *false*
is one of the last five words?	*true*, *false*
is one of the last ten words?	*true*, *false*
is one the first five or last five words	*first 5*, *middle*, *last 5*

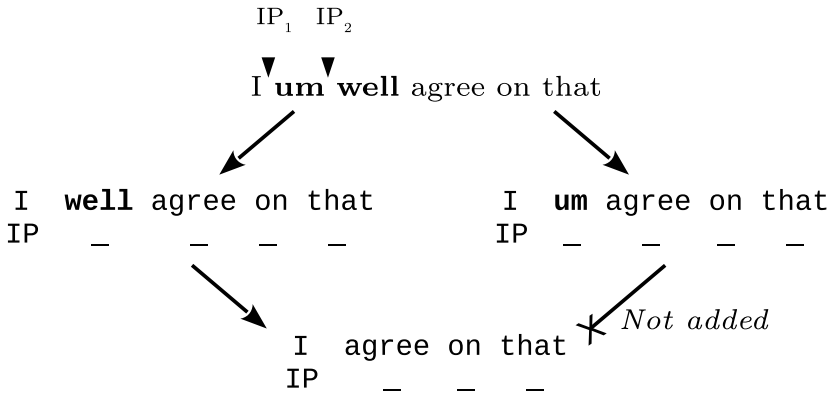


Figure 5.5: Example of deriving training sequences from individual IPs.

out in such way that no duplicated training sequences are produced. As the result of this process, 2684 utterances were extracted for pauses and 1125 for repetitions

5.3.4 Evaluation methodology

Disfluency generation work has completely been conducted in a speaker-independent way, that is utterances from different speakers are mixed in our datasets. The reason is that the amount of disfluent words is rather small, so it would have been difficult to train speaker-dependent models. As stated in the previous section, four datasets are considered, i.e., two datasets for language models (one for each disfluency type) from which two others are derived for CRFs. All these datasets are partitioned in the same way into a training set (60% of the utterances), development set (20%), and a test set (20%).

In our experimental setup, the method is evaluated after each step. The IP prediction step is evaluated using recall, precision, and F-measure over the predicted IP labels (*cf.* Section 3.4.1). Precision can be meant as the accuracy of our method to insert IPs at the same exact position as the speakers in the corpus, while recall is related to the number of these reference IPs that are inserted by the method. Nonetheless, as already shown in [Dall et al., 2014], one reference can not be considered as a reliable ground truth in our problem, since IPs predicted by the method might be perfectly valid, whereas not in the reference. To overcome this problem, we propose a new metric referred to as *IP ratio*. Given a set U of fluent utterances, it compares the proportion of IPs in disfluent utterances predicted by our method with respect to those from the reference. Formally, given a fluent utterance $u \in U$, a disfluent utterance derived from u can be reduced to its set of IPs. If this set is noted as $\mathcal{X}_u$, we can define the degree

of disfluency d of that disfluent utterance as follows:

$$d(\mathcal{X}_u) = \frac{|\mathcal{X}_u|}{|u|}, \tag{5.5}$$

where $|.|$ refers to the cardinality of the set of IPs and of the word sequence. Then the IP ratio of U is defined as the function of the average difference over the degree of disfluency between a tested hypothesis and the disfluent reference, as formulated below:

$$\textit{IP ratio}(U) = 1 + \frac{\sum_{u \in U} d(\mathcal{H}_u) - d(\mathcal{R}_u)}{\sum_{u \in U} d(\mathcal{R}_u)}, \tag{5.6}$$

where $\mathcal{H}_u$ and $\mathcal{R}_u$ are the sets of IPs in the hypothesis and the reference respectively. IP ratio is normalized according to the degree of disfluency in the reference and centered on 1. Thus, any value between 0 and 1 denotes an under-prediction, for instance, 0.3 means 70% less predicted IPs in the hypothesis compared to the reference, while 1 means that they have the same exact number of IPs, and values greater than 1 mean over-prediction, e.g., , 2 means an over-prediction by 100%.

Regarding word insertion step, the returned disfluent utterances are measured in terms of perplexity. This measure returns high values for completely unlikely word sequences, for instance, utterances where disfluencies have been inserted in unusual places, thus strongly breaking the syntax, or the inserted tokens badly fit the sentence, or the proportion of predicted disfluencies is far from those in the reference. In practice, perplexities are given by the language models built on the training set. Since the word insertion function also relies on language model probabilities, it is logical that utterances with a high proportion of disfluencies will get low perplexities. Hence, it only makes sense to compare perplexities of utterances with IP ratios of the same range.

All these objective measures have been used to configure and study various parameters of the whole method. Still, these measures cannot replace human judgments, therefore, perceptual tests have also been conducted to complete the objective results. Details about their protocol will be given along with their results.

5.4 Training of CRFs

The goal of this section is to provide an initial validation of the proposed method. This goes mainly through the IP prediction CRFs the validation of the generated disfluent utterances by the process. To perform validation in a rigorous way and to get a fine understanding of the results, experiments on each disfluency type have been carried out separately. In the remainder of this section, details of the tunning process of CRFs are given, objective evaluation of these models is performed on the test set, and perceptual

Table 5.2: Repetition generation experiments on the development set.

Feature	Window	Recall (%)	Precision(%)	F-measure	IP ratio
Word	None	3.7	4.2	3.9	0.5
Word	Best	4.6	**19.4**	7.5	0.4
Word + POS	None	3.3	7.0	4.5	0.3
Word + POS	Best	**6.8**	18.1	**9.9**	0.5

tests are conducted to validate the whole proposed process.

5.4.1 Feature and window size selection

Two main parameters in CRF training are the features used to describe the words and the size of the context around each word. These two aspects have been studied on the development set to investigate which aspects are important to train reasonably good CRFs. No fully automated approach has been followed. Instead, several possibilities have been tried and decisions have been manually made by observing the results of objective measures.

Considering the word as the mandatory feature, each time a new features is added to the list of selected features to find the best complementary one. Then, the idea is to repeat this process iteratively until no feature improves the CRFs. Regarding context size, optimization has been performed on windows of different sizes instead of single words. Precisely, windows are made of few words (1 up to 3) before and after each word under examination by the CRFs.

As shown in Tables 5.2 and 5.3, the overall precisions and recalls are low. Still differences can be noted. Conclusions of the feature selection process are that small feature sets perform better than large ones. This is probably due to the small amount of data. More precisely, words and POS seem to be the best set of features. Then, it appears that considering windows always improves the results. For repetitions, the best configuration is to include one word from the left, and three words from the right, while for pauses, the best window is one word from the left and none from the right. The results also show that all models are under-predicting disfluencies. This is can be particularly observed on repetitions where the best models under-predicts by 50%. These configurations are those selected for validations on the test sets. In addition to these ones, the feature telling whether a word was originally in the fluent utterance or has been inserted during the process will be studied. The reason is that, we think this feature can help in predicting new disfluencies, for instance, predicting a pause in the neighborhood of a repetition.

Table 5.3: Pause generation experiments on the development set.

Feature	Window	Recall (%)	Precision (%)	F-measure	IP ratio
Word	None	7.7	27.6	12.1	0.5
Word	Best	**12.1**	23.1	**15.9**	0.7
Word + POS	None	5.5	**36.1**	9.5	0.4
Word + POS	Best	11.8	24.4	**15.9**	0.7

Table 5.4: Repetition generation experiments on the test set.

Feature	Window	Rec.	Prec.	F-meas.	IP ratio	Perplexity
Fluent utterances					0.0	241
Disfluent reference utterances					1.0	236
Word	None	0.8	3.8	1.3	0.1	236
+ POS	Best	**6.2**	**17.1**	**9.2**	0.4	**231**

5.4.2 Objective evaluation

Initial experiments on repetition and pause generation are presented. CRFs are trained on the features and window sizes selected in the previous section. Apart from recall, precision, F-measure and IP ratio, perplexities are now also used to validate the final disfluent utterances. CRFs and language models are trained on the training sets and results are produced on the test sets. Repetitions have been generated from fluent utterances, while pauses have been generated on top of utterances which may contain repetitions from the corpus, i.e., reference repetitions and not those which have been predicted by a previous iteration of our method. This will allow us to validate the performance for repetitions and pauses separately.

The experiments have been conducted using several configurations. The first one uses the simplest CRF model with only the word as a feature and no window. The second model has been trained with the word and POS features plus the selected window sizes. Lastly, only for pause experiments, we tested the additional configuration which includes information about previously predicted disfluencies. The results for repetitions are presented in Table 5.4 and for pause in Table 5.5.

First, the results observed on the test set are consistent with those of the development set. This tends to validate the feature and window size selection. Then, generated utterances are compared with the fluent and disfluent reference utterances. For both repetitions and pauses, fluent utterances always have the highest perplexity, showing that this metric can be seen as a good way of assessing the quality of disfluent utterances. Regarding repetitions, perplexity variation is low[2], still, the best result is achieved by

[2]We think that this low variation is because the number of repetitions in the language model training set is low and repetitions can have many different forms as opposed to pauses, where only few tokens

Table 5.5: Pause generation experiments on the test set.

Feature	Window	Rec.	Prec.	F-meas.	IP ratio	Perplexity
Fluent utterances					0.0	242
Disfluent reference utterances					1.0	**172**
Word	None	8.2	29.4	12.8	0.5	209
+ POS	Best	17.9	33.6	23.3	0.7	191
+ Prev. pred. disf.	Best	**19.8**	**34.5**	**25.1**	0.7	188

our optimized setting. Nonetheless, it is difficult to entirely conclude on this point, since the IP ratio shows that this setting under-predicts repetitions by around 60%. The consequences of this under-prediction can be clearly seen on the recall numbers since the best repetition model is able to only retrieve 6.2% of the reference repetitions correctly.

On the side of pauses, our predicted utterances also brings low perplexities, while not outperforming this time the disfluent reference. However, the IP ratios are again low. This highlights the need for controlling the proportion of inserted disfluencies as studied in Section 5.5. The last configuration where previously predicted disfluencies are added as a feature, brings a significant improvement. Despite the fact that his latter model under-predicts pauses by 30%, it is able to retrieve 19.8% of the reference pauses correctly with a precision of 34.5%. This is all the more interesting that this feature was not selected from the development set. This proves that there is a dependency between prediction of disfluencies. Moreover, this highlights that the objective measures are not reliable enough to assess the quality of disfluent utterances, and that perceptual tests are needed.

Overall, it can be said that the results for both repetitions and pauses are not good enough, however, as it has been suggested in the literature, this might be due to the low number of disfluencies used to train our models. For instance, in [Tomalin et al., 2015], a lattice-based approach of pause insertion is presented, where their models are trained on 20M words as opposed to 150K in our case. Although their approach performs well in general, for certain pause types, like “AH”, “HM”, “UHU”, and “UHUM”, where the training data is small, very low scores are achieved. This proves that having a large training data is critical for generating disfluencies and in this perspective our results are comparable with those of the literature

In the next section, a perceptual evaluation of our disfluency generation models is presented.

are possible.

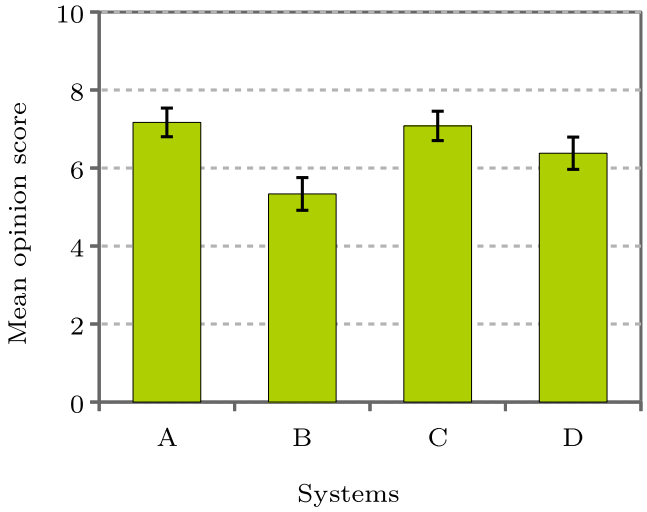


Figure 5.6: Mean opinion scores for tested systems for repetitions. Confidence intervals are given for $\alpha = 0.05$.

5.4.3 Perceptual tests

We believe that even if models predict disfluencies at wrong positions, there is still a chance that some of these disfluencies are possible considering a spontaneous conversation. Therefore, two MOS tests were conducted for repetitions and pauses separately with 24 non-native subjects using 40 utterances sampled from the test set[3]. These utterances are the same when testing repetitions or pauses. Utterances were presented to the subjects in their textual form and no speech synthesis has been used here. Since our TTS system has not been trained to synthesize disfluencies, highly disfluent utterances would have been penalized against less disfluent ones. At each step, the subject was presented with a fluent utterance and several other utterance propositions, and asked the following question:

> "Imagine someone tells the text below during a spontaneous conversation. How likely do you judge the following spoken propositions?"

Subjects could score proportions on a 10 degree scale (0=impossible, 10=perfectly possible).

Results on repetitions are presented by a bar chart in Figure 5.6. Tested systems are the same as in Table 5.4 and are noted as: (A) fluent utterances, (B) reference utterances, (C) utterances generated with the simplest model (word only), and (D)

[3]The 40 utterances have been selected from the test set such that the reference utterances contain a mixed number of repetitions and pauses. Very short and long utterances were discarded.

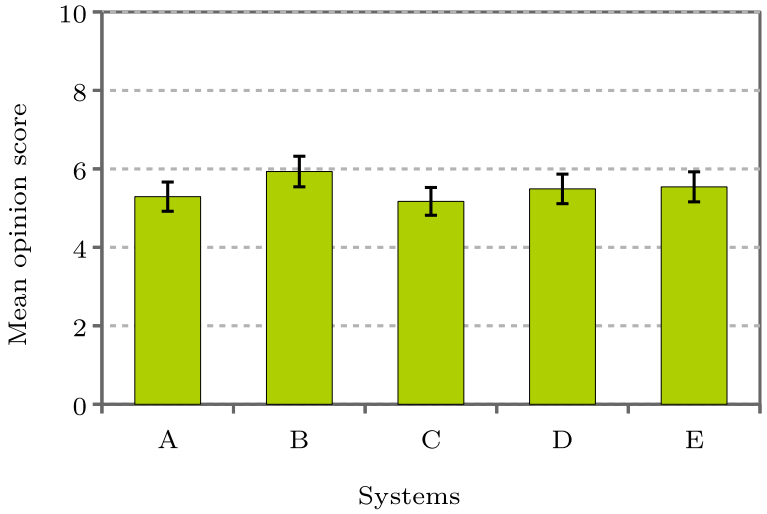


Figure 5.7: Mean opinion scores for tested systems for pauses. Confidence intervals are given for $\alpha = 0.05$.

utterances generated using the optimized model (words+POS+window). The results are surprising since the reference (B) has received the lowest score. The next lowest score is from the optimized model (D), which over-generates repetitions. In contrary, the other two systems with no repetitions (A) and with a small amount of repetitions (C) have received the highest scores. These results can be interpreted such that the lower a system generates repetitions, the most possible it is perceived. We think that this surprising conclusion is due to the fact that presented utterances do not contain any pause, which make repetitions look less natural.

Next, the results of the pause perceptual test are summarized in Figure 5.7. The tested systems are (A) fluent utterances in the sense that they do not contain any pause, (B) reference utterances, (C) utterances generated with the simplest model (word only), (D) utterance generated using the optimized model, and (E) utterances generated using the model accounting for the previously predicted disfluencies.

First, it is worthing noting that the range of the results here is even more narrow than those of repetitions. The only statistical differences are between systems A/C and system B. The difference between A and B confirms that including pauses confers a more spontaneous style to utterances, while the difference between B and C demonstrates that our simplest model is not enough to properly integrate pauses. Then, systems D and E seems to perform better, but no definitive conclusion can be drawn.

From these first series of experiments, we conclude that the overall method is effective in transforming fluent utterances into disfluent ones. Particularly, adding features

and context to IP prediction CRFs seems to be necessary. Then, the process should be improved by systematically combining repetitions with pauses. Finally, some limitations can be highlighted. Among them, the evaluation of disfluent utterances is a difficult task for humans, probably because utterances at most only differ by very few words. Further, depending on the configuration, systems generate utterances with greatly different IP ratios, which makes their comparison difficult. These limitations lead us to the questions on how to make differences stronger and in a controlled manner.

5.5 Controlling spontaneousness

The degree of disfluency may vary in spontaneous speech. Some speakers may be very disfluent, e.g., if they are stressed, disturbed by their interlocutor, or they are not very confident about the topic of conversation. The goal of this section is to study how to control the number of predicted disfluencies to be able to model such situations. In our proposed method, this depends on the criterion used in our algorithm to stop the insertion of disfluencies (*cf.* Section 5.2.1). In this section, we first discuss different possible stopping criterion and study their properties, before evaluating how different controlled degrees of disfluency are perceived.

5.5.1 Stopping criteria

Different reasons can be given to stop insertion of disfluencies in our algorithm. Among possible reasons, given that our algorithm picks IPs among a set of candidates, a first one can be attributed to (i) the fact that there are no more candidates to consider, or (ii) it may also be decided because the considered IP candidate is not good enough according to the IP prediction model, lastly, (iii) one may want to stop the algorithm when a sufficient amount of disfluencies have been inserted. In this work, we limited our stopping criterion to these possible reasons, since they seem to be the most relevant ones to our mind.

We have identified the underlying parameters behind each of these reasons and studied how different values of these parameters impact the number of inserted disfluencies. In practice, this means to examine the correlation between these parameters and the IP ratio.

First, the size of IP candidates can be controlled by varying the number of hypotheses requested to the IP prediction CRFs. Figure 5.8 shows the IP ratios of generated disfluent utterances when setting different values for this number of hypotheses. It appears that there is almost no effect in changing the number of hypotheses, except when only one or two hypotheses are considered. Then, even when limiting to one hypothesis, there is no way to generate utterances with low number of disfluencies, as it appears

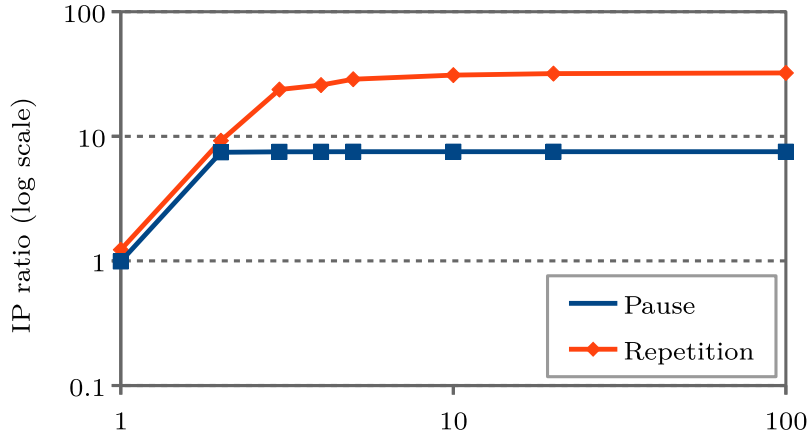


Figure 5.8: Correlation between number of hypotheses and IP ratio.

that IP ratio never reaches values lower than 1. As a conclusion, there is apparently no reason to limit the number of hypotheses. Thus, in the further experiments, the number of hypotheses is set to 100.

Second, the quality of the predicted IPs can be controlled by examining their posterior probabilities returned by CRFs and stopping as soon as an IP with a probability lower than a given threshold is encountered. This controlling mechanism does not consider sequence-level probabilities but rather those of individual IPs since a sequence with a high probability might still contain IPs with very low probabilities. Hence, the main idea here is to not include poor quality IPs. The same idea could have been applied to filter out those IPs instead of stopping as soon as the first low probability IP is met. Figure 5.9 presents how setting different thresholds on IP probabilities impacts the IP ratios. The results suggest that this threshold has some sort of log-linear effect on the IP ratios. When this threshold is put at around 0.5, IP ratio reaches 1, which means that the reference and the generated utterances have the same proportion of IPs. However, the figure cannot tell us about the real impact of filtering poor quality IPs on the quality of the final disfluent word sequences. This latter point is addressed through perceptual tests in the next section.

Lastly, the degree of disfluency can be directly measured and controlled while building disfluent utterances. By setting a maximum threshold on this degree of disfluency, the algorithm can be stopped as soon as this threshold is reached. Having such kind of threshold can be used particularly when expert or empirical knowledge about the style or the context of the speech is available. For instance, in our datasets, the proportion of repetitions and pauses are 1% and 12% respectively. In Figure 5.10, results of the IP ratios are shown when setting different values on this threshold. It logically appears

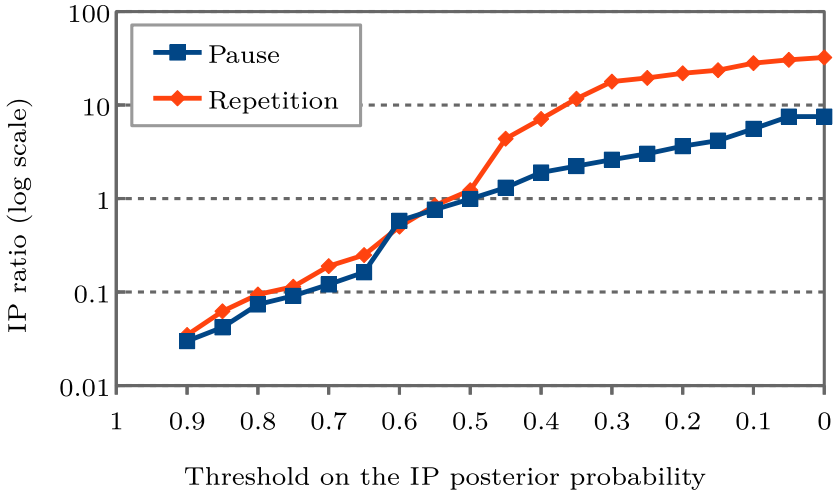


Figure 5.9: Correlation between threshold on posterior probabilities and IP ratio.

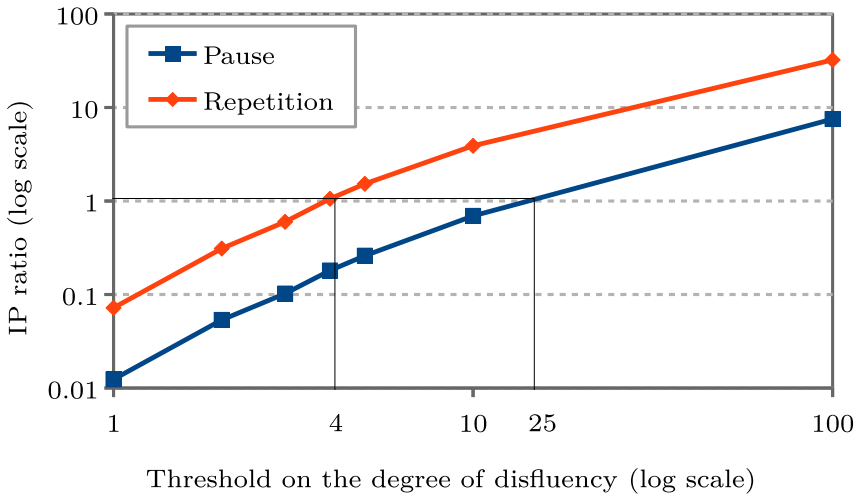


Figure 5.10: Correlation between threshold on degree of disfluency and IP ratio.

that there is a linear correlation between these two parameters. It can be seen that the IP ratio is close to 1 when the threshold is around 4% for repetitions and 25% for pauses. It means that for a given target degree of disfluency, the threshold on the maximum has to be set to a much higher value.

To sum up, the presented studies show that some of the parameters have direct correlations with the IP ratio, particularly, the probability of IPs and degree of disfluency. This information can be used to generate utterances with varying degrees of IP ratios. However, in complementary experiments we conducted, it appears that when different CRF experimental setups are used, these correlation factors change. This means that thresholds have to be determined according to the CRF setup, especially when the number of features varies. Overall, it is clear that further investigation should be conducted

to find out more generic stopping criteria.

In the next section, we will use the results obtained here to generate disfluent utterances with different degrees, and examine how these different disfluency degrees are perceived.

5.5.2 Perception of the degree of spontaneousness

We conducted a MOS test aiming at examining the perception of different disfluency degrees. Utterances with different IP ratios were generated by fixing the number of hypotheses at 100 and setting different threshold values for the two other parameters from the previous section. Three levels of disfluency degrees were considered for repetitions and pauses separately: zero, medium, and high. "Medium" denotes an IP ratio of 1, while "high" stands for respective IP ratios of 4 and 2.5 for repetitions and pauses. All possible combinations of these levels were tested, leading to nine systems.

The test was conducted with 26 non-native speakers using the same 40 utterances (in textual form) that were used in Section 5.4.3. At each step, the subject was presented with a reference utterance and utterances proposed by the nine systems and asked the same question as in the previous perceptual tests.

Results are presented by the bar chart of Figure 5.11. Similarly to our previous perceptual tests, differences between the tested systems are not statistically significant. The most representative example of this is that the disfluent reference utterances and the fluent ones are perceived in a similar way. Even though no obvious trends can be seen across systems, it looks like systems having a medium amount of pauses, i.e., B, E, and H, are judged slightly more acceptable than those of zero and high amounts. Further, utterances with a high amount of repetitions and pauses (system I), seems to be perceived as the most unlikely. In spite of this evaluation issue, manual observations show that the generated disfluent utterances are reasonably good. As illustrated in Table 5.6, pauses and repetitions are inserted in plausible places in the automatically generated utterances (E and I).

The main conclusion here is that such a test might be too difficult for the testers. This evaluation issue should definitely be more deeply investigated. For instance, it would be interesting to conduct the test with linguists. Alternatively, the asked question about the plausibility of the utterances should be replaced with a more direct one about the perceived degree of spontaneousness. AB tests could also have been done, permitting one to one comparison of the systems. Finally, although we intentionally decided to not synthesize utterances, we can now wonder if this would not ease such a test.

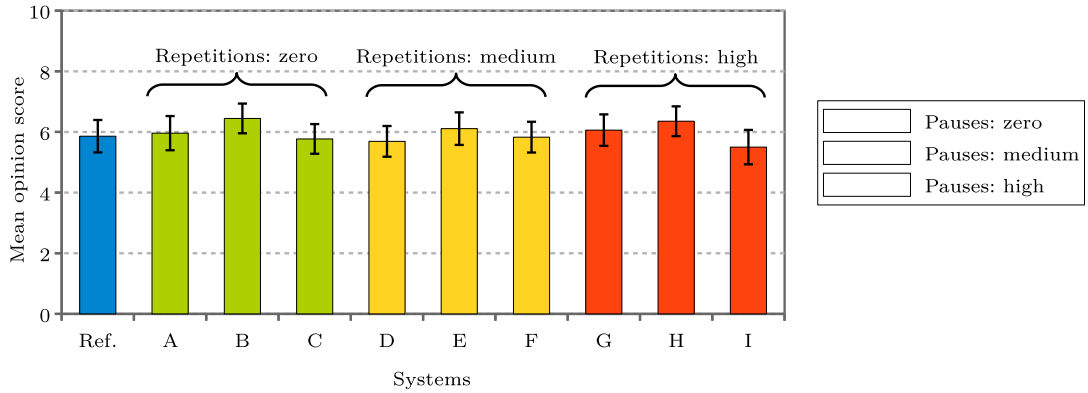


Figure 5.11: Mean opinion scores for tested systems. Confidence intervals are given for $\alpha = 0.05$.

Table 5.6: Different versions of the same utterance: disfluent reference, fluent (A), generated disfluent ones with medium (E) or high (I) degree of disfluency.

Ref.	yeah in fact he **uh** let's see I ***silence*** I suppose I don't remember
A	yeah in fact he let's see I suppose I don't remember
E	***silence*** yeah ***silence*** in fact he **uh uh** let's see I suppose ***silence*** **I don't I don't** remember
I	***silence*** yeah ***silence*** in fact he **uh uh** let's see ***silence*** I suppose ***silence*** **I don't I don't** remember

5.6 Conclusion

To sum up, in this section we presented a disfluency generation method. After studying the usually admitted Shriberg's disfluency schema, we derived a new process that formalizes disfluency generation. The main idea is to separately model disfluency types and to compose them. For each type, we firstly predict the IPs and then insert disfluent words at the spotted positions. A proof of concept of this process was given through an implementation based on CRFs and language models.

The main originality of the work lies in the fact that it can generate multiple types of disfluencies. Then, the proposed formalization clearly highlights different components, which can be independently improved, for instance, by using most recent machine learning techniques, which was not the aim of our work. Moreover, it makes it possible to study some particular aspects of the generation process. The preliminary results also indicate that it could be possible to control the degree of disfluency.

Limits and perspective can be drawn out of this exploratory work. First, results have emphasized that weaknesses in the perceptual evaluation, since it was difficult for the testers to feel the differences between the utterances. Then, the size of the datasets used

to train models was probably too small for rare disfluencies like repetitions or discourse markers. This is particularly true for revisions. Although they were formalized, revisions were too complex and too few to consider. Thus, the main perspective concerning the method is to integrate them.

General conclusion

In this thesis, we addressed the issue of expressivity in speech synthesis. Since expressivity covers a wide area, the main focus was put on spontaneous speech with emphasis on pronunciation variants and speech disfluencies. In this direction, the first two chapters of the thesis gave a background of the problem area and a survey of the previous studies, while the last three chapters described the corpus and exposed our contributions. In the following two sections, we first provide a summary of the main findings and contributions of the thesis, then, some perspectives for the continuation of the pronunciation variants and disfluency work are discussed.

5.7 Summary of the thesis and contributions

The first chapter of this thesis was devoted to providing an overview of the basics of speech and expressivity. Starting by explaining the human speech production mechanism, we then described some of the important concepts which are believed to be of significant impact for expressivity like phonetics, phonology, and prosody. We finished the first chapter by studying the impacts of expressivity on pronunciation and speech fluency. In the second chapter, mainly the different ways of exploiting expressivity to make speech applications more natural and expressive were studied. In addition, we surveyed some of the previous works in the area of pronunciation variants and speech disfluencies. The third chapter described the data and the evaluation methodology used for generating pronunciation variants and speech disfluencies. A statistical analysis of the data was also provided detailing different aspects like most common types of variations and disfluencies.

The last two chapters of the thesis were devoted to presenting our main contributions on pronunciation variants and speech disfluencies respectively. In the fourth chapter, we proposed a new method for generating pronunciation variants by adapting standard phoneme sequences to spontaneous ones. Along with this process, we studied various factors that influence the performance of this adaptation such as related features, context information, phoneme dependency and different ways of training

adaptation models, i.e., speaker-dependent and independent adaptations. The method was first tested in the context of spontaneous speech adaptation and then we showed that it can be extended to other adaptation tasks as well, for instance, to solve the problem of inconsistency between phoneme sequences handled in TTS systems. Objective and subjective evaluations of the method on spontaneous speech showed that adapted spontaneous pronunciations using a combination of features significantly better reflect spontaneous speech than standard ones. Moreover, the phonological reranking had a great impact in bringing the pronunciations closer to spontaneous speech. These initial experiments allowed us to validate the effectiveness of the proposed method and led us to corpus-specific adaptation experiments. Similarly, the results here also showed that the proposed method brings a significant improvement in terms of PER. Likewise, perceptual tests showed an improvement in the quality of generated synthetic speech. Hence, we showed that pronunciation adaptation helps to reduce inconsistencies between phonemes labeled by their underlying speech corpus and those generated by the phonetizer during synthesis.

In the last chapter, we gave details of an exploratory but novel disfluency generation approach which, unlike most of the previous work in the literature, was able to generate several types of disfluencies. An algorithm was developed for this purpose which relied on an improved version of Shriberg's schema, where the disfluency generation task was broken down into the application of a series of transformation functions, each specific to a disfluency type. This approach provided the advantage of generating disfluencies in a more deterministic way. To make the method perform as desired, we prepared the data in a unique way such that several training samples could be generated from a single disfluent utterance. Then the method was tested extensively with respect to various aspects such as the accuracy and quality of inserted disfluencies and their positions as well as the impact of the stopping criteria. On the one hand, initial objective evaluation results showed that repetitions achieved very low scores, while pauses models performed better in general. Further, the results also showed that adding features and context size to prediction models improve their performance. On the other hand, perceptual tests demonstrated that evaluating disfluent utterances is a difficult task for humans. This is mainly because most utterances only differ by very few words, and they were presented in their textual form. Lastly, our studies on stopping criteria yielded the important parameters and showed what sort of values should be assigned to them in order to generate utterances with varying IP ratios.

5.8 Perspectives

Although both contributions provide a useful insight into the different ways of exploiting expressivity for speech synthesis, we believe both works could be improved in various ways. In the following sections, proposed perspectives on pronunciation variants and disfluencies, as well as more general ones are provided.

5.8.1 Pronunciation variants

One of the main advantages of our pronunciation variants work is the utilization of a wide range of features. However, among those features, prosodic ones were extracted in an oracle manner, leading to optimistic adaptation results. As stated in the manuscript, generating this kind of features is still a research problem and was out of our scope. Thus, an interesting perspective would be to test the effect of prosodic features when they are generated using a prosodic feature predictor. Of course we do not anticipate them to have the same effect as extracting the oracle ones, however, this will still bring some more insights into questions concerning their effectiveness.

Furthermore, the current method uses a fixed number of hypotheses for the phonological reranking. Thus, adaptation results when considering few hypotheses for long utterances are limited. Similarly, having many hypotheses degrades the results in the case of short utterances. A reasonable trade-off would be to consider a variable number of hypotheses based on the length of the utterance. That is, to take few hypotheses when the utterance is short, and to take a larger number of hypotheses in the case of long utterances. This will enable better handling of N-best hypotheses and hopefully lead to better results.

5.8.2 Speech disfluencies

One of the main challenges when working with speech disfluencies was the low number of disfluent utterances in the corpus. This problem became more obvious during the objective evaluations as we were trying to measure how the models perform in finding the exact position of IPs. A larger corpus of disfluencies would let CRF models better learn patterns to predict IP positions and let the language models find the best fitting disfluency with the places of the IPs. Another perspective is to complete the annotation of revisions and integrate them into the current method. This would permit to fully validate the proposed disfluency generation method as well as to enrich generated utterances with even more disfluency types and thus, making them more expressive.

5.8.3 Common perspectives

The first perspective that concerns both works is the possibility to combine the results of both works. This can be achieved by generating disfluencies for a given utterance and then passing it through our pronunciation adaptation process. This way, we can generate even more expressive synthetic speech. Further, more robust evaluations have to be considered. In the pronunciation variants work, adapted pronunciations were perceptually evaluated using an HTS trained on non-spontaneous speech data. This led to several phoneme inconsistencies since the HTS system was not able to synthesize all the existing adapted spontaneous pronunciations. In our disfluency work, the perceptual tests did not even include the generated speech, forcing testers to imagine the utterances and judging them accordingly. For both works, it could be interesting to test the generated speech samples using actual spontaneous speech data and examine how this would affect the results. A final perspective concerns the definition of expressivity that we considered in both works. Basically, we conditioned expressivity solely on the linguistic side without taking prosody into account. However, prosody is a critical element of expressive speech. Thus, to have fully expressive speech, pronunciation variants and disfluencies have to be generated alongside their prosody. Therefore, we believe that an additional study on the prosody side of pronunciation variants and disfluencies would definitely lead to better results in an expressivity perspective.

Publications

(1) R. Qader, G. Lecorvé, D. Lolive and P. Sébillot. Probabilistic Speaker Pronunciation Adaptation for Spontaneous Speech Synthesis Using Linguistic Features. In Proceedings of the International Conference on Statistical Language and Speech Processing (SLSP), 2015.

(2) M. Tahon, R. Qader, G. Lecorvé, D. Lolive. Improving TTS with corpus-specific pronunciation adaptation. In Proceedings of Annual Conference of the International Speech Communication Association (Interspeech), 2016.

(3) M. Tahon, R. Qader, G. Lecorvé, D. Lolive. Optimal feature set and minimal training size for pronunciation adaptation in TTS. In Proceedings of the International Conference on Statistical Language and Speech Processing (SLSP), 2016.

(4) R. Qader, G. Lecorvé, D. Lolive and P. Sébillot. Adaptation de la prononciation pour la synthèse de la parole spontanée en utilisant des informations linguistiques. Actes des Journées d'Études sur la Parole (JEP), 2016.

Bibliography

Martine Adda-Decker and Lori Lamel. Pronunciation variants across system configuration, language and speaking style. Speech Communication, 29, 1999.

Martine Adda-Decker, P Boula de Mareüil, and Lori Lamel. Pronunciation variants in french: schwa & liaison. In International Congress of Phonetic Sciences, 1999.

Martine Adda-Decker, Philippe Boula de Mareüil, Gilles Adda, and Lori Lamel. Investigating syllabic structures and their variation in spontaneous french. Speech Communication, 46, 2005.

Jordi Adell, Antonio Bonafonte, and David Escudero. Filled pauses in speech synthesis: towards conversational speech. In Proceedings of Text, Speech and Dialogue (TSD), 2007.

Jordi Adell, Antonio Bonafonte, and David Escudero Mancebo. On the generation of synthetic disfluent speech: local prosodic modifications caused by the insertion of editing terms. In Proceedings of Annual Conference of the International Speech Communication Association (Interspeech), 2008.

Jordi Adell, David Escudero, and Antonio Bonafonte. Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence. Speech Communication, 54, 2012.

Sebastian Andersson, Kallirroi Georgila, David Traum, Matthew Aylett, and Robert AJ Clark. Prediction and realisation of conversational characteristics by utilising spontaneous speech for unit selection. In Speech Prosody, 2010.

Levent M. Arslan and John HL Hansen. Language accent classification in american english. Speech Communication, 18, 1996.

Michael Ashby and John A. Maidment. Introducing Phonetic Science. Cambridge University Press, 2005.

Jo-Anne Bachorowski. Vocal expression and perception of emotion. Current Directions in Psychological Science, 8, 1999.

Rebecca Bates and Mari Ostendorf. Modeling pronunciation variation in conversational speech using prosody. In Proceedings of ISCA Workshop on Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology (ITRW), 2002.

Alan Bell, Daniel Jurafsky, Eric Fosler-Lussier, Cynthia Girand, Michelle Gregory, and Daniel Gildea. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. The Journal of the Acoustical Society of America, 113, 2003.

Alan Bell, Jason M Brenier, Michelle Gregory, Cynthia Girand, and Dan Jurafsky. Predictability effects on durations of content and function words in conversational English. Journal of Memory and Language, 60, 2009.

Grégory Beller. Transformation of expressivity in speech. Linguistic Insights, 97, 2009.

Jacob Benesty, M. M. Sondhi, and Yiteng Huang. Springer Handbook of Speech Processing. Springer Science & Business Media, 2007.

Christina L. Bennett and Alan W. Black. Using acoustic models to choose pronunciation variations for synthetic voices. In Proceedings of Annual Conference of the International Speech Communication Association (Interspeech), 2003.

Christopher M Bishop. Neural networks for pattern recognition. Oxford university press, 1995.

Susan E Brennan and Michael F Schober. How listeners compensate for disfluencies in spontaneous speech. Journal of Memory and Language, 44, 2001.

Keith Brown and Keith Allan. Concise encyclopedia of semantics. Elsevier, 2010.

Jim Byrnes. Signal Processing for Multimedia. IOS Press, 1999.

B Cannas, F Cau, A Fanni, P Sonato, MK Zedda, JET-EFDA contributors, et al. Automatic disruption classification at jet: comparison of different pattern recognition techniques. Nuclear fusion, 46, 2006.

CASANA. The childhood apraxia of speech association (casana), 2013. URL `http://www.apraxia-kids.org/`.

Ken Chen and Mark Hasegawa-Johnson. Modeling pronunciation variation using artificial neural networks for English spontaneous speech. In Proceedings of Annual

Conference of the International Speech Communication Association (Interspeech), 2004.

Jonathan Chevelu, Damien Lolive, Sébastien Le Maguer, and David Guennec. How to compare tts systems: A new subjective evaluation methodology focused on differences. In Proceedings of Annual Conference of the International Speech Communication Association (Interspeech), 2015.

Noam Chomsky. Syntactic Structures. Bod Third Party Titles, 2002.

Herbert H. Clark. Using Language. Cambridge University Press, 1996.

Herbert H Clark. Speaking in time. Speech Communication, 36, 2002.

Herbert H. Clark and Jean E. Fox Tree. Using uh and um in spontaneous speaking. Cognition, 84, 2002.

Beverley Collins and Inger M. Mees. Practical Phonetics and Phonology: A Resource Book for Students. Routledge, 2013.

Ronald Comer and Elizabeth Gould. Psychology Around Us. John Wiley & Sons, 2010.

Martin Corley and Robert J Hartsuiker. Hesitation in speech can... um... help a listener understand. In Proceedings of Meeting of the Cognitive Science Society, 2003.

Rasmus Dall, Marcus Tomalin, Mirjam Wester, William J Byrne, and Simon King. Investigating automatic & human filled pause insertion for speech synthesis. In Proceedings of Annual Conference of the International Speech Communication Association (Interspeech), 2014.

Robert I. Damper and John FG Eastmond. Pronunciation by analogy: Impact of implementational choices on performance. Language and Speech, 40, 1997.

Michael J. Dedina and Howard C. Nusbaum. PRONOUNCE: a program for pronunciation by analogy. Computer Speech & Language, 5, 1991.

Elisabeth Delais-Roussarie, Mathieu Avanzi, and Sophie Herment. Prosody and Language in Contact: L2 Acquisition, Attrition and Languages in Multilingual Situations. Springer, 2015.

Philip Dilts. Modelling phonetic reduction in a corpus of spoken English using Random Forests and Mixed- Effects Regression. PhD Thesis, University of Alberta, 2013.

Thomas Drugman, Alexis Moinet, and Thierry Dutoit. On the use of machine learning in statistical parametric speech synthesis. In Proceedings of Benelearn, 2008.

Danielle Duez. Silent and non-silent pauses in three speech styles. Language and Speech, 25, 1982.

Eric Fosler-Lussier and Nelson Morgan. Effects of speaking rate and word frequency on conversational pronunciations. In Modeling Pronunciation Variation for Automatic Speech Recognition, 1998.

Eric Fosler-Lussier and Nelson Morgan. Effects of speaking rate and word frequency on pronunciations in convertional speech. Speech Communication, 29, 1999.

Eric Fosler-Lussier et al. Multi-level decision trees for static and dynamic pronunciation models. In Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), 1999.

John Eric Fosler-Lussier. Dynamic pronunciation models for automatic speech recognition. PhD thesis, University of California, Berkeley Fall 1999., 1999.

Carol A Fowler and Jonathan Housum. Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. Journal of Memory and Language, 26, 1987.

Jean E. Fox Tree and Josef C. Schrock. Discourse markers in spontaneous speech: Oh what a difference an oh makes. Journal of Memory and Language, 40, 1999.

Jean E. Fox Tree and Josef C. Schrock. Basic meanings of you know and i mean. Journal of Pragmatics, 34, 2002.

William J. Frawley. International encyclopedia of linguistics. Oxford university press, 2003.

Toshiaki Fukada, Takayoshi Yoshimura, and Yoshinori Sagisaka. Automatic generation of multiple pronunciations based on neural networks. Speech Communication, 27, 1999.

Prasanta Kumar Ghosh and Shrikanth Narayanan. Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion. The Journal of the Acoustical Society of America, 130, 2011.

Egidio Giachin, Aaron Rosenberg, and Chin-Hui Lee. Word juncture modeling using phonological rules for HMM-based continuous speech recognition. In Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1990.

Randy Goldberg and Lance Riek. A Practical Handbook of Speech Coders. CRC Press, 2000.

Masataka Goto, Katunobu Itou, and Satoru Hayamizu. A real-time filled pause detection system for spontaneous speech recognition. In Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), 1999.

D. Govind and SR Mahadeva Prasanna. Expressive speech synthesis: a review. International Journal of Speech Technology, 16, 2013.

Steven Greenberg. Speaking in shorthand–A syllable-centric perspective for understanding pronunciation variation. Speech Communication, 29, 1999.

Steven Greenberg, Hannah Carvey, and Leah Hitchcock. The relation between stress accent and pronunciation variation in spontaneous american english discourse. In Proceedings of Speech Prosody, 2002.

David Guennec and Damien Lolive. Unit selection cost function exploration using an a* based text-to-speech system. In Proceedings of Text, Speech and Dialogue (TSD), 2014.

Isabelle Guyon and André Elissef. An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 2003.

Ke-Song Han and Gui-Lin Chen. Letter-to-sound for small-footprint multilingual tts engine. In Proceedings of International Conference on Spoken Language Processing (ICSLP), 2004.

William J. Hardcastle and Nigel Hewlett. Coarticulation: Theory, Data and Techniques. Cambridge University Press, 2006.

William J. Hardcastle, John Laver, and Fiona E. Gibbon. The Handbook of Phonetic Sciences. John Wiley & Sons, 2010.

Peter A Heeman and James F Allen. Speech repairs, intonational phrases, and discourse markers: modeling speakers' utterances in spoken dialogue. Computational Linguistics, 25, 1999.

Keikichi Hirose and Jianhua Tao. Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis. Springer, 2015.

Harald Höge, Zdravko Kacic, Bojan Kotnik, Matej Rojc, Nicolas Moreau, and Horst-Udo Hain. Evaluation of modules and tools for speech synthesis: the ecess framework. In LREC, 2008.

Wendy Holmes. Speech synthesis and recognition. CRC press, 2001.

Matthias Honal and Tanja Schultz. Automatic disfluency removal on recognized spontaneous speech-rapid adaptation to speaker dependent disfluencies. In Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2005.

Xuedong Huang, Alejandro Acero, and Hsiao-Wuen Hon. Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice Hall PTR, 2001.

Xuedong D Huang, Yasuo Ariki, and Mervyn A Jack. Hidden Markov models for speech recognition, volume 2004. Edinburgh university press Edinburgh, 1990.

Stéphane Huet, Guillaume Gravier, and Pascale Sébillot. Morpho-syntactic post-processing of n-best lists for improved french automatic speech recognition. Computer Speech & Language, 24, 2010.

Akemi Iida, Nick Campbell, Fumito Higuchi, and Michiaki Yasumura. A corpus-based speech synthesis system with emotion. Speech Communication, 40, 2003.

International Phonetic Association, editor. Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet. Cambridge University Press, 1999.

Per-Anders Jande. Phonological reduction in swedish. In Proceedings of ICPhS, 2003.

Sittichai Jiampojamarn. Grapheme-to-phoneme conversion and its application to transliteration. PhD thesis, University of Alberta, 2011.

Dan Jurafsky, Wayne Ward, Zhang Banping, Keith Herold, Yu Xiuyang, and Zhang Sen. What kind of pronunciation variation is hard for triphones to model? In Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2001.

Panagiota Karanasou. Phonemic variability and confusability in pronunciation modeling for automatic speech recognition. PhD thesis, Université Paris Sud-Paris XI, 2013.

Penny Karanasou, François Yvon, Thomas Lavergne, and Lori Lamel. Discriminative training of a phoneme confusion model for a dynamic lexicon in ASR. In Proceedings of Annual Conference of the International Speech Communication Association (Interspeech), 2013.

Reima Karhila, Ulpu Remes, and Mikko Kurimo. Noise in hmm-based speech synthesis adaptation: Analysis, evaluation methods and experiments. IEEE Journal of Selected Topics in Signal Processing, 8, 2014.

Mayank Kaushik, Matthew Trinkle, and Ahmad Hashemi-Sakhtsari. Automatic detection and removal of disfluencies from spontaneous speech. In Proceedings of the Australasian International Conference on Speech Science and Technology (SST), 2010.

Anne K Kienappel and Reinhard Kneser. Designing very compact decision trees for grapheme-to-phoneme transcription. In Proceedings of Annual Conference of the International Speech Communication Association (Interspeech), 2001.

Simon King. A beginners' guide to statistical parametric speech synthesis. The Centre for Speech Technology Research, University of Edinburgh, UK, 2010.

Simon King. An introduction to statistical parametric speech synthesis. Sadhana, 36, 2011.

Simon King and Vasilis Karaiskos. The Blizzard Challenge 2012. In Proceedings of Blizzard Challenge 2012 Workshop, 2012.

Katrin Kirchhoff. Robust speech recognition using articulatory information. PhD thesis, University of Bielefeld, 1999.

Paul R Kleinginna Jr and Anne M Kleinginna. A categorized list of emotion definitions, with suggestions for a consensual definition. Motivation and Emotion, 5, 1981.

Hans Kruschke. Simulation of speaking styles with adapted prosody. In Proceedings of Text, Speech and Dialogue (TSD), 2001.

Julian Kupiec. Robust part-of-speech tagging using a hidden markov model. Computer Speech & Language, 6, 1992.

Gitta PM Laan. The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style. Speech Communication, 22, 1997.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of International Conference on Machine Learning (ICML), 2001.

Willem JM Levelt. Monitoring and self-repair in speech. Cognition, 14, 1983.

Philip Lieberman. Some effects of semantic and grammatical context on the production and perception of speech. Language and Speech, 6, 1963.

Yang Liu, Elizabeth Shriberg, Andreas Stolcke, Dustin Hillard, Mari Ostendorf, and Mary Harper. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. Transactions on Audio, Speech, and Language Processing, 14, 2006.

Karen Livescu, Preethi Jyothi, and Eric Fosler-Lussier. Articulatory feature-based pronunciation modeling. Computer Speech and Language, 36, 2016.

GF Mahl. Disturbances in the patient's speech as a function of anxiety. Eastern Psychological Association, Atlantic City, NJ. Reprinted in I. Pool (Ed.), Trends in content analysis, 1959.

Philippe Boula de Mareüil, Benoît Habert, Frédérique Bénard, Martine Adda-Decker, Claude Barras, Gilles Adda, and Patrick Paroubek. A quantitative study of disfluencies in french broadcast interviews. In Proceedings of Disfluency in Spontaneous Speech Workshop, 2005.

Lutz Marten. At the Syntax-pragmatics Interface: Verbal Underspecification and Concept Formation in Dynamic Syntax. Oxford University Press, 2002.

Corey Miller. Individuation of postlexical phonology for speech synthesis. In The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis, 1998.

Louisa Moats. LETRS, Language Essentials for Teachers of Reading and Spelling. Sopris West Educational Services, 2004.

Terry Myers, John Laver, and John Anderson. The Cognitive Representation of Speech. Elsevier Science, 1981.

Beatrice Oshika, Victor W. Zue, Rollin Weeks, Helene Neu, and Joseph Aurbach. The role of phonological rules in speech understanding research. Transactions on Acoustics, Speech and Signal Processing, 23, 1975.

Vincent Pagel, Kevin Lenzo, and Alan Black. Letter to sound rules for accented lexicon compression. arXiv preprint cmp-lg/9808010, 1998.

Alok Parlikar. Style-Specific Phrasing in Speech Synthesis. PhD thesis, Carnegie Mellon University, 2013.

Nabankur Pathak and P. H. Talukdar. The basic grapheme to phoneme (G2P) rules for bodo language. International Journal, 2, 2013.

John F Pitrelli, Raimo Bakis, Ellen M Eide, Raul Fernandez, Wael Hamza, and Michael A Picheny. The ibm expressive text-to-speech synthesis system for american english. Transactions on Audio, Speech, and Language Processing, 14, 2006.

Mark A Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. Speech Communication, 45, 2005.

Kishore Prahallad, Alan W Black, and Ravishankhar Mosur. Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis. In Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2006.

José C. Príncipe, N.R. Euliano, and W.C. Lefebvre. Neural and adaptive systems: fundamentals through simulations. Wiley, 2000.

Ville Pulkki and Matti Karjalainen. Communication Acoustics: An Introduction to Speech, Audio and Psychoacoustics. Wiley, 2015.

J. Ross Quinlan. Induction of decision trees. Machine Learning, 1, 1986.

K. Sreenivasa Rao. Predicting Prosody from Text for Text-to-Speech Synthesis. Springer Science & Business Media, 2012.

Michael Riley, William Byrne, Michael Finke, Sanjeev Khudanpur, Andrej Ljolje, John McDonough, Harriet Nock, Murat Saraclar, Charles Wooters, and George Zavaliagkos. Stochastic pronunciation modelling from hand-labelled phonetic corpora. Speech Communication, 29, 1999.

Sherry R Rochester. The significance of pauses in spontaneous speech. Journal of Psycholinguistic Research, 2, 1973.

Ralph Leon Rose. The communicative value of filled pauses in spontaneous speech. PhD thesis, University of Birmingham, 1998.

Antti-Veikko I. Rosti and Spyros Matsoukas. Combining outputs from multiple machine translation systems. In Proceedings of NAACL-HLT, 2007.

James A. Russell. A circumplex model of affect. Journal of Personality and Social Psychology, 39, 1980.

Peter A Schreiber. Understanding prosody's role in reading acquisition. Theory into Practice, 30, 1991.

Marc Schröder. Expressive speech synthesis: Past, present, and possible futures. In Affective information processing. Springer, 2009.

Tanja Schultz and Katrin Kirchhoff. Multilingual Speech Processing. Elsevier Science, 2006.

Terrence J. Sejnowski and Charles R. Rosenberg. Parallel networks that learn to pronounce english text. Complex systems, 1, 1987.

Elizabeth E Shriberg. Phonetic consequences of speech disfluency. Technical report, DTIC Document, 1999.

Elizabeth Ellen Shriberg. Preliminaries to a theory of speech disfluencies. PhD thesis, University of California, 1994.

Paul Skandera and Peter Burleigh. A Manual of English Phonetics and Phonology: Twelve Lessons with an Integrated Course in Phonetic Transcription. Gunter Narr Verlag, 2011.

Andreas Stolcke and Elizabeth Shriberg. Statistical language modeling for speech disfluencies. In Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1996.

Andreas Stolcke, Elizabeth Shriberg, Rebecca A Bates, Mari Ostendorf, Dilek Hakkani, Madelaine Plauche, Gökhan Tür, and Yu Lu. Automatic detection of sentence boundaries and disfluencies based on recognized words. In Proceedings of International Conference on Spoken Language Processing (ICSLP), 1998.

Andreas Stolcke, Jing Zheng, Wen Wang, and Victor Abrash. Srilm at sixteen: Update and outlook. In Proceedings of Automatic Speech Recognition and Understanding Workshop (ASRU), 2011.

Andreas Stolcke et al. Srilm-an extensible language modeling toolkit. In Proceedings of Annual Conference of the International Speech Communication Association (Interspeech), 2002.

Helmer Strik, Judith M Kessens, and Mirjam Wester. Modeling pronunciation variation for automatic speech recognition. In Proceedings of the European Speech Communication Association (ESCA) Workshop, 1998.

Shiva Sundaram and Shrikanth Narayanan. An empirical text transformation method for spontaneous speech synthesizers. In Proceedings of Annual Conference of the International Speech Communication Association (Interspeech), 2003.

Charles Sutton and Andrew McCallum. An introduction to conditional random fields for relational learning. Introduction to statistical relational learning, 2006.

Marc Swerts, Anne Wichmann, and R-J Beun. Filled pauses as markers of discourse structure. In Proceedings of International Conference on Spoken Language Processing (ICSLP), 1996.

Gary Tajchman, Eric Foster, and Daniel Jurafsky. Building multiple pronunciation models for novel words using exploratory computational phonology. In Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), 1995.

Paul Taylor. Hidden markov models for grapheme to phoneme conversion. In Proceedings of Annual Conference of the International Speech Communication Association (Interspeech), 2005.

Paul Taylor. Text-to-speech synthesis. Cambridge university press, 2009.

Marcus Tomalin, Mirjam Wester, Rasmus Dall, W Byrne, and Simon King. A lattice-based approach to automatic filled pause insertion. In Proceedings of the Disfluency in Spontaneous Speech (DiSS) Workshop, 2015.

Jean E. Fox Tree. The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. Journal of Memory and Language, 34, 1995.

Jean E Fox Tree. Listeners' uses ofum anduh in speech comprehension. Memory & cognition, 29, 2001.

Shu-Chuan Tseng. Grammar, prosody and speech disfluencies in spoken dialogues. Unpublished doctoral dissertation. University of Bielefeld, 1999.

Jack V Tu. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. Journal of Clinical Epidemiology, 49, 1996.

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In Proceedings of ISCA Speech Synthesis Workshop.

Bahram Vazirnezhad, Farshad Almasganj, and Seyed Mohammad Ahadi. Hybrid statistical pronunciation models designed to be trained by a medium-size corpus. Computer Speech & Language, 23, 2009.

Robert L Weide. The cmu pronouncing dictionary. URL: http://www.speech.cs.cmu.edu/cgibin/cmudict, 1998.

Colin W Wightman and Mari Ostendorf. Automatic labeling of prosodic patterns. Transactions on Speech and Audio Processing, 2, 1994.

Junichi Yamagishi, Koji Onishi, Takashi Masuko, and Takao Kobayashi. Acoustic modeling of speaking styles and emotional expressions in hmm-based speech synthesis. Transactions on Information and Systems, 88, 2005.

Shi Yin, Chao Liu, Zhiyong Zhang, Yiye Lin, Dong Wang, Javier Tejedor, Thomas Fang Zheng, and Yinguo Li. Noisy training for deep neural networks in speech recognition. EURASIP Journal on Audio, Speech, and Music Processing, 2015, 2015.

Hong You, Abeer Alwan, Abe Kazemzadeh, and Shrikanth Narayanan. Pronunciation variations of spanish-accented english spoken by young children. In Proceedings of Annual Conference of the International Speech Communication Association (Interspeech), 2005.

Yuli You. Audio Coding: Theory and Applications. Springer US, 2010.

Heiga Zen, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan W Black, and Keiichi Tokuda. The HMM-based speech synthesis system (HTS) version 2.0. In Proceedings of SSW, 2007.

Heiga Zen, Keiichi Tokuda, and Alan W Black. Statistical parametric speech synthesis. Speech Communication, 51, 2009.

Heiga Zen, Andrew Senior, and Mike Schuster. Statistical parametric speech synthesis using deep neural networks. In Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013.

List of Figures