Intonation Modelling for Speech Synthesis and Emphasis Preservation

THIS IS A TEMPORARY TITLE PAGE
It will be replaced for the final print by a version
provided by the service academique.

Thèse n. 7520
présentée le 22 Décembre 2016
à la Faculté des Sciences et Techniques de l'Ingénieur
Programme Doctoral en Génie Électrique (EDEE)
Laboratoire LIDIAP (Idiap Research Institute)
École Polytechnique Fédérale de Lausanne

pour l'obtention du grade de Docteur ès Sciences
par

Pierre-Edouard Honnet

devant le jury composé de:

Dr. Jean-Marc Vesin, président du jury
Prof. Hervé Bourlard, directeur de thèse
Dr. Philip N. Garner, co-directeur de thèse
Prof. Jean-Philippe Thiran, rapporteur
Dr. Junichi Yamagishi, rapporteur
Dr. Antonio Bonafonte, rapporteur

Lausanne, EPFL, 2016

Abstract

Speech-to-speech translation is a framework which recognises speech in an input language, translates it to a target language and synthesises speech in this target language. In such a system, variations in the speech signal which are inherent to natural human speech are lost, as the information goes through the different building blocks of the translation process. The work presented in this thesis addresses aspects of speech synthesis which are lost in traditional speech-to-speech translation approaches.

The main research axis of this thesis is the study of prosody for speech synthesis and emphasis preservation.

A first investigation of regional accents of spoken French is carried out to understand the sensitivity of native listeners with respect to accented speech synthesis. Listening tests show that standard adaptation methods for speech synthesis are not sufficient for listeners to perceive accentedness. On the other hand, combining adaptation with original prosody allows perception of accents.

Addressing the need of a more suitable prosody model, a physiologically plausible intonation model is proposed. Inspired by the command-response model, it has basic components, which can be related to muscle responses to nerve impulses. These components are assumed to be a representation of muscle control of the vocal folds. A motivation for such a model is its theoretical language independence, based on the fact that humans share the same vocal apparatus. An automatic parameter extraction method which integrates a perceptually relevant measure is proposed with the model. This approach is evaluated and compared with the standard command-response model.

Two corpora including sentences with emphasised words are presented, in the context of the *SIWIS* project. The first is a multilingual corpus with speech from multiple speaker; the second is a high quality speech synthesis oriented corpus from a professional speaker.

Two broad uses of the model are evaluated. The first shows that it is difficult to predict model parameters; however the second shows that parameters can be transferred in the context of emphasis synthesis. A relation between model parameters and linguistic features such as stress and accent is demonstrated. Similar observations are made between the parameters and emphasis. Following, we investigate the extraction of atoms in emphasised speech and their transfer in neutral speech, which turns out to elicit emphasis perception. Using clustering methods, this is extended to the emphasis of other words, using linguistic context. This approach is validated by listening tests, in the case of English.

Keywords: intonation modelling, regional accents, intonation synthesis, intonation-based emphasis synthesis, prosody, text-to-speech synthesis, speech-to-speech translation

Résumé

La synthèse vocale est une partie inhérente de la traduction parole à parole. Elle permet de restituer un signal de parole dans la langue voulue, après reconnaissance et traduction d'un signal de parole dans une langue source. Malgré l'existence de systèmes capables de traduire la parole, la parole synthétique en sortie de ces systèmes reste générique et ne reflète pas des nuances de l'utilisateur, présentes dans le signal d'entrée. Cette thèse traite des aspects de la communication que la représentation textuelle, vecteur de l'information dans un systeme de traduction parole à parole, ne contient pas.

La ligne de recherche centrale des travaux présentés dans cette thèse est l'étude de la prosodie pour la synthèse et la restitution d'accents sur certains mots dans la phrase.

La nécessité d'une prosodie correcte lors de la synthèse est d'abord motivée par une étude sur la perception d'accents régionaux du français. Cette étude, conduite à l'aide de tests d'écoute, de méthodes de synthèse et de modifications de la prosodie, démontre que les méthodes d'adaptation standard d'un système de synthèse sont insuffisantes pour percevoir les accents régionaux, et que l'utilisation de la prosodie originale permet à des natifs de déceler l'intensité des accents des locuteurs.

Dans le cadre multilingue de la traduction parole à parole, un modèle d'intonation théoriquement indépendant de la langue est proposé. Similaire au modèle commande-réponse, qui décompose l'intonation en plusieurs éléments pour décrire les variations à court et moyen termes, le modèle proposé s'appuie sur une description du contour de l'intonation par des éléments de base, inspirés des réponses musculaires à des impulsions nerveuses. Cette formulation rend le modèle plausible d'un point de vue physiologique, et vise à prendre en compte la formation de l'intonation par les cordes vocales. Le modèle est proposé avec une méthode automatique d'extraction des paramètres prenant en compte l'aspect perceptif de l'intonation. Cette approche, évaluée et comparée au modèle commande-réponse standard, permet d'atteindre un niveau de modélisation avec la précision voulue.

Deux corpus de parole incluant des mots spécifiquement accentués sont construits afin de permettre l'étude et l'application de nos méthodes d'accentuation. Le premier contient des phrases prononcées dans plusieurs langues par plusieurs locuteurs. Le second est un corpus de haute qualité enregistré par une actrice professionnelle, destiné entre autres à la synthèse vocale du français.

Deux applications du modèle proposé sont evaluées par la suite : la prédiction d'intonation pour la synthèse de parole, qui s'avère difficile, et la synthèse d'intonation de mots accentués, dans le cadre de la restitution d'accents sur les mots voulus dans une phrase. Une étude des

paramètres du modèle démontre leur relation avec des descriptions linguistiques du texte, telles que l'accentuation des syllabes, et avec l'accentuation de mots spécifiques. Après un simple transfert de composantes dans l'intonation de mots accentués vers les mots neutres prononcés dans le même contexte, l'utilisation d'arbres de décision pour générer ces composantes afin de les substituer à celles d'un mot neutre permet de susciter la perception de l'accentuation sur ce mot neutre. Des tests perceptifs corroborent l'efficacité de cette méthode dans un scénario monolingue.

Mots clefs : modélisation de l'intonation, accents régionaux, synthèse de l'intonation, synthèse de mots accentués, prosodie, synthèse vocale, traduction parole à parole

Acknowledgements

I would first like to thank Hervé Bourlard, Phil Garner and Junichi Yamagishi for their supervision. More specifically, Hervé for his work to make Idiap such a great working environment, Phil for his wisdom, guidance and availability as well as for fun moments outside of Idiap, and Junichi for giving me the opportunity to do an internship in his lab in Japan, sharing with me his experience, his dynamism and for very fruitful discussions. I would also like to thank the Swiss National Science Fundation for funding my work via the SIWIS project.
I am very grateful to my thesis committee, Antonio Bonafonte, Junichi Yamagishi, Jean-Marc Vesin and Jean-Philippe Thiran, for reading my thesis and providing constructive feedback which improved the quality of the thesis.
The work presented in this thesis could not have been achieved without the collaboration and help of many colleagues, from Idiap and elsewhere. To mention only them, I would like to thank Branko, Alexandros, Milos, Xingyu, Gyorgy, Aleksandar, Yang, and from Japan Shinji Takaki, Kameoka Sensei and Kobayashi Sensei.
At Idiap, things go flawlessly most of the time, which makes working efficiently easy. For this I would like to thank the great system team and administration of Idiap.
The coffee breaks on the 3rd floor at Idiap revolve around various topics, and although some work related discussions sometimes happened there, most of the time it was more of an escape from it. I would like to thank all the people who participated in these breaks.
Tuesdays are synonym of TAM for most Idiap people, but they were also synonym of unihockey for some of us. I am glad we almost always managed to find enough people to do some minimal amount of sports in the middle of the week. Many skiing and hiking events made the weekends very enjoyable, and discovering places around Martigny was a lot of fun. All the beer outings and other social gatherings in the town were also a nice break from work. Thank you to all the people involved in these activities.
I started to share an apartment with friends for the first time in 2010, since then I have had many different flatmates, that I would like to thank to make my mood go up after a bad day at work. From Grenoble to Martigny, passing by Dublin and Tokyo, I have had interesting discussions and spent some good time with all my flatmates. I have also made many friends at Idiap and I would forget some if I gave names, so I rely on you to recognise yourselves.
Finally, I am grateful to my family who supported me, and more specifically to my parents who lead me there by letting me choose my own ways and helping me achieving my goals.

Martigny, December 2016 P.-E. H.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

1.1.1 The Swiss Context

Switzerland is surrounded by countries with different languages and cultures: Germany, France, Italy, Liechtenstein and Austria. As a consequence, like a few other countries, it is a multicultural nation which has multiple official languages: German, French, Italian, and Romansh[1].

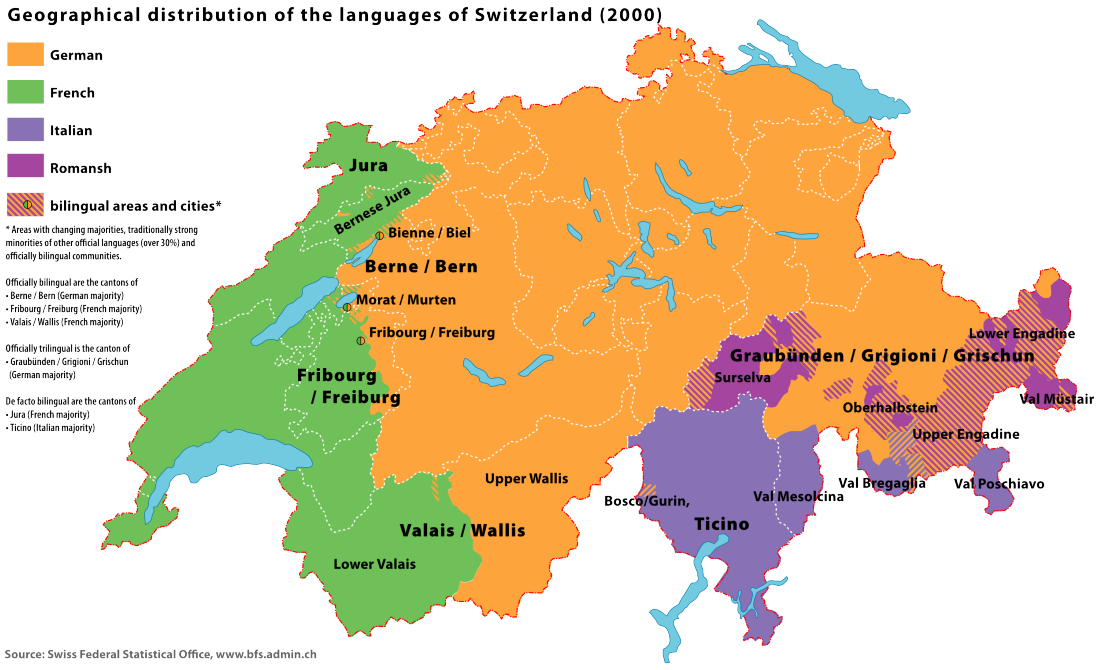

Figure 1.1 – Map of Switzerland with language spoken in each region.[2]

[1] Art. 4 of the Federal Constitution of the Swiss Confederation, see https://www.admin.ch/ch/e/rs/1/101.en.pdf
[2] Source: Swiss Federal Statistical Office http://www.bfs.admin.ch

Figure 1.1 shows the distribution of languages spoken within Switzerland. Although there are some bilingual or even trilingual areas, most of the regions are monolingual. The fact that some people are not able to speak the language of a region when they travel in their own country makes evident the need for a translation medium. More generally, at the international level, there are thousands of languages spoken in the world (about 6000 are identified) and learning them all is practically impossible. Speech-to-speech translation (S2ST) consists of recognising speech in a source language, translating it and synthesising the translation in the target language. The input and output of such a system are both speech. The following section introduces S2ST.

1.1.2 Speech-to-Speech Translation

Speech-to-speech translation enables communication between individuals speaking different languages, through some device (e.g. Figure 1.2 between French and German).

Figure 1.2 – Speech-to-speech translation. The boy speaks in French, it gets translated to German by a device[3].

The 3 main building blocks of such a system are:

- Automatic speech recognition (ASR), which converts speech into text

[3] © Coralie Jenny, http://ateliercocoshi.wixsite.com/portfolio, used with permission.

- Machine translation (MT), which translates the text from one language to another

- Text-to-speech (TTS) synthesis, which converts text into speech.

Automatic Speech Recognition

Automatic speech recognition is the start of the S2ST chain. Its role in a classical setting is to disregard the variability between different speakers' speech, and the variability in the voice of a speaker uttering the same sentence in different ways, in order to capture only *what* the speaker said. The performance of state of the art ASR systems is very good in clean conditions. Following the deep learning trend in various research fields such as image classification [Krizhevsky et al., 2012] or natural language processing [Collobert and Weston, 2008], most of the systems have evolved from hidden Markov model (HMM)-based ASR [Rabiner, 1989] towards deep neural network (DNN)-based ASR [Hinton et al., 2012].

Machine Translation

Machine translation is probably the most difficult problem in the context of S2ST, as it will start from an imperfect input, as errors in ASR exist. Some of the difficulties that can be encountered in this task include the pronunciation of words, which varies considerably in a dialogue compared with the pronunciation of read and well articulated speech. Another aspect is the variability in language structures, e.g. if English and French are relatively close in terms of sentence structure — meaning that some mapping can be done between words to some extent — it is not the case between English and Mandarin, where some ideas can be expressed in very different ways. The MT problem is not addressed in this work, and MT is assumed to work flawlessly for our work on other modules.

Text-To-Speech Synthesis

Text-to-speech synthesis can be achieved by many different techniques, the most widely used ones being unit selection and statistical parametric speech synthesis (SPSS). The first one consists of selecting speech units from natural speech recordings and concatenating them to produce the desired output in the most natural way as possible, based on target and concatenation costs. This is one of the best methods for producing natural sounding speech, however it requires a fair amount of data, and cannot handle the generation of sound combinations which are not in the database. Additionally, this method is solely speaker dependent hence inflexible. SPSS, on the other hand, relies on the parameterisation of speech (acoustic features) and statistical models which can predict the acoustic parameters to synthesise. As for ASR state of the art systems, in TTS the most commonly used statistical models are HMM [Tokuda et al., 2002b; Zen et al., 2009] and more recently DNN [Zen et al., 2013] and their variants, e.g. [Zen and Sak, 2015; Zen and Senior, 2014].

1.1.3 Personalised Speech-to-Speech Translation

As described above, the building blocks of an S2ST system are generally dealt with separately. That means that the personality, or voice characteristics of the user — who inputs speech in the system — are suppressed to obtain better performance for the first two tasks (ASR and MT). This implies a generic output, with the same TTS voice for all the users, which for a human to human conversation would sound rather unnatural. Some commercial applications have recently achieved online S2ST, e.g. Skype™ [2], but this solution is not personalised. In the last decade, personalised S2ST received some attention from the speech research community, with projects such as EMIME [Kurimo et al., 2010][3]. Personalising S2ST consists of being able to produce an output that is specific to the user, meaning that if speaker X speaks in language A, then the output in language B should have the same voice as speaker X. The key to the realisation of such a system relies on the adaptation of the output voice using data from the input language. Adapting the TTS module in this context is referred to as cross-lingual speaker adaptation (CLSA), and consists of adapting a system in one language with data in another language [Chen et al., 2009; Gibson et al., 2010; Liang, 2012; Liang and Dines, 2011; Oura et al., 2010; Peng et al., 2010; Wu et al., 2008, 2009; Yoshimura et al., 2013].

1.1.4 Prosody in Translation

As described earlier, in a classical S2ST system, information about the speaker would be lost when the speech is processed by ASR. CLSA aims at "fixing" this issue by adapting the TTS to the speaker's voice. In addition to the speaker's voice characteristics, his/her prosody would be lost in the process as well. Prosody, which is not about *what* was said but rather *how* it was said, carries information about the speaker identity, mood, social background, emotions, speaking style and intentions. If some of the parameters controlling prosody can be easily adapted locally for a TTS system, such as the average and ranges of speaking rate, intonation and intensity, it is rather difficult to translate prosody in some specific given context. Agüero et al. [2006] have investigated prosody generation for S2ST by exploiting intonation, and pauses [Agüero et al., 2008]. Later, a few approaches to intent transfer have been proposed [Anumanchipalli et al., 2012; Do et al., 2015a,b, 2016a,b]. Transferring — or translating — intent basically means to be able to emphasise correctly the words in the TTS output, to reproduce the underlying meaning expressed through prosody in the input speech. A more complete literature survey on these aspects is included later.

1.2 Scope of the Thesis

The goal of this thesis is to investigate how TTS synthesis systems could be improved in the context of S2ST by using information present in the input of the speech translation system which is traditionally lost in ASR and MT steps.

[2] http://www.skype.com

[3] *Effective Multilingual Interaction in Mobile Environments*, more details at http://www.emime.org/

As brought up earlier, there are several aspect of TTS models that could benefit from using some information in the source speech, although it is in a different language from the target speech. In the context of Swiss languages, regional accents play an important role in the identification of speakers' origins. It is then worth investigating how speech synthesis models can be adapted to generate regional accents, and what aspects are important for the listeners to perceive accentedness. Along with speaker identity, the way things are said is an important aspect of human communication, as it allows disambiguation, or focus on some information. The variations observed can partly be attributed to the speaker's intonation. Different languages use intonation in different ways, however all humans produce it using the same vocal apparatus. This thesis focuses on the modelling of intonation in a theoretically language independent manner. Later, the modelling of intonation in TTS is investigated, with the goal of having a flexible system whose parameters can be controlled. The synthesis of word-level focus, combined with focus detection in the input speech, can be a way to translate intention in the context of S2ST. In this work, we restrict ourselves to the synthesis of emphasis, which aims at being further used in an end-to-end system.

1.3 Main Contributions

The main contributions of the thesis, presented in the following chapters, can be summarised as follows:

1. A French speaker dependent speech corpus was recorded. As no high quality speaker dependent speech database freely available for the French language, a corpus design was made and a voice talent was recorded, with specific instructions. The database can be used for TTS purposes, includes multiple styles, and has an acted emphasis component, useful for studies on emphasised speech.

2. The adaptation of standard French TTS models to Swiss regional accented speech is not sufficient for listeners to perceive the Swiss accent. Using correct prosody alone with standard French pronunciation increases the perception of Swiss accent, and when combined with model adaptation is not perceived significantly differently from original Swiss speech in terms of accentedness.

3. A new intonation model was proposed, which is a generalisation of the command-response model. The components of the model differ from the original command-response model and are less restrictive. The model is physiologically plausible, which makes it theoretically language independent. It is proposed with an extraction method, based on the matching pursuit algorithm. A perceptually relevant measure is proposed and integrated in weighted matching pursuit method to extract meaningful components.

4. The aforementioned model was used in an intonation prediction scenario: several attempts to use statistical modelling to predict its parameters were made. The peculiarities of the model turned out to make the prediction of the parameters from standard

TTS contextual labels difficult.

5. The relevance of the generalised command-response model components with respect to prosodic contexts reveals that the components share mutual information with accent and stress information. Emphasis is also found to share mutual information with the parameters of the model components. Intonation-based emphasis synthesis is investigated using the model and modification of synthetic speech. The generation of emphasis-specific word-level intonation proves able to produce emphasis.

1.4 Outline

The thesis has 7 chapters. Chapter 2 introduces some notions and previous work, as well as databases and methods used in the thesis.

Chapter 3 presents the procedure for recording a new database which is coherent with existing databases, and with some additional features.

Chapter 4 is an investigation of the regional accent perception for the case of French in Switzerland as opposed to the French accent from France. We use TTS, speaker adaptation and analysis-synthesis to modify segmental and suprasegmental parameters to evaluate how segmental and suprasegmental aspects are perceived for this specific regional accent.

In Chapter 5, a new intonation model is proposed. Intonation modelling background is introduced, and work on generalising the command-response model is presented. Our method to decompose intonation into prosodic elements is explained, and its performance evaluated. The model is compared with the standard command-response model.

In Chapter 6, we investigate the use of the generalised command-response model for the task of intonation generation. Several statistical methods are tested to predict the parameters of the command-response model.

Chapter 7 tackles intonation-based emphasis synthesis, using the model introduced in Chapter 5. After an analysis of model parameter relevance, clustering methods are used to generate emphasis-specific intonation components.

Finally, Chapter 8 gives some summary and conclusions on the work presented in the preceding chapters, with some limitations and possible future directions that could be built up on this work.

2 Background

In this chapter, we introduce several notions and methods on which the work of the thesis is based. Text-to-speech synthesis is introduced as it is at the core of the work, and the various aspects investigated revolve around it. As we are particularly interested in prosody, its general aspects from a signal processing point of view are presented, and more specifically, an overview of the state of the art of intonation modelling is given. In the context of S2ST, emphasis conservation in translation is introduced. Finally, we present the speech databases and evaluation methods used in this work.

2.1 Text-to-Speech Synthesis

A text-to-speech (TTS) synthesis method, as its name indicates it, converts text into speech, in other words, characters into an acoustic signal. The two main categories of speech synthesiser which are currently used are concatenative TTS, and statistical parametric speech synthesis (SPSS) including hidden Markov model-based and deep neural network-based speech synthesis.

The former consists of finding speech segments in a database and concatenating them. Unit-selection TTS is a subcategory of concatenative systems, which chooses and concatenates speech segments of variable lengths according to target and concatenation cost functions [Hunt and Black, 1996]. It generally results in very high quality synthetic speech, as long as the recorded database — which has to be from a single speaker — is big enough to cover the possible cases for the system usage, and is of high quality. The main inconveniences of this method are that it is costly, as a lot of data needs to be recorded, and that it is not flexible, i.e. only speech from the same speaker, with the same speaking style can be synthesised.

SPSS approaches, on the other hand, model speech parameters using knowledge of speech signals and speech production, and use statistical models to connect text — or a representation of text, like a sequence of phonemes — and parameters that compose speech. Generative models are suitable to predict speech parameters from linguistic context — derived from text

by rule-based methods. Then speech can be reconstructed from the synthetic parameters. One of the advantages of SPSS systems is the flexibility introduced by the parameterisation of speech: parameters can be modified to result in a modification in the output speech; another advantage is the ability to build systems from a relatively small amount of data. On the other hand, SPSS does not achieve the same quality of speech as the best concatenative techniques, although recently some methods allow the synthesis of decent quality speech. The remainder of the thesis deals with SPSS, unless stated otherwise.

2.1.1 Statistical Parametric Speech Synthesis

Statistical parametric speech synthesis consists of using a representation of speech which captures its characteristics, and a statistical model which can learn links between text and the speech parameters from data. This section gives a sufficient overview of state of the art methods to support the research in the rest of the thesis.

Source-Filter Model

Speech production can be modelled by a source-filter model, where the source represents characteristics of signals produced at the vocal folds, such as periodicity which results from regular opening and closing of the vocal folds; and the filter is a representation of the vocal tract which shapes the resonance of the source waves. In traditional SPSS systems, the source-filter model is used to parameterise speech (with various representations); a simple hypothesis is that voiced and unvoiced excitation signals could be represented by a periodic impulse train (the fundamental frequency, F_0) and white noise, respectively. The vocal tract filter represents the spectral envelope of speech. The speech signal is then the convolution of these two components:

$$s(n) = e(n) * h(n) \tag{2.1}$$

where $s(n)$ is the speech signal, $e(n)$ is the excitation and $h(n)$ is the impulse response of the vocal tract system.

The source-filter model is one accepted parameterisation of speech, and most representations of speech that are used in parametric TTS follow the same principle. Some standard representations of the spectral information often used in SPSS are mel-(generalised) cepstral coefficients [Tokuda et al., 1994], or line spectrum pairs [Soong and Juang, 1984]. The ratio between impulses and white noise used in mixed excitation can lead to buzziness while resynthesising speech from extracted parameters — caused by erroneous voiced / unvoiced decision and the presence of strong harmonics at higher frequencies. To overcome this, some vocoders — tools that enable decomposition into a parametric form and reconstruction of speech from these parameters — introduce various methods which "*soften*" the voiced / unvoiced decision. Some typical methods are band aperiodicity, introduced by Kawahara

et al. [1999], or harmonic to noise ratio (HNR) [Garner et al., 2013]. More recently, sinusoidal vocoders have been shown to give more consistent performance than source-filter vocoders. For more details, the reader can refer to an experimental comparison of vocoder types by Hu et al. [2013].

In this work, multiple F_0 extractors were used and are mentioned where necessary. For the spectral parameters, unless stated otherwise the STRAIGHT vocoder was used [Kawahara et al., 1999].

HMM-based Speech Synthesis

A hidden Markov model is a finite state machine which can generate a sequence of observations. It is generally constructed with multiple hidden states, and defined by its transition probabilities between states and the emission probability distribution of each state. The states are hidden, because we do not know the sequence of states, but we know the observation which is emitted by the states. Figure 2.1 shows a 5 state left-to-right HMM without skips.

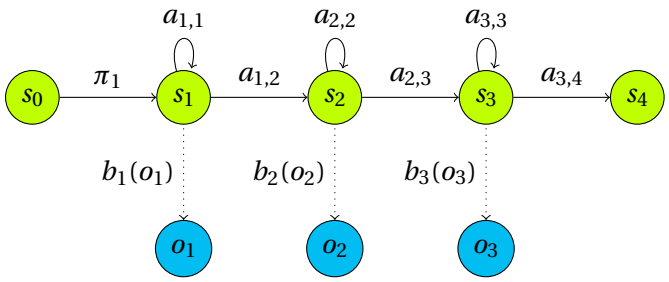

Figure 2.1 – A 5 state left-to-right HMM with no skip. $\{s_i\}_{i=0}^{4}$ are the states (only $\{s_i\}_{i=1}^{3}$ are emitting states). Outputs $\{o_k\}$ are emitted following emission probability densities.

In this figure, a_{ij} is the transition probability from state s_i to s_j, b_i is the output probability distribution of state s_i, o_k the observation emitted by state s_i with the probability $b_i(o_k)$. Intuitively, if we know the state sequence, i.e. how many time steps are spent in each state, generating an observation consists of generating parameters following the output distribution of each state. An N-state HMM is then described by $\mathbf{A} = \{a_{ij}\}_{i,j=1}^{N}$, $\mathbf{B} = \{b_i(\mathbf{o})\}_{i=1}^{N}$ and $\Pi = \{\pi_i\}_{i=1}^{N}$.

The set of parameters describing HMMs is denoted for convenience:

$$\boldsymbol{\lambda} = (\mathbf{A}, \mathbf{B}, \Pi) \tag{2.2}$$

Using HMMs for Synthesis In contrast to the case of ASR, where the goal is to retrieve the most probable sequence of words (or sub-word units), in synthesis, the goal is to retrieve samples from the model given a word sequence. This can be defined as:

$$\hat{\mathbf{o}} = \underset{\mathbf{o}}{\operatorname{argmax}}\, p(\mathbf{o} \mid w, \hat{\boldsymbol{\lambda}}) \tag{2.3}$$

$$= \underset{\mathbf{o}}{\operatorname{argmax}} \sum_{\forall \mathbf{q}} p(\mathbf{o}, \mathbf{q} \mid w, \hat{\boldsymbol{\lambda}}) \tag{2.4}$$

$$\approx \underset{\mathbf{o}}{\operatorname{argmax}} \max_{\mathbf{q}} p(\mathbf{o} \mid \mathbf{q}, \hat{\boldsymbol{\lambda}})\, p(\mathbf{q} \mid w, \hat{\boldsymbol{\lambda}}) \tag{2.5}$$

with $\hat{\mathbf{o}}$ the estimated observation, w the word or sentence, $\hat{\boldsymbol{\lambda}}$ the estimated model parameters, $\mathbf{q}$ the sequence of states, and $p(.)$ the probability density. Equation (2.5) can be approximately divided into two problems:

$$\hat{\mathbf{q}} = \underset{\mathbf{q}}{\operatorname{argmax}}\, P(\mathbf{q} \mid w, \hat{\boldsymbol{\lambda}}) \tag{2.6}$$

$$\hat{\mathbf{o}} = \underset{\mathbf{o}}{\operatorname{argmax}}\, p(\mathbf{o} \mid \hat{\mathbf{q}}, \hat{\boldsymbol{\lambda}}) \tag{2.7}$$

In short, one should first estimate the state sequence, given the word or sentence and model parameters; then estimate the output observation given the state sequence and model parameters. The state sequence is obtained by estimating state durations. Some constraints on continuity and derivatives are added to ensure that there are no discontinuities at state transitions. The introduction of dynamic features (1^{st} and 2^{nd} order derivatives of the static features in most cases) then allows generation of parameter trajectories which evolve smoothly in time, as continuous speech signals do. In HMMs, the duration modelling is quite simple but not very reasonable, therefore hidden *semi*-Markov models (HSMMs) are typically used, as they enable explicit modelling of state duration distributions (replacing state transition probabilities). Various distributions can be used to model state durations, the most commonly used for convenience in the HMM framework being the Gaussian distribution. The training aspects of HMM-based synthesis are not discussed in this work, but comprehensive and detailed introductions are given by Zen et al. [2009] and Yamagishi [2006b].

Modelling Context In ASR, the context around a representation of a phoneme is generally limited to triphones, i.e. the preceding and succeeding phonemes are taken into account in text representation. For TTS, a much larger context is needed. This difference can be explained by the fact that in ASR, the long term dependencies do not increase the performances of the system while in TTS, we are trying to model the high level variability observed — such as different prosodic patterns — in different contexts. This extended context usually includes several preceding and succeeding phonemes, type of phoneme (vowel, fricative, plosive, etc.), syllable level information (accented or not, stressed or not, etc.), word-level features, relative positions of phonemes, syllables, words in the higher unit, etc. The number of

possible combination of contexts becomes huge when the context window increases (meaning that contextual factors are given for the current phoneme, syllable, word, but also for the neighbouring units). As a consequence, the data required to learn parameters for each of the combinations will practically not be available.

To deal with the huge number of possible HMM states, clustering techniques have been proposed to share model parameters among states in each cluster. Using decision trees is one of the standard methods introduced by Young et al. [1994] and Yoshimura et al. [1999]. In a decision tree, at each node except the leaf nodes a binary contextual question, such as "is the current phoneme a vowel?" or "is the next syllable accented?", allows splitting the data into two, based on some measure between the two sets described by the question. The leaf nodes have state output distributions. In the case of unseen context (i.e. if one wants to synthesise some context which was not in the training data), going down in the tree starting from the root node allows reaching one of the leaf nodes, which in the case of a "good" tree will have similar context to the target context. The big set of questions allows generating big trees which will take into account many aspects of the context.

Due to the way HMMs are trained, i.e. in a multi-stream way, it is possible to cluster different parameters in different ways. Thus, the F_0 stream may be clustered differently from the spectral parameters. This makes sense as different parameters are related differently to the context; it allows capturing higher level relations between prosodic features and linguistic context.

DNN-based speech synthesis

Deep neural networks (DNNs) have recently become the state of the art method for many applications, not only in speech processing, but also in image processing or natural language processing for example. A DNN is an artificial neural network (ANN) with many layers. Although ANNs have been proposed for ASR tasks some 30 years ago (e.g. the work of Bourlard and Wellekens [1987]) they have become successful only recently, due to the advances in hardware and availability of large amounts of data, which allow training of such systems. DNNs have started to be used in TTS only very recently, with the first published research presented in 2013 by Zen et al. [2013]. Figure 2.2 shows an example of DNN architecture typically used for TTS. Note that in this example, the number of units per layer is rather small for displaying purposes, but thousands of units are normally used for this task.

The $\{x_i\}_{i=1}^{N}$ correspond to the input layer units, with N the number of input features, $\{h_{i,j}\}_{i,j=1}^{L,K_i}$ is the j^{th} hidden unit from layer i, with L the number of hidden layers, and K_i the number of nodes in layer i. $\{o_i\}_{i=1}^{No}$ are the outputs, with No the number of output features. At each node (here, the hidden units, also called neurons), the input consists of a linear combination of the values from the previous layer, and a non-linear activation function gives the output.

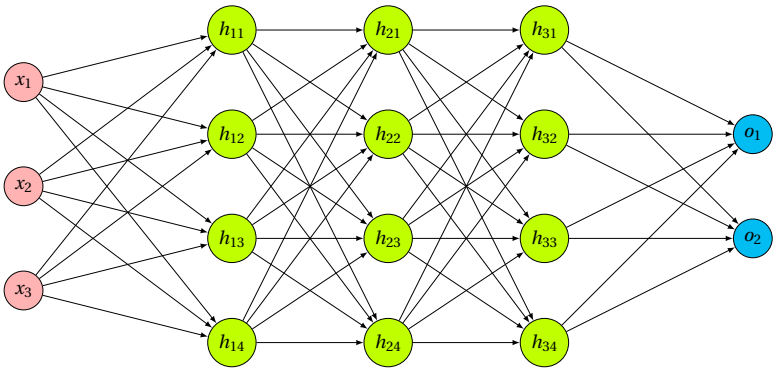

Figure 2.2 – Example of feed-forward DNN with 3 hidden layers, 4 units per layer.

For instance, at the output of unit i of layer l, the value would be:

$$h_{i,l} = f\left(\sum_{k=1}^{K_{l-1}} w_{k,l-1} h_{k,l-1} + b_{i,l}\right) \tag{2.8}$$

where f is the activation function, $w_{k,l-1}$ is the weight between unit $h_{k,l-1}$ and $h_{i,l}$, K_{l-1} is the number of units of the previous layer $(l-1)$ and $b_{i,l}$ is a bias term. One standard activation function used in DNN approaches is the *sigmoid* function:

$$f(y) = \frac{1}{1+\mathrm{e}^{-y}} \tag{2.9}$$

where y is typically the sum in equation (2.8).

In SPSS, DNNs are generally used with the same type of input features as in HMM-based synthesis. To fit the DNN architecture, these features are converted to binary classes, corresponding to the answer to a binary question used in clustering in HMM-based TTS or numerical values (for instance the number of words in the sentence). DNNs can then be seen as a way to replace the decision trees of HMM-based synthesis, as their architecture allows going directly from contextual features to acoustic features.

Other approaches combining SPSS and unit-selection methods, known as hybrid approaches, are at the core of commercial systems, and produce high quality speech. At the time of writing, a new method called “Wavenet” (a generative model for raw audio) was proposed by Google’s DeepMind team, which allows direct synthesis of speech samples, without having to use a representation of speech [van den Oord et al., 2016]. Based on causal convolutional networks, it has shown ability to produce highly natural sounding speech (outperforming both SPSS and unit-selection methods), but the method is still in its infancy and has a high computational cost at the moment.

2.1.2 Speaker Adaptation - Adaptive Training

One of the advantages of statistical parametric synthesis is the possibility to modify model parameters to generate modifications in the output speech. This is interesting as one can imagine transforming a synthetic voice to sound like a particular speaker. In a unit selection system, that would imply collecting a lot of data from this target speaker. By using statistical modelling, it is possible to estimate transforms to go from one voice (for instance an average voice) to another (the target speaker's voice). In the TTS context, it is useful as it allows adapting the trained models to sound like a target speaker using only little data. There are several methods to perform speaker adaptation, based on linear regression (LR), maximum *a posteriori* (MAP), or a combination of them. The most common methods are maximum likelihood linear regression (MLLR), constrained MLLR (CMLLR), structural maximum *a posteriori* linear regression (SMAPLR) and constrained SMAPLR (CSMAPLR). The transform functions used rely on maximum likelihood and MAP criteria [Yamagishi et al., 2009]. In the unconstrained case, only mean vectors are adapted, while in the constrained scenario, the covariance matrix is also adapted to estimate the transforms. Yamagishi et al. [2009] showed that using gender dependent models (training average voice using only speakers from one gender) and adapting them with a combination of CSMAPLR and MAP adaptation was giving better and more stable performance.

Speaker adaptive training (SAT) consists of training an average voice using speech from many different speakers, and requires a smaller amount of data per speaker than regular speaker dependent training. However, because of the speakers' voice characteristics, there is a variability that should be taken into account to estimate the models. This is done by expressing the difference between each training speaker's voice and the average voice as a linear regression function of mean vectors of the state output and duration distributions. MLLR is typically used as a speaker normalisation technique to reduce the influence of speaker differences in terms of voice characteristics. For a better understanding of speaker adaptive training, a detailed and comprehensive analysis is given by J. Yamagishi's PhD thesis [Yamagishi, 2006a], or Yamagishi and Kobayashi [2007]. An average voice trained using SAT can be adapted using standard speaker adaptation methods, and it was shown to obtain better speech quality than using only speaker dependent model when having only a small amount of data from the target speaker [Yamagishi and Kobayashi, 2007].

Speaker adaptation and speaker adaptive training for DNN-based synthesis have started only recently to receive some attention, and the research on these topics is still taking its first steps, with only few successful approaches reported so far [Fan et al., 2015].

2.2 Prosody

The word *prosody* comes from the Greek word προσῳδία [prosɔ:(i)día:], "song sung to music, tone of a syllable", later evolving in the Latin word prosodia, "accent of a syllable", taking its current form in the late 15th century. A definition of the current word could be: "the

components of speech that are not individual phonetic segments"[1]. In other words, it is an aspect which is not related to the sounds that are emitted — and perceived — for each phoneme, but rather the way they are produced and perceived. As it was underlined in Chapter 1, it is not *what* is said, but rather *how* it is said. It encompasses characteristics about the speaker's voice, emotional state, speaking style, socio-linguistic background, intentions.

2.2.1 Prosody in the Speech Signal

In a speech signal, prosody is generally associated with 3 components: the intonation — or speech melody —, the rhythm — including speech rate and its variations — and intensity. The intonation of the speech signal is characterised by the fundamental frequency of speech, denoted F_0, which can be related to the frequency of the vocal fold vibrations. In the source-filter model, this corresponds to the harmonic source. Various methods can extract F_0 from a speech signal. Many models trying to represent intonation exist, we present some of them in Section 2.2.2.

Intensity corresponds to the loudness of speech. It is the energy of the speech signal. In the case of digital speech signal, it is simply calculated as the sum of squared amplitude of each sample in the desired time window.

The rhythm is the speed with which the speaker utters syllables, words, and sentences. The speech rate, which can be calculated at the phoneme, syllable, word or sentence level is quite variable according to the language, speaker and speaking style. Related to the speaking rate, the pauses and their durations are another very important aspect of prosody.

2.2.2 Intonation Modelling

Modelling intonation in the context of this thesis consists of finding a representation of F_0 that allows analysis and possibly synthesis of it. Many intonation models have been proposed, each of them having pros and cons. We give a review of the most pertinent ones to this thesis, which is by no means exhaustive.

There are two main approaches to model intonation: directly modelling pitch, or trying to simulate the pitch production process. If the former category counts numerous models, e.g. Bailly and Holm [2005]; Hirst and Espesser [1993]; Pierrehumbert [1981]; Taylor [2000]; the latter has only few models, e.g. Fujisaki and Nagashima [1969]; Kochanski and Shih [2003]. Most of the models allow analysis and representation of the pitch, however they cannot be used directly with existing generative models in the same fashion as other features, such as in the HMM synthesis framework.

Hirst and Espesser [1993] model the intonation contour as a sequence of specific F_0 target points. In this approach, the intonation contour is approximated by a (piece-wise) quadratic

[1] https://en.wikipedia.org/wiki/Prosody_(linguistics)

spline function, using an algorithm called *Momel* (for "modelling melody"). This yields some smooth continuous version of the *macromelodic* characteristics of intonation. The parameters of the spline functions can be used to get back this version of F_0, which makes Momel reversible. A further step which was taken was to extract from this spline some "interesting" points, defining intonation as a sequence of *tonal segments* which are labelled using an alphabet of 8 symbols. This labelling system, presented by Hirst and Di Cristo [1998], is called *INTSINT*, for INternational Transcription System for INTonation. The possible segment labels are assumed to be absolute tones (top, mid or bottom), relative tones (higher, same or lower) or iterative relative tones (upstepped or downstepped). These labels, associated with the *key* and *span* of the speaker which defines their minimal and maximal pitch values in a logarithmic scale, allow characterising a curve modelled by Momel.

The *Tilt* model describes F_0 as a sequence of events with specific shapes that can be automatically extracted with an obvious synthesis step [Taylor, 2000]. Each *tilt* component is a measure of the shape and amplitude of an event in the prosodic (intonation) contour. The intonational stream is then a sequence of events which each have a rising and falling component. They are characterised by their position, amplitude and some tilt parameter, shaping the event, which varies between -1.0 for a fall and $+1.0$ for a rise, where 0.0 has equal sized rise and fall components. The amplitude is directly linked to the phonetic prominence of the event, while its duration is simply the sum of the rising and falling component durations. In the reconstruction phase, the contour is interpolated linearly between events.

Another model derived by Bailly and Holm [2005], called Superposition of Functional Contours (SFC), is a data driven approach, based on the superposition of elementary contours extracted with the use of neural networks. The first two models try to directly model intonation with no attempt to understand its underlying production process. SFC, on the other hand, mostly relies on metalinguistic information, trying to link high level linguistics to acoustic realisations. It is a data-driven method which uses artificial neural networks to learn how to generate the superpositional contours given metalinguistic functions. The intonation contour is then decomposed into a variable number of function contours. There is no constraint on the shape of these elementary contours, the model should learn from multiple instances of different functions in the database. Although it is an explicit model of prosody, this model is close to recent integrated modelling of intonation in TTS systems, as it tries to learn structure from the data.

Only a few models actually try to explain the intonation by investigating its production aspect. The most popular model in this category is the command response (CR) model of Fujisaki and Nagashima [1969]. This model, based on earlier work from Öhman [1967], decomposes the intonation into additive physiologically meaningful components. Two types of components allow modelling an intonation contour: phrase components, which are long term components, assumed to model the effect of the translation of the thyroid cartilage on the vocal fold tension; and accent components, which are short term components, assumed to model the effect of the rotation of the thyroid cartilage on the vocal fold tension. The relation between the vocal

fold tension and the fundamental frequency is then straightforward [Fujisaki, 2006]. This model is discussed in further detail in Section 5.3, Chapter 5.

Another example, proposed by Kochanski and Shih [2000] is *Stem-ML*, which stands for Soft template Markup Language. It is a model of intonation which, although it was originally developed to model Chinese tones, was designed to be language independent. It is somewhere between linguistics and the physical tool that produces F_0. The intonation production process is described as an optimisation problem where the maximisation of two functions is sought: the *ease of production* and the *speaker's estimate of the efficiency of the specific prosody on the listener*. The former is approximated by a function of the pitch derivatives; the latter is based on the error between prosody targets (defined by tags) and the realised prosody.

Some of these models have been successfully implemented in the context of TTS, as external prosody models [Bailly and Gorisch, 2006; Kameoka et al., 2010; Yoshizato et al., 2012; Zhang et al., 2006]. The following section discusses state of the art methods to model intonation in TTS, namely methods which are well integrated in the framework used for other acoustic features.

2.2.3 Intonation Modelling for TTS

In the context of speech synthesis, intonation is generally modelled within the framework of the method used, namely HMMs and DNNs.

In the first steps of HMM-based synthesis, a multi-space probability distribution (MSD)-HMM was developed and became a standard way of handling the fact that speech can be voiced or unvoiced [Tokuda et al., 2002a]. In the unvoiced regions, F_0 does not actually exist, then it can be represented by a unique symbol whose meaning is "no value". In a traditional HMM, the state output distribution is generally modelled by a Gaussian or another probability distribution. MSD-HMMs allow generating values from different spaces. In the case of F_0 modelling, the observation sequence can be seen as a mixed sequence of outputs from a space with dimension one (single continuous value) for voiced regions and a space with dimension zero (a unique symbol) for unvoiced regions.

In DNN-based synthesis, F_0 is modelled frame by frame, along with other acoustic parameters (spectral parameters) with some smoothing generally done after generating the parameters. The voiced-unvoiced decision is typically modelled by a binary value and is learnt along with other features in the DNN training, or by using MSD as in the HMM case. The linguistic context, input of the DNN, being at the frame level, contains information of higher level than phoneme and word, but also intra-phone information, with some features such as "position in the phoneme", to model patterns.

Recently, some work was done using continuous F_0 and it was shown that continuous F_0 improves the perceived naturalness of synthesis [Latorre et al., 2011; Yu and Young, 2011]. This was further improved by hierarchical modelling using continuous wavelet decomposition to

separate the different levels of variation in F_0 into multiple continuous features, on different scales [Suni et al., 2013]. In this work, the authors exploit the multi-stream architecture of an HMM-based TTS framework to cluster these different temporal scale components with different decision trees. This method was later extended and exploited in DNN-based TTS by Ribeiro et al. [2016a].

Another approach consists of post-processing the synthetic F_0, output of the HMMs. An example of what can be done to improve the output of HMM synthesis is given by Hirose et al. [2011, 2012]. Based on the command response (CR) model [Fujisaki and Nagashima, 1969], the idea is to estimate the F_0 model commands from linguistic information, and then optimise them according to the F_0 generated by HMMs. By modifying the estimated parameters, it becomes possible to increase the expressivity of the synthetic speech. Another attempt to integrate the CR model in HMM-based TTS was made by Hashimoto et al. [2012], where parameterised F_0, generated by the CR model, was used for training the HMM intonation features. This improved the quality of the synthetic speech as modelling the F_0 smoothed its contour before training.

Some interesting work on prosody in speech synthesis, dealing mainly with models which are external to the TTS framework, can be found in a recent collection of work by Hirose and Tao [2015].

2.3 Emphasis

Emphasis in speech can be defined as a specific segment of a sentence (often one word, or group of words) which is given particular importance, something which "stands out". It is usually referred to as *prominence*, *focus*, or *saliency*.

According to Nakajima et al. [2014]:

> "*In human speech, emphasis can be regrouped at least into four functions based on analysis in conventional literature [...] (bold portions show emphasized words and phrases).*
>
> 1. *Expressing linguistic "focus":*
>   e.g., "**Taro** did." (as an answer to "who did ...?")
> 2. *Expressing "contrast":*
>   e.g., "not A **but B**"
> 3. *Expressing "element of surprise":*
>   e.g., "I heard he was sick, but he had **much energy**."
> 4. *Disambiguating grammatical structure: clarifying parallel and dependency structure:*
>   e.g., to distinguish "{old men} and women" from "**old** {men and women}" in "old men and women""

In other words, a prominent word indicates new information, disambiguation, attention to a particular aspect of the sentence or contrast. In the context of this work, we focus on the emphasis of isolated word / group of words.

2.3.1 Prominence of Words

For humans, there is not *one way* of uttering a sentence. In some sentences, most of the words can be emphasised, to imply different underlying meanings. For instance:

- **Jack** didn't eat fish yesterday. (Someone else ate fish)
- Jack **didn't** eat fish yesterday. (He did not eat it)
- Jack didn't **eat** fish yesterday. (He did not eat fish, he may have done something else related to fish)
- Jack didn't eat **fish** yesterday. (He ate something else)
- Jack didn't eat fish **yesterday**. (He ate fish another day)

In a speech signal, aspects related to prosody are the ones which are the most concerned with emphasis. An emphasised word will manifest itself by pauses before and / or after the word, a slower speaking rate in the word, a higher energy in the word, an increased activity in the intonation, or a combination of these features. Emphasis production is language and speaker dependent: different speakers use different cues to express emphasis, and emphasis is delivered differently according to the spoken language.

2.3.2 Word Emphasis Detection and Synthesis

In the context of automatic speech processing, two tasks related to emphasis are generally pertinent: the detection of these prominent words, and their generation. In an S2ST scenario, emphasis detection can be seen as a complementary task for ASR, and emphasis synthesis as a complementary task for TTS.

Emphasis Detection

The literature provides many examples of emphasis detection, with very different approaches. Some techniques are based on pitch, e.g. the work of Kennedy and Ellis [2003] and Arons [1994]. Heldner et al. [1999] used a combination of intensity and spectral tilt. Liang [2016] used a data-driven method, by considering emphatic words as outliers with respect to prosody. Recently, in the context of the *SIWIS* project and this thesis, a method using multi-level demodulation was proposed by Cernak and Honnet [2015], combining stress and syllable modulation peaks to detect emphatic words. Cernak et al. [2016] later proposed an interesting

approach where a phonological vocoder is used to detect differences in the phonological realisation between emphasised and neutral words. Gerazov et al. [2016] investigated the use of a prosodic model originally developed for intonation, by modelling the three main aspects of prosody (intonation, intensity and rhythm) to detect emphasis.

Emphasis Synthesis

Generating emphasis means bringing the listener to perceive emphasis on some desired target word(s). Strom et al. [2007] designed a corpus with many different phonetic contexts, with the same "emphasised" context: A set of three sentences which requires specific emphasis on a name, repeated with about 1100 different names, to cover all the possible diphones, in the context of an emphasised syllable. An example given in their paper is (bold letters indicate words to be emphasised):

> "It was **Erwin** who did it!"
> "No, it was **Eliza** who did it!"
> "It was **Eliza**, not **Erwin**!"

Then, using this corpus, in the context of unit-selection TTS, the authors integrate prominence labels in their target cost function.

Another approach, proposed by Ochi et al. [2009], exploits the command-response intonation model of Fujisaki and Hirose [1984], as it predicts F_0 using this model, and then alter the commands to control the emphasis at will. Hirose et al. [2012] followed the same idea by proposing an intonation contour reshaping method, based on the command-response model as well, in the context of HMM-based synthesis.

Yu et al. [2010] proposed two approaches to emphasis synthesis, using features of decision trees: a two-pass decision tree state clustering, which clusters states first using emphasis related context and then extends the tree using standard contextual features; factorised decision trees, which allow the exploitation of all the data rather than fragmenting it into emphasised/non-emphasised subsets.

2.3.3 Emphasis in Translation

In the context of speech-to-speech translation, there have been attempts to preserve intentions in the translation, in other words, trying to translate emphasis of some words in the input language to the output language. A traditional S2ST system would simply lose any emphasis information at the stage of ASR, the goal of these approaches being to retrieve and use this information.

One way to preserve emphasis in translation is to detect emphasis in the input language, use machine translation as a carrier, and synthesise the sentence with emphasis in the output

language. In that sense, the methods described in the previous sections can be applied in such a scenario. This general approach is investigated in this thesis, in Chapter 7 where emphasis synthesis is tackled, and was at the core of the *SIWIS* project with several methods for emphasis detection proposed by Cernak and Honnet [2015], Cernak et al. [2016] and Liang [2016], or Gerazov et al. [2016] in the SP2 project[2].

Other methods try to directly exploit the input data to reproduce emphasis in the output language. Anumanchipalli et al. [2012] proposed an intent transfer method for S2ST. In this work, a conversion function between accent vectors of the source and target language is trained, based on *tilt* accent vectors. At synthesis time, for the neutral words, word-level *tilt* vectors are predicted, and using the conversion function, the emphasised word *tilt* vector is predicted using the emphasised word vector from the input sentence.

More recently, Do et al. [2015a] proposed a more complete and integrated framework with the use of linear regression HSMMs to preserve word-level emphasis in S2ST. An emphasis weight sequence is estimated for an observation sequence and its transcription using the expectation maximisation algorithm [Dempster et al., 1977]. The translation of emphasis is then performed by estimating the emphasis weight sequence in the target language given the sequence from the input language with conditional random fields (CRF). At the synthesis stage, the speech parameter vector sequence maximises the likelihood function given the state sequence, the word-level emphasis sequence and the model parameters (it is the same as in equation (2.7), with in addition the word emphasis sequence for conditioning). Later, Do et al. [2015b] investigated pause prediction to improve emphasis in the context of S2ST. Following their previous work, the authors used a similar method, where CRFs were employed to predict the pauses in the output language given the pauses in input language. Later, with the increasing interest of the research community in deep learning, Do et al. [2016a] proposed to use long short term memory (LSTM) neural networks to encode and decode emphasis, which slightly improved upon the performance of the CRF-based method. The same authors also proposed to use cluster adaptive training (CAT) with a continuous emphasis representation (as opposed to the binary or multi-level representation that they used before) [Do et al., 2016b].

In the more general context of prosody in S2ST, Agüero et al. [2006] proposed to exploit information which was present in the input language speech signal and to use it for the synthesis of speech in the output language. By learning the relations between intonation in both languages, using this approach improved the naturalness of the TTS output. Pause transfer in S2ST was also proposed by Agüero et al. [2008].

2.4 Datasets

This section gives a brief introduction of the speech databases used in this work and their application in the context of the thesis.

[2] See https://www.idiap.ch/scientific-research/projects/sp2

2.4.1 Multi Speaker Databases

To train HMM-based TTS systems using adaptive training, speech from multiple speakers can be used. Here we describe three datasets that were used for this purpose.

WSJ

The Wall Street Journal (WSJ) corpus [Paul and Baker, 1992] is an American English speech database, consisting of read speech from multiple speakers. It was originally recorded for research on ASR. The recorded sentences were taken from news from the Wall Street Journal news text.

SI84 set This dataset consists of speech from 83 (43 male and 40 female) speakers, originally designed for speaker independent (SI) model building. It contains 7085 sentences amounting to 13.66 hours of speech. The SI84 subset of the WSJ corpus was used to evaluate our intonation model (Chapter 5) and to train an English HMM-based TTS system (Chapter 7).

The CMU Arctic Databases

The CMU Arctic databases, by Kominek and Black [2004], were designed and recorded for the purpose of speech synthesis research. They are high quality single speaker speech databases of about 1200 utterance each. The sentences are phonetically balanced, and recorded in studio conditions. The speech comes with corresponding transcriptions in the Festival format [Black et al., 1997], that can easily be used to build a TTS system. They are widely used in TTS research and have been integrated in speech synthesis system demos[3]. Some of the CMU Arctic databases were used in the evaluation of our intonation model in Chapter 5.

BREF

BREF [Lamel et al., 1991] is another speech database dedicated to ASR research. It is a French read-speech corpus designed for speech recognition model training and testing. The sentences to be read by the speakers were chosen from the French newspaper *Le Monde*. It consists of recordings from 120 selected speakers (55 males and 65 females), recorded in a sound-proof room. The complete database represents more than 100 hours of speech. In this work, we used BREF to build some French TTS models to evaluate French accents (Chapter 4) and to evaluate our intonation model (Chapter 5).

[3] e.g. in HTS demo, see http://hts.sp.nitech.ac.jp/; and more recently in an Idlak Tangle recipe, see https://github.com/bpotard/idlak

PhonDat

PhonDat [Hess et al., 1995] is a corpus of German speech from 201 different speakers. Each speaker read a sub-corpus of 450 different sentences (including single words and two short passages of prose text). The corpus contains a total of 21,681 recorded utterances. It is provided with a phonological segmentation by hand of a small subset and an automatic alignment of the whole corpus. Phondat was used to evaluate our intonation model (Chapter 5).

2.4.2 The PFC Corpus

The PFC project ("Phonologie du Français Contemporain")[4] consisted of recordings from multiple speakers from various French speaking regions, in an attempt to collect speech data from a vast geographical area. It gathers read speech and guided or free conversational speech, from many different locations from French speaking regions of the world. The recordings were done at the homes of the speakers, in an attempt to avoid collecting "lab speech". The dataset from the PFC database [Durand et al., 2009] used in this work consists of read speech by Swiss French speakers and French speakers from Paris. The data was recorded in 5 cities: Paris (France) and 4 cities in 4 different Swiss cantons: Martigny (Valais), Nyon (Vaud), Neuchâtel (Neuchâtel), and Geneva (Geneva). For each location, 4 male and 4 female speakers born and raised in the city were recorded. This corpus was used to investigate Swiss French accents (Chapter 4).

2.4.3 The SIWIS Multilingual Database

One of the goals of the *SIWIS* project was to investigate Swiss languages in the context of speech-to-speech translation. To this end, a multilingual multi-speaker database with emphasis was recorded [Goldman et al., 2016]. As a simple overview, the data consists of speech from bilingual and trilingual speakers, in at least two of the database languages: English, French, German and Italian. The sentences recorded in each language have the same meaning, therefore the corpus is parallel in the language sense, in a similar fashion to the EMIME bilingual corpus [Wester, 2010]. The recorded data consists of news read speech, and a short excerpt of a novel. Another aspect of this corpus is its emphasis component: for a fraction of the sentences, the speakers were asked to emphasise a specific predetermined word in the sentence. As these sentences were also recorded in a neutral fashion, it makes the corpus parallel from an emphasis point of view. In total, about 24 hours of speech were recorded from 36 speakers. More details on the design of the database are provided in Chapter 3, Section 3.2 as it is a contribution of this thesis.

[4] See http://www.projet-pfc.net

2.4.4 Blizzard Challenge Databases

The Blizzard challenge[5] is an annual competition whose goal is to evaluate the quality of synthetic speech using various data-driven methods on the same speech corpus [Black and Tokuda, 2005]. For this, high quality speech databases are built and made available to the participants.

The Blizzard Challenge 2008

In 2008, the competitors had to build a UK English voice and a Mandarin voice [Karaiskos et al., 2008]. *Roger* is the name of the UK English data that was recorded to this end — being the name of the speaker who gave his voice for it. It consists of about 15 hours of recordings with a "fairly standard RP accent". Designed by Strom et al. [2006, 2007], the first purpose of the corpus was to build unit-selection TTS systems with expressive prosody, namely by integrating prominence and emphasis modelling. It contains, among other things, children's stories, read news and word lists. A part of the corpus also consisted of sentences with an explicit structure, to elicit emphatic productions in specific contexts. This corpus was used for emphasis studies in Chapter 7.

The Blizzard Challenge 2011

For the 2011 edition of the Blizzard challenge [King and Karaiskos, 2011], the participants were asked to build a voice using US English data provided by Lessac Technologies. The data, described by Wilhelms-Tricarico et al. [2011], consists of 16.6 hours of speech from a professional female voice talent known as *Nancy*. The text to be read by the speaker was annotated using "*Lessemes*", which are annotations which "explicitly capture the musicality of speech, [avoiding] the artificial separation of prosodic and linguistic features of speech." Eventually, the speaker was presented with text and musical score-like representation of Lessemes. The speaker is a trained singer and had to follow intonation patterns depicted by the annotation. This corpus was used in the context of intonation synthesis in Chapter 6, and for analyses of our intonation model in Chapter 7.

2.4.5 French Female Voice

There is a fair number of high quality speaker dependent speech databases available for English language. This is not true for French: although some multi-speaker databases exist, and one can find relatively good quality free French audiobooks online[6], to the best knowledge of the author, there is no free French speech corpus allowing building high quality speech synthesis systems. The main limitation of audiobook speech is that it is not segmented and, despite some tools such as *ALISA* [Stan et al., 2016], developed in the context of the Simple4All

[5] See http://www.festvox.org/blizzard/

[6] e.g. *Candide*, by Voltaire freely available on LibriVox: https://librivox.org/candide-by-voltaire/

project, requires a lot of work to be well segmented. Within the *SIWIS* project, a French voice talent was recorded, aimed at several use cases. It contains 10 hours of speech, recorded by a native French female speaker selected from a voice talent bank. The data consists of read speech from multiple styles, with text coming from French parliament debates and from novels. Similarly to the *SIWIS* database, the speaker was asked to emphasise specific words for a set of about 1600 sentences. More details on the corpus design, text selection and recordings are given in Section 3.3 of Chapter 3, as it is a contribution of this thesis.

2.5 Evaluation Methods

In the field of text-to-speech synthesis, there are two complementary ways of evaluating the output of a system: objective evaluations, and subjective evaluations. The former consists of comparing synthetic speech to natural human speech via mathematical quantities. The latter consists of asking listeners' opinion about, for instance, quality of speech, intelligibility, or their preference between samples.

2.5.1 Objective Measures

When it comes to evaluate speech synthesiser outputs, the standard procedure is to measure the distance between synthetic parameters and parameters extracted from real speech. These typically corresponds to *mel cepstral distortion* for the spectral part when mel cepstral coefficients are used, and *root mean square error* (RMSE) and *correlation* for F_0. In the context of this thesis, the objective measures concern the accuracy of the intonation that can be synthesised, therefore, RMSE and correlation are the standard objective measures. Some other measures, relevant to specific tasks, are introduced in some chapters when used.

2.5.2 Listening Tests

The complexity of the speech signal makes it rather difficult to evaluate speech quality using only objective measures. Also, measuring synthetic features before vocoding does not take into account the vocoder flaws in the evaluation. A typical method to evaluate the quality or other aspects of synthetic speech is to ask some subjects to listen to samples and to rate the desired aspect.

Standard Listening Tests

There are many ways to design listening tests, depending on the purpose of the evaluation. When evaluating a system, mean opinion score (MOS) tests are usually carried out: the listeners have to rate some aspect of speech on a given scale (typically, the naturalness between 1 and 5 where 1 is very unnatural, 5 completely natural). A typical method to compare two systems consists of using a preference test where listeners have to decide for pairs of sentences

which one they prefer, or if they have no preference. The MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) test has recently started to be used more and more to compare multiple systems and rank them: listeners have to rank all the systems on a continuous scale for the same stimulus generated by each of them. In the case where the intelligibility is investigated, transcription tests are employed: as the name suggests, the listeners have to transcribe what they hear. Another commonly used test is the same/different evaluation, where listeners are asked for each pair if they perceive both versions as exactly the same, or different; this type of test was conducted in Chapter 6.

Other Listening Tests

Some other more specific listening tests can be employed, where the instructions are related to the evaluation of particular aspects of speech. In this thesis, we conducted some listening tests related to:

1. Accentedness: in the context of regional accents, the listeners were asked the degree of accent of the speaker, i.e. how strong they perceived the accent of the speaker. This type of test was conducted in Chapter 4.
2. Emphasis word identification: the listeners were asked to choose which word or group of words was the most emphasised in the sentence they had to listen to.
3. Emphasis strength: related to the previous task, the listeners had to precise how strong they thought the emphasised word they had chosen was. The last 2 tests were carried jointly in Chapter 7.

A Note about Listening Tests

Last year, a very interesting survey was done by Wester et al. [2015] about the quality and validity of listening tests in TTS research. The main conclusion of this piece of work was that to obtain a stable level of significance, 30 listeners at least should take part in the evaluation. The paper also gives some useful guidelines to listening tests for speech synthesis evaluation, that the listening tests conducted in the context of this thesis follow.

3 Data

With the evaluation of speech-to-speech translation and emphasis studies in mind, this chapter presents the design and recording of two speech databases. Both were created in the context of the *SIWIS* project.

The first database is a multilingual corpus which contains parallel speech in several language pairs, from several speakers. It also contains sentences with emphasised words, which are then available from multiple speakers.

The second database is a French high quality speech corpus, aimed at building TTS systems, investigate multiple styles, and emphasis. The speech comes from a French voice talent, and contains about ten hours of speech, including emphasised words in many different contexts.

The first database was joint work with a project partner and was published in the following paper:

– Jean-Philippe Goldman, Pierre-Edouard Honnet, Rob Clark, Philip N. Garner, Maria Ivanova, Alexandros Lazaridis, Hui Liang, Tiago Macedo, Beat Pfister, Manuel Sam Ribeiro, Eric Wehrli, and Junichi Yamagishi. The SIWIS database: a multilingual speech database with acted emphasis. In *Proceedings of Interspeech*, pages 1532–1535, San Francisco, CA, USA, September 2016

The second database was also joint work, although specified mainly by the candidate. It has not been published yet.

3.1 Motivation

The *SIWIS* project is a personalised speech-to-speech translation (S2ST) project, which aims at investigating S2ST in Swiss languages. One of the goals of the project is to enable the multilinguality of the ASR and TTS systems. Another line of research at the core of the project is the modelling of prosody for synthesis in the context of S2ST.

In an attempt to study intonation in the context of intent transfer, based on the detection and synthesis of word emphasis, the two databases presented in this chapter were designed to contain some speech with emphasis. In Section 3.2, we introduce a database which should serve both multilingual study and emphasis study purposes. In Section 3.3, we introduce a database that can be used to build high quality French speech synthesis systems, and allows investigations of the emphasis aspect in various contexts.

3.2 The SIWIS Database: a Multi Speaker Multilingual Data

The EMIME bilingual database, by Wester [2010], is a bilingual database containing several language mappings: English/Finnish and English/German. This bilingual data was aimed at investigation of cross lingual speaker adaptation.

Towards the same ends but with an inclination towards prosody aspects, and a bias towards Swiss languages, a multilingual database was recorded for the *SIWIS* project. As it was inspired by the EMIME bilingual database, it consists of speech from bilingual or trilingual speakers, with parallel content. The following pairs of languages were recorded to cover all the possible mappings amongst the four languages of interest: English, French, German and Italian.

3.2.1 Text Material

For each language, 3 different scenarios were proposed, with a similar architecture. Each set of 171 prompts for each language was divided into 5 parts as follows:

- **europarl**: 25 Europarl statements among which 20 declaratives and 5 interrogatives. The Europarl corpus was used to have a parallel meaning across languages [Koehn, 2005].

- **news**: 100 sentences from newspapers: 80 declaratives and 20 interrogatives.

- **sus**: 20 semantically unpredictable sentences, which can be used for intelligibility assessment.

- **focus**: 25 Europarl statements. This is a subset of the **europarl** part, but one word, written in capital letters in the prompt, was emphasised, i.e. pronounced with a focus.

- **prince**: A selected continuous passage from *Le Petit Prince*, with a length of about 240 words with some interrogative sentences and some direct and indirect discourse. The text was presented as a single prompt to ensure consistency in the prosody. The speaker was asked to read it with involvement.

3.2.2 Speaker Selection

Candidates were asked to read a short excerpt of *Le Petit Prince* of Antoine de Saint-Éxupéry in at least two of the four languages. The speakers were then selected based on their foreign accentedness: at least 3 experts in linguistics rated the strength of the candidates' foreign accent on a scale from 0 to 3 (0 for strong foreign accent, 1 for noticeable foreign accent, 2 for very slight foreign accent and 3 for no foreign accent). To be selected, the speakers had to obtain an average evaluation of at least 2.5 and no evaluation below 2. Eventually, 36 speakers were selected : 22 bilingual and 14 trilingual. Figure 3.1 gives the number of speakers recorded for each language pair.

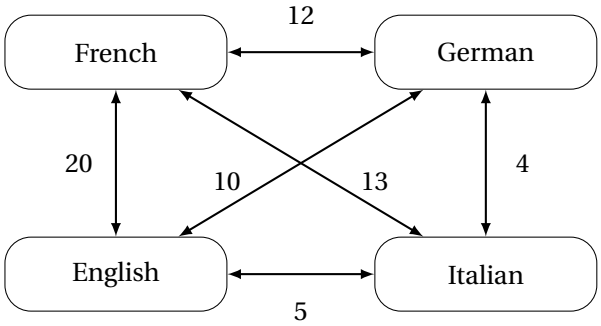

Figure 3.1 – Mapping between language pairs. The number of bilingual speakers recorded for each pair is indicated on the arrows.

3.2.3 Recordings

The recordings took place in a anechoic booth in which were placed a dynamic microphone MX418/C SHURE at 10-20 cm from the speaker with a pop shield, and a keyboard to control the prompts scrolling. The prompts were visible to the subject on a screen outside of the booth. A clone screen was visible to the operator to supervise the session. The sound device USBPre2 was used to record the signal into a 44.1 kHz mono 16bit format.

The SpeechRecorder 4 software (from the Institute of Phonetics and Speech Processing of the Ludwig-Maximilians-Universität München) was used to present the prompts one by one. The prompts were randomly mixed within the 4 first parts (i.e except the **prince** part which was presented as a single prompt). The prompt was presented to the speaker who could take a few seconds to read it mentally. Then, he or she pronounced it and had to press a key to either jump to the next prompt or re-record the same prompt. Redoing the same prompt was done in case of stuttering, hesitation or wrong reading. The speakers usually realised they had to restart the same prompt by themselves. Nevertheless, the operator could also ask them to do so.

3.2.4 Database Content

Table 3.1 summarises the content of the database. Note that the labels correspond to HTS-like labels[1], and their automatic alignment was done using TTS average voice models for each language. The emphasis marks on the labels were added manually, at the word level.

Table 3.1 – Recording numbers, durations and aligned labels.

Language	Sessions	Prompts	Total dur.	Labels	With emphasis	Words
French	31	5332	512 min.	4474	440	61815
English	22	3771	350 min.	3597	303	41023
German	17	2903	266 min.	2561	276	25660
Italian	16	2738	287 min.	—	—	—
Total	86	14744	1415 min.	10632	1019	128498
			~ 23.6 hrs.			

3.3 Single Speaker French Database

Few databases exist with sentences containing emphasised words. In English, one of the recent Blizzard Challenge datasets provided a high quality somewhat expressive speech corpus [Karaiskos et al., 2008]. The speaker, *Roger*, was asked to utter some of the sentences with a specific emphasis on one or two words, as described in Chapter 2, Section 2.3. While their approach allows covering all the possible diphones in an emphasised word, the context of this emphasised word is limited to a specific scenario.

To enable studies of emphasis intonation in a more variable context, we designed a speech corpus which contains emphasised words in many different contexts. The emphasised data collected was a part of a bigger high quality French speech database whose primary target application is speech synthesis. Below, we describe the database design, recordings and give some statistics about the data.

3.3.1 Text Material

The text selected for recordings consists of six parts:

- **parl**: consists of 4500 isolated sentences from French parliament debates.
- **book**: consists of 3500 isolated sentences from French novels.
- **siwis**: consists of 75 sentences from the *SIWIS* database.
- **sus**: consists of 100 semantically unpredictable sentences.
- **emph**: consists of 1575 sentences taken from the 4 other sets.

[1] A description is available in the documentation of the HTS demo, available at http://hts.sp.nitech.ac.jp/

- **chap**: consists of a full book chapter.

These 6 subsets serve different purposes. A detailed explanation on the text collection and preparation, as well as the purpose of each set is given below.

parl With the primary goal being the construction of TTS systems, the text was taken from debates between June 2012 and July 2015 at the French parliament. The main advantages of using such data are its free and open access, and its contemporary aspect: the vocabulary used in the debates reflects the current language in the society. The text data was downloaded from the French national assembly website[2]. A first selection was done to get the best diphone coverage as possible, using a greedy algorithm. A selection based on the sentence length, removing the shortest and longest sentences, allowed reducing the sentence set. Many sentences with the same recurrent structure were manually removed, for instance sentences like "La parole est à Monsieur David Douillet."

book In order to have copyright free material, five French books were selected from two authors: *Zadig* and *Micromegas* from Voltaire, and *Voyage au centre de la Terre*, *L'Île mystérieuse* and *Vingt mille lieues sous les mers* from Jules Verne[3]. After removing very short and very long sentences, a first random sentence selection was made. A manual checking was done in order to remove the sentence containing names which pronunciation could be ambiguous.

siwis This set corresponds to the sentences from the *SIWIS* database for which the speakers were asked to emphasise specific words. Recording the same data allows us to have parallel data from one database to the other. The sentences come from the Europarl corpus.

sus These sentences were partly taken from the *SIWIS* database, for which 20 such sentences were selected and recorded. The remaining 80 sentences were generated on a sentence generator website[4]. The generator randomly produces sentences respecting grammatical rules. The set of generated sentences was manually checked to remove sentences which made sense. This subset was recorded for testing, as semantically unpredictable sentences are good candidates to evaluate intelligibility for instance.

emph These sentences are aimed at studies on emphasis. In a similar fashion to the *SIWIS* database, this allows collecting the same sentences in 2 different styles: one in a neutral style, and one in which the speaker is asked to emphasise a specific word more than the rest of the

[2]The URLs follow the syntax http://www.assemblee-nationale.fr/14/cri/2012-2013/20130001.asp, where 14 stands for the 14th term of office, cri for integral report, 2012-2013 for the current session year, 2013 for the current civil year, 0 for ordinary session, 001 indicates that it is the first session of the year.

[3]The books are freely available on http://beq.ebooksgratuits.com

[4]http://romainvaleri.online.fr/

utterance. The 1575 sentences consist of 800 sentences from **parl**, 600 sentences from **book**, all the 75 sentences from **siwis**, and all the 100 sentences from **sus**.

chap This part is a full chapter taken from the French book *Vingt mille lieues sous les mers* from Jules Verne (Chapter III of the book). It amounts to approximately 1800 words. The fact that the sentences are not isolated like in the other sets makes it attractive for studies on higher level prosody. The style of reading, because of the material which contains both dialogues and narration, makes this part of the database more expressive by design.

3.3.2 Speaker and Recordings

The recordings were conducted by a professional voice recording agency called Voice Bunny[5]. The speaker was selected from a pool of native female French voice talents, by the author of this thesis who is a native French speaker. The instructions were provided to the speaker as follows:

- For **parl**, **book**, **siwis**, **sus**: read each sentence in an isolated manner, with a long pause (> 2 s.) between each sentence.
- For **emph**: read the sentence with focus on the indicated word, e.g.:
  "**Lourde** [read out this bold word with emphasis] *erreur, madame la ministre !*"
- For **chap**: read the full chapter in an expressive manner, without long pauses between sentences.

The sentences with emphasis were recorded after their neutral version, in order not to influence the speaker to reproduce the patterns that they were asked to produce in that case. Finally, the book chapter was read in one session, in order to have the dependencies that one can expect when reading a long text, e.g. the gradual downdrift of intonation along a paragraph, and reset when starting a new paragraph.

The data was recorded and provided in 44.1kHz mono 16 bits. Adobe Audition[6] was used for processing. 30 minutes of recordings generally required 90 minutes of processing, editing and checking from the voice actress. In total, 23 sessions were necessary to complete the recordings.

3.3.3 Database Content

Table 3.2 gives the amount of recorded speech in terms of utterances and time. The times without silences were estimated based on the automatic alignment performed on the contextual labels, and correspond to the times between the start and end of speech (meaning

[5] https://voicebunny.com
[6] http://www.adobe.com/products/audition.html

that the mid-sentence pauses are counted as speech). The first five parts of the database were segmented by finding long silences, and keeping short silences before and after actual speech; the full chapter was not segmented.

Table 3.2 – Amount of speech data recorded.

Style	# of sentences (sessions)	Time (inc. silences)	Time (no silence)
parl	4500 (6)	4h02	3h44
book	3500 (11)	5h00	4h45
siwis	75 (1)	3.8min	3.5min
sus	100 (1)	4.5min	4min
emph	1575 (5)	1h33	1h26
chap	—[7] (1)	10.5min	10.5min
Total	9750 (23[8])	10h54	10h13

Figure 3.2 illustrates the position of the emphasised words, both in an absolute and relative manner. The relative position was simply obtained by dividing the position by the total number of words, including intermediate silences. Many emphasised words are located in the second half of the sentence, but their absolute position is generally lower than 10, mostly because the majority of the sentences are short. There are in total 1695 annotated emphasised words, for 1575 sentences (some sentences had multiple emphasised words).

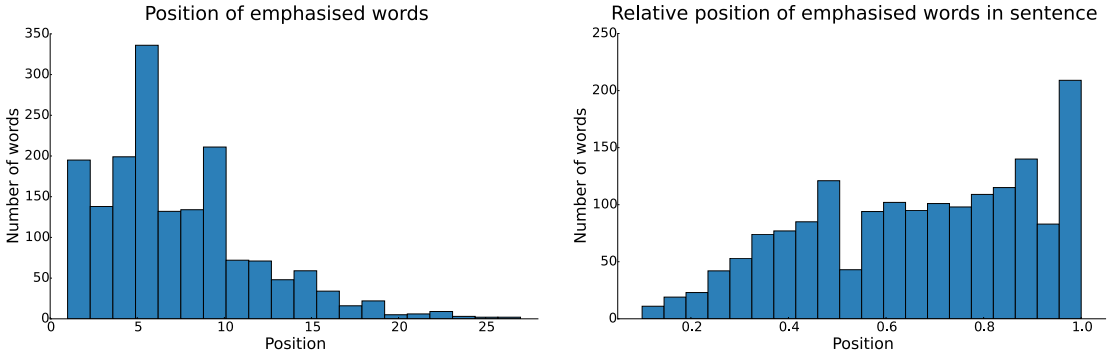

Figure 3.2 – Emphasised word positions. Left: absolute position, right: relative position in the sentence.

Table 3.3 gives the distribution of the emphasised words by number of syllables. 95% of the emphasised words contain 3 or fewer syllables.

Table 3.3 – Number of syllables in emphasised words.

# syllables	1	2	3	4	5	6	Total
# words	803	568	246	67	10	1	1695

When looking at the word-level context as defined by the HTS label format, the 1695 words

[8] This part was not segmented in sentences.

[8] sus and emph were recorded in the same session as one of the book sessions.

correspond to 1450 different contexts (to have the same context, 2 words need to have exactly the same number of syllables, position in the phrase and utterance, part-of-speech, etc.).

3.4 Summary

This chapter presented the design, recording and content of two new freely available speech databases.

The first one contains about 24 hours of speech, from 36 bilingual or trilingual speakers who did not have a foreign accent in the recorded language. The languages of the database are English, French, German and Italian. Each speaker uttered about 170 sentences in two or three languages, where the sentences had the same meaning, making the corpus parallel for languages. Another feature of the corpus is the word-level emphasis acted by the speakers, in a parallel manner — both neutral and emphasised versions of the sentences are available. The database is freely available for research [9].

The second database is a high quality French database, containing speech from multiple styles. Its primary purpose is speech synthesis, but it also contains sentences with emphasis on specific words, in many contexts. It will be released with no restrictions on the usage.

Both databases are used in the following chapters of the thesis, for studies on intonation and emphasis, in the context of speech-to-speech translation and intent preservation.

[9] Available at: http://bit.ly/siwisData.

4 Swiss French Accents in TTS Adaptation

In this chapter, we take a look at the regional accents of French native speakers in Switzerland. For S2ST in the Swiss context, regional accents (and dialects, in the case of Swiss German) can have an important role in interactions. Although there are no big differences between standard French accent and Swiss French accent, native listeners are generally able to distinguish easily the origin of the speaker.

French speech resources are not rare, as far as metropolitan France French accent, which is considered as standard accent and defined as "Français de référence" by Morin [2000], is concerned. On the other hand, Swiss accented French speech is more scarce. For this reason, we are interested in the possibility to adapt standard French TTS systems to Swiss accent.

We assess the accentedness of natural and synthetic speech through subjective evaluation. Using speech synthesis and analysis by synthesis methods, we modify both segmental and suprasegmental aspects of standard French synthetic acoustic parameters, and show the limits of speaker adaptation techniques to emulate Swiss French accent with little data. We compare standard French HMM-based TTS output with models adapted to Swiss French accent, and with the same synthetic speech augmented with natural Swiss French prosody.

The contributions presented in this chapter were originally published in the following conference papers:

- Pierre-Edouard Honnet, Alexandros Lazaridis, Jean-Philippe Goldman, and Philip N. Garner. Prosody in Swiss French accents: Investigation using analysis by synthesis. In *Speech Prosody*, Dublin, Ireland, May 2014

- Pierre-Edouard Honnet and Philip N. Garner. Importance of prosody in Swiss French accent for speech synthesis. In *Nouveaux cahiers de linguistique française*, September 2014

4.1 French Accents in Swiss Regions

4.1.1 Regional Accents in Automatic Speech Processing

In ASR, regional or foreign accents and dialects of a language bring variations that decrease performance of systems. It was shown by Huang et al. [2001] that the two main sources of inter-speaker variation were gender and accent. Kat and Fung [1999] proposed two solutions to overcome the variability introduced by accent: using a wide training database which includes accented data, or building accent-specific systems which will be used according to the accent of the speech to be recognised. In the literature, there are many other attempts to tackle the accent issue (mainly for non-native accented speech) in ASR by using adaptation techniques [Aalburg and Hoege, 2004; He and Zhao, 2003; Liu and Fung, 2000]. More generally, ASR systems are often confronted with non-native accents, and need to counteract effects of the accent component.

Conversely, in TTS, producing accented speech is desirable for some applications like S2ST, foreign language learning and dialect synthesis. Synthesising accented speech is still a quite new and challenging area. In most cases, different accents are modelled separately using different training data. There is only limited recent work on regional accent adaptation in TTS. Astrinaki et al. [2013] proposed interpolation of TTS models using closest speakers to a chosen geographical position. In this way, the English voice has average characteristics of these speakers, representing the specific regional accent. Another work by Gutierrez-Osuna and Felps [2010] consists of generating intermediary accent transformations between native and foreign speakers, to evaluate pronunciation of learners (in the context of computer assisted pronunciation training). This research, which is some kind of interpolation between native and non native accented speech, can be seen as part of TTS for under-resourced languages and cross-lingual speaker adaptation for TTS.

There are some other areas of automatic speech processing that are concerned with regional accents, for instance accent identification [Hanani et al., 2013; Huang et al., 2007; Omar and Pelecanos, 2010; Teixeira et al., 1996]. Swiss French accents have been investigated in this aspect by Lazaridis et al. [2014a], using speaker identification techniques. Prosodic features for accent discrimination were also investigated [Lazaridis et al., 2014b].

4.1.2 Peculiarities of Swiss Accents

The perception of different regional accents in a language can result from several sources. In French, the accents vary because of different factors according to the regions. For instance, there are noticeable differences at the pronunciation level of some phones between “Français de Référence” (FR) defined by Morin [2000] as standard pronunciation, and Canadian French (or Quebec French) [Côté, 2012]. As far as Swiss accents are concerned, between FR and Swiss French the differences in pronunciation are limited, and would rather express through prosody than in pronunciation. Speakers from these two categories are geographically close,

and “Romandie”, also called “Suisse romande” (i.e. the French speaking part of Switzerland) would not be distinguished from Eastern and Southeastern France linguistically according to Knecht [1979].

We are interested in the adaptation of standard French TTS systems to regional Swiss accents. The following sections investigate both segmental and suprasegmental variations between FR and Swiss French, and their perception by native French speakers.

4.2 Segmental Variation Perception

4.2.1 Pronunciation of Swiss French

Segmental variations correspond to the pronunciation of basic sound units, most of the time phonemes. In this section we investigate the adaptation of a standard average TTS voice trained with FR speech to Swiss accented speakers. This enables the study of how segmental variations are perceived when adaptation is performed.

It is commonly agreed that there is not *one* Swiss French accent, but Swiss *accents* [Martinet, 1971]. Accents inside Suisse Romande vary from one canton (administrative region) to another. It is even known that people from one city can distinguish accents from others cities in the same canton [Andreassen and Lyche, 2009]. On the other hand, the various Swiss accents share some common peculiarities.

Métral [1977] gave an overview of the segmental aspects of Swiss accents. The main differences that are observed between FR and Swiss accent at a global level, meaning by considering all the Swiss accents together, concern the openness of some vowels. An example is the pronunciation of [œ] which becomes [ø], as in “jeune” (young, in English) which becomes “jeûne” (which in standard French is the word for fasting in English). This type of confusion happens in final syllables, when the [œ] is not before [r], for all the regions of Romandie. The survey conducted by Métral [1977] is somehow biased as acknowledged, due to the social status of the persons interrogated on the distinction between open and closed vowels: the subjects with less education actually tend to confuse vowels more than the majority of the subjects in that work. There are other vowel pairs for which the distinction is different between FR and Swiss accents, either on the open / close or on the front / back dimension. Table 4.1 gives a few examples of pronunciations that can be found in Swiss regions as opposed to FR underlined by Metral.

Differences can also be observed in the prosody of both types of French, these aspects are discussed in Section 4.3.

Table 4.1 – Examples of pronunciation difference between FR and Swiss French vowel pronunciations and how they differ.

Vowel in FR (diff)	Vowel in Swiss French (diff)	Example (Swiss pronunciation)
[œ] (open-mid)	[ø] (close-mid)	jeune (jeûne)
[ɛ] (front open-mid)	[ə] (central close-mid)	messe (meùsse)
[e] (close-mid)	[ɛ] (open-mid)	école (ècole)
[o] (back)	[œ] (front)	abricot (abrikeù)

4.2.2 Perception of Swiss vs French Pronunciations

We want to evaluate how Swiss and French pronunciation of French are perceived when performing model adaptation, in terms of regional accentedness. Our hypothesis is that segmental level adaptation is not enough to emulate the Swiss accent. To demonstrate it, we compare multiple systems, which are first adapted with Swiss speech data and then supplemented with real prosody. The evaluation is done through subjective listening tests. We expect that listeners will perceive the Swiss accent more when adapting the system with Swiss speech, but that the "true" natural prosody is necessary to perceive the accent as strongly as in the original speech.

Adaptation to Swiss Accent

To evaluate the effect of adaptation of a TTS system to Swiss regional speech, we use a standard French statistical parametric speech synthesis (SPSS) system and adapt its acoustic parameters using a speaker adaptation strategy. Then, to assess how correct prosody affects accentedness, prosodic parameters are modified to match those of original speech. Figure 4.1 gives an overview of the complete procedure. Raw data is in grey, features in red, the models are in green and the final outputs are in blue. The three outputs correspond from left to right to:

- the output of the models after adaptation to Swiss accent
- the output of the models after adaptation when providing time alignment
- the output of the models after adaptation when providing time alignment and the original F_0 (extracted from natural speech).

These three outputs are compared against vocoded original speech from the Swiss speakers, and with the average output of the standard French system.

Data

The data used comes from two databases: the BREF database [Lamel et al., 1991] and a part of the PFC database [Durand et al., 2009] with additional content [Avanzi, 2014].

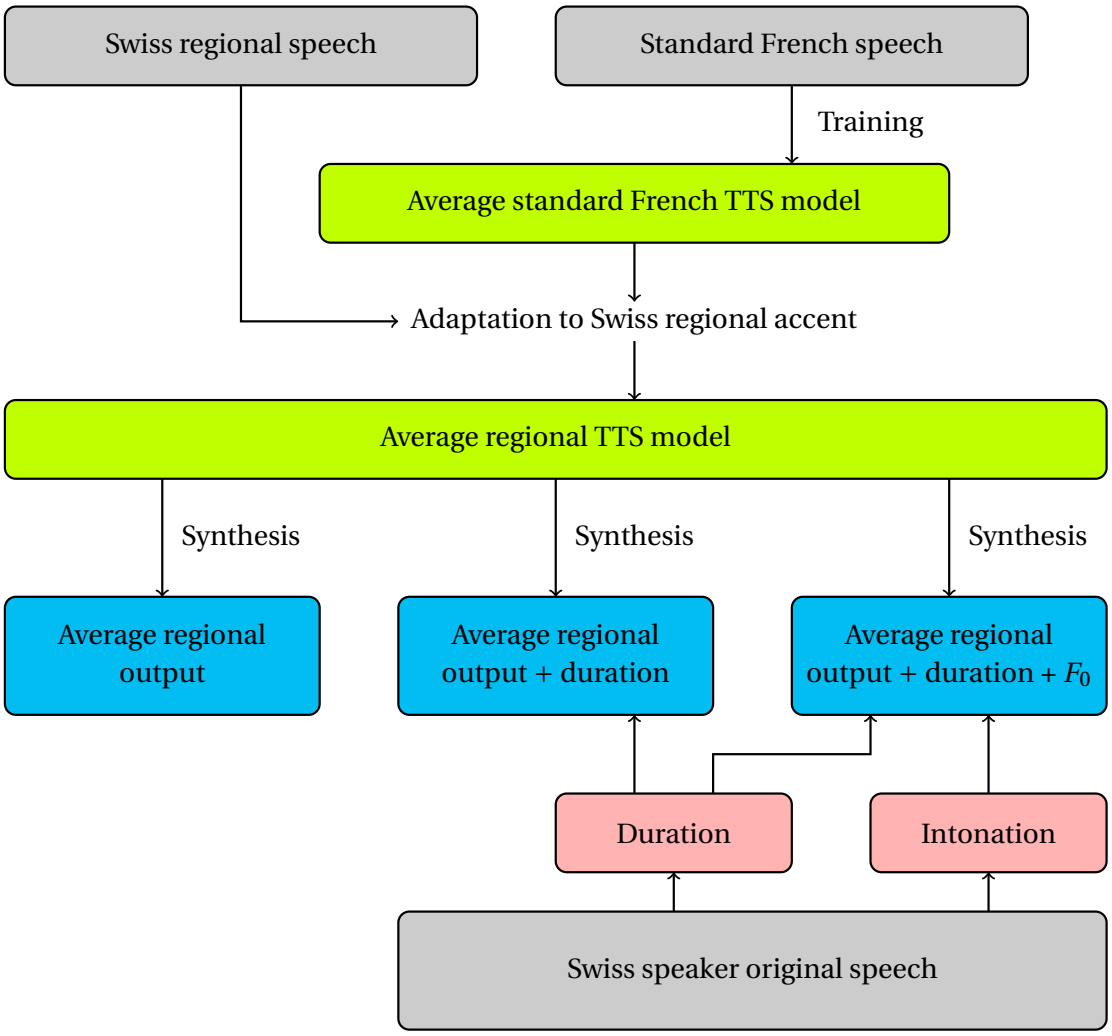

Figure 4.1 – Adaptation of standard French TTS system to Swiss French accent, using original prosody. Raw data is in grey, features in red, the models are in green and the final outputs are in blue.

These two databases are described in more details in Chapter 2. As a quick reminder, BREF is a read speech corpus with text from the French newspaper *Le Monde*. The speech from 10 male speakers was used for this work. The PFC corpus is composed of read speech and free conversational speech. In this work, the common content read speech was used; 12 speakers were selected among the 20 male speakers available, originating from 5 cities: Paris (France), Geneva (Geneva canton), Martigny (Valais), Neuchâtel (Neuchâtel canton) and Nyon (Vaud).

Experimental Settings

Training The average TTS models were trained on a subset of the BREF database composed of 6857 sentences (about 12 hours of speech) from 10 male speakers. We used 39 mel cepstral

coefficient with energy coefficient, log F_0, 21 band aperiodicities extracted every 5 milliseconds with the STRAIGHT vocoder [Kawahara et al., 1999] and their first and second order derivatives. 5 emitting state left-to-right HSMMs with no skip were trained with full-context labels using the version 2.1 of HTS [Zen et al., 2007] and speaker adaptive training [Yamagishi and Kobayashi, 2007]. The training resulted in an average male standard French voice.

Adaptation The adaptation was done for each city, yielding 5 systems. For each group of speakers, we used 20 sentences per speaker, leaving a test sentence out for evaluation. For Paris, 2 speakers were used, for Geneva 3 speakers, for Martigny 1 speaker (in this case it is a standard speaker adaptation), for Neuchâtel 2 speakers and for Nyon 4 speakers.

Synthesis The same sentence was synthesised for each of the 5 adapted systems:

> "*La côte escarpée du mont St Pierre, qui mène au village, connaît des barrages chaque fois que les opposants de tous les bords manifestent leur colère.*"

The choice of the sentence was done according to previous studies on Swiss accent evaluation [Avanzi et al., 2013; Racine et al., 2013]. It was segmented manually and the orthographic transcriptions were corrected manually before full-context label creation (adding pauses and hesitations). Features were extracted from Swiss French data the same way as for the training data. The trained TTS models were then used to estimate the duration of Swiss speech data.

For each synthetic file, three versions were created as depicted in Figure 4.1.

To add duration information, we first extracted the duration information from the original waveforms using forced alignment: given the speech features, their corresponding transcription (full-context phonetic labels in our case) and some French TTS models, the Viterbi algorithm was used to estimate phone and state boundaries. Using the state duration information, a forced-aligned synthesis was performed, i.e. parameter generation given the known state sequence (it means, if we refer to the way synthesis is done from the models, as described in chapter 2, estimating only equation (2.7), as $\mathbf{q}$ is imposed). The resulting speech was composed of synthetic parameters, but aligned in time with the original speech, i.e. the phoneme durations were the same as original ones.

For the system with duration information and intonation, time alignment was also performed, and we replaced the synthetic intonation ($\log F_0$) with the original one. After vocoding, the output was a speech signal composed of synthetic spectrum and aperiodicity coupled with original duration and intonation. The reason for using both original intonation and duration is that it is not possible to use only original intonation, because the other parameters (spectral information) have to be aligned with the excitation part to reconstruct the speech signal.

The vocoded version, which just decomposes the speech and reconstructs it, was used to have a reference of the accentedness of original files. We used a vocoder to emulate the best possible parametric speech synthesis.

Subjective Evaluation

A listening test was conducted in order to evaluate the degree of accent of each version of the sentence for each system and each speaker. For this purpose, a webpage was built enabling subjects to listen to:

- the average model output (1 sample)
- the adapted model output (5 samples)
- the adapted model output with original duration (12 samples)
- the adapted model output with original duration and intonation (12 samples)
- the vocoded file (12 samples)

The vocoded version is perceptually very close to the original recorded speech as only an analysis and resynthesis is performed. For each file, the listeners had to give a degree of Swiss accent between 1 and 5, 1 being "no accent" and 5 "strong accent" (in the instructions, "no accent" was defined as *standard accent* and close to Paris accent). The listeners could listen to the files as many times as they wanted.

4.2.3 Results

19 native French speakers participated in the study. 7 subjects were Swiss (mainly from Vaud and Valais), the 12 remaining were all French. Among the participants, 4 were females and 15 were males. Figure 4.2 shows the average perceived degree of accent for the different systems. For each speaker, the left-most bar (in black) corresponds to the output of the TTS models adapted to the speaker's regional accent, and the right-most (in green) corresponds to the original sentence from the speaker vocoded. The red bar corresponds to the average perceived accent when adding duration information, and the blue one when adding duration and intonation. For comparison, the average voice output had an average score of 1.42 (standard deviation 0.49).

The first general observation that can be made is that in all the cases, adding duration increases the perceived degree of accent compared to the adapted model outputs. In the majority of the cases, adding original intonation increases the perceived accentedness compared to the version with duration information. The version with original prosody (duration and intonation) are the closest to the original vocoded files in 67% of the cases (8/12).

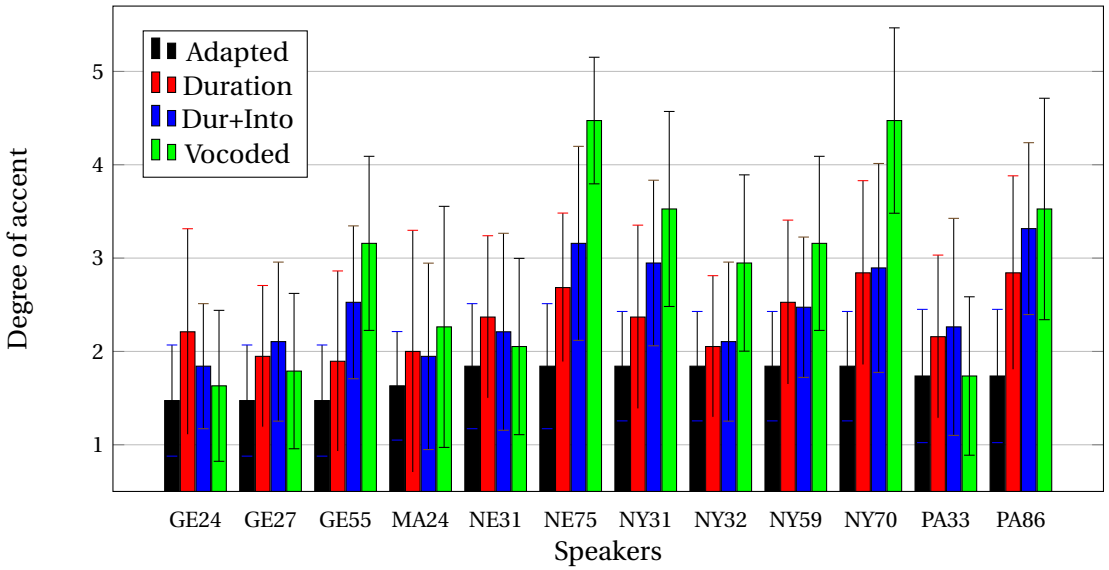

Figure 4.2 – Mean degree of accent for each version of the sentence for the 12 speakers. Black is adapted output, red with duration, blue with duration and intonation, green is original vocoded.

A Wilcoxon signed rank test was performed for each speaker between each pair among the four versions, as the data is ordinal [Clark et al., 2007]. Because the points on the scale have an order but cannot be considered as equally spaced, it is valid to calculate means for the data, but not statistically meaningful to compare them. However, the comparison can be done on the medians. We found that for every speaker, there was no significant difference between the scores obtained for the vocoded version and for synthesis with duration and intonation information, no significant difference between vocoded version and for synthesis with duration information, and no significant difference between version with duration and version with duration and intonation. That means that adding duration or adding duration and intonation allows an approximation of the real accentedness level for all the speakers.

In all the cases, there was a significant difference between the vocoded version and the adapted output with no prosodic information ($p-\text{value} < 0.05$). For all the speakers from Geneva and Neuchâtel, there was a significant difference between the adapted version, and the version with duration information. The same observation was made for 2 out of the 4 speakers from Nyon. Speakers from Paris and the speaker from Martigny did not show significant difference between these two systems, which can be expected as the system already has “Paris accent” for the Parisian speakers, and standard speaker adaptation was performed for the Martigny speaker, so duration models were adapted to his voice. Finally, when comparing the adapted output with the version with duration and intonation, all the speakers from Geneva, Nyon and Paris showed significantly different scores, while for speakers from Neuchâtel and Martigny, the scores did not differ significantly. The same reason as for the duration case could explain the similarity for speaker from Martigny, however it does not explain why speakers from

Neuchâtel are not perceived significantly differently in these two settings.

Table 4.2 gives the mean distance between the different combinations for each speaker. It is simply calculated using the sum of absolute difference for each sample on all listeners' scores:

$$\text{Distance}_k(S1,S2) = \frac{1}{L}\sum_{l=1}^{L}\left|\text{score}_{k,l}(S1) - \text{score}_{k,l}(S2)\right| \tag{4.1}$$

where $S1$ and $S2$ are the two systems being compared, k the speaker, L the number of listeners, and $\text{score}_{k,l}(S)$ the score of the sample from system S for speaker k given by listener l.

Table 4.2 – Mean distances between configurations per speaker

Systems	*GE*24	*GE*27	*GE*55	*MA*24	*NE*31	*NE*75
ave-dur	0.95	1.16	1.63	0.89	1.00	1.11
ave-int	1.11	1.00	1.79	0.74	0.58	1.47
dur-int	0.89	1.11	0.89	0.58	0.95	0.89
ada-dur	0.89	1.11	1.47	0.89	0.89	1.00
ada-int	1.16	1.16	1.63	0.63	0.58	1.26
ave-ada	0.47	0.47	0.47	0.74	0.53	0.53
ave-voc	1.68	1.68	1.47	1.42	1.32	1.74
ada-voc	1.63	1.63	1.53	1.21	**0.89**	1.53
dur-voc	1.16	1.26	**0.89**	**1.16**	1.37	1.47
int-voc	**0.89**	**1.00**	1.05	1.21	0.95	**1.00**

Systems	*NY*31	*NY*32	*NY*59	*NY*70	*PA*33	*PA*86
ave-dur	1.21	1.00	1.16	1.00	0.89	1.79
ave-int	1.53	1.21	1.47	0.79	1.00	1.63
dur-int	0.63	0.42	0.74	0.53	0.63	1.21
ada-dur	1.00	0.89	0.74	0.89	0.79	1.37
ada-int	1.21	1.11	0.95	0.89	0.79	1.53
ave-ada	0.53	0.53	0.53	0.53	0.63	0.63
ave-voc	1.89	1.53	1.58	1.58	1.68	2.63
ada-voc	1.58	1.11	1.16	1.26	1.58	2.21
dur-voc	1.00	1.16	0.95	1.42	1.00	1.37
int-voc	**0.79**	**0.95**	**0.84**	**1.11**	**0.89**	**1.21**

"ave" corresponds to average voice output, "ada" to adapted model output, "dur" to adapted voice with original duration, "int" to adapted voice with original duration and intonation, and "voc" corresponds to vocoded.

The last 4 lines are the distances between each of the system under test and the vocoded samples, which can be seen as targets. The average trend is a reduction of the distance with the vocoded version when adapting, a further reduction when adding duration information

and finally when adding intonation. This is in accordance with the absolute scores presented in Figure 4.2. In 9 cases, the closest to vocoded is the system adapted along with all prosodic information. In 2 cases, adding only duration yields the closest score to vocoded speech. Finally, in one case, adaptation alone gives the closest result to vocoded speech: the speech from this speaker ($NE31$) was perceived as slightly accented only, and his intonation was relatively neutral compared to other speakers. If we compare with the other speaker from the same region, $NE71$, we can see that that speaker has a very strong accent, which may have influenced the adaptation enough to perceive the same degree of accent as $NE31$.

Table 4.3 gives the mean differences between systems over all speakers, showing the average trends presented in the previous table; with the reduction of the distance between scores of the different systems with the vocoded samples when adding prosodic cues.

Table 4.3 – Mean distances between configurations

	average	adapted	duration	intonation	vocoded
average	0	0.55	1.15	1.19	1.68
adapted		0	0.99	1.08	1.44
duration			0	0.79	1.18
intonation				0	0.99
vocoded					0

If we measure the absolute differences between scores per speaker, we observe that from average standard French output to regional accent adapted output with original duration and intonation, the mean distance is reduced by 41%. If we only use duration, the reduction is of 30%. In comparison, with only adaptation, and no correction of prosody, the distance is reduced by only 14%.

We see that perceptually, adding prosodic cues increases significantly the degree of accent of the synthetic speech. Even though we cannot conclude that there is no difference between our adapted models supplanted with original prosody, in terms of accentedness, the systems are not perceived significantly differently. As far as adapted models are concerned, the significant difference observed between the output of the models adapted to regional accents and the original vocoded speech demonstrates our hypothesis, which was that adaptation is not enough to perceive accents as strong as the original.

4.3 Suprasegmental Variation Perception

After looking at the segmental variations between accents, we observed that prosodic aspects play an important role in accent perception. In this section we investigate the fusion of standard French pronunciation with Swiss French prosody.

4.3.1 Perception of Swiss Prosody

There are some divergences, as often in the area of prosody, on the rhythm topic, i.e. Swiss speakers are known to speak slower than French. Miller [2007] showed that on read speech samples, the speaking rate was the same for French and Swiss (from Vaud canton) speakers, but the articulatory rate (excluding pauses) was slower for Swiss speakers. French speakers use more pauses, which decreases their speaking rate. Schwab and Racine [2013] recently led an empirical study to verify whether Swiss people indeed speak slower than French people or not. The findings showed that pause frequency and duration were not different among some French, Belgian and Swiss speakers. However, articulation rate was found to be slower for Swiss speakers.

Schwab et al. [2012] compared two Swiss regional accents with French accent, regarding penultimate accentuation, showing that Swiss speakers are more likely to accentuate penultimate syllables than French speakers. Variations were also observed among Swiss regions with different strategies in expressing prominence on these syllables.

These two aspects of regional accents were further investigated by Avanzi et al. [2012] using the PFC corpus: irrespective of the style of speech (read speech or free conversational speech), for French speakers, 5.25% of the penultimate syllables in clitic groups were identified as prominent, while for Swiss speakers, the percentage of prominent penultimate syllables was between 11.47% and 15.13%, being 2 to 3 times more than for French speakers. The accentuation of penultimate syllables is mainly expressed through intonation and energy. To be rigorous, one should actually say that Swiss tend to accentuate both penultimate and ultimate syllables at the intonation level, with an increase of intensity on the syllable preceding the *tonic*, i.e. the ultimate syllable in standard French. Métral [1977] affirmed that the intonation patterns vary from one canton to another, and that the realisation of penultimate accentuation is less present in Valais French, where the intonation patterns are different from the rest of Romandie. Concerning speech rate, French and Swiss from Martigny were found to speak faster than the other Swiss speakers (Geneva, Neuchâtel, Nyon). Another general observation made was that older Swiss speakers tend to speak slower than younger ones, while for French variety no significant different was found related to age. Finally, one of the conclusions was that slower speech rate correlates with the perception of more penultimate syllables, probably because when the speech is faster it is more difficult to perceive them.

As a consequence of a higher tendency to accentuate penultimate syllables, Swiss speakers are often said to produce more variations in their intonation, however it is hard to study the phenomenon due to the variety of intonation patterns. By accentuating different syllables, they generate different intonation patterns that may sound more “lively” or “singing” to French listeners.

4.3.2 Simulating Swiss Prosody in French Speech

We are now interested in the perception of Swiss accent when its prosody is mixed with standard French pronunciation. Our hypothesis is that by having the original Swiss prosody, even if the pronunciation comes from FR models, listeners will perceive the Swiss accent. In other words, as we previously observed that adapting TTS models was not enough to perceive Swiss accents, we now evaluate how using standard French pronunciation with Swiss prosody is perceived.

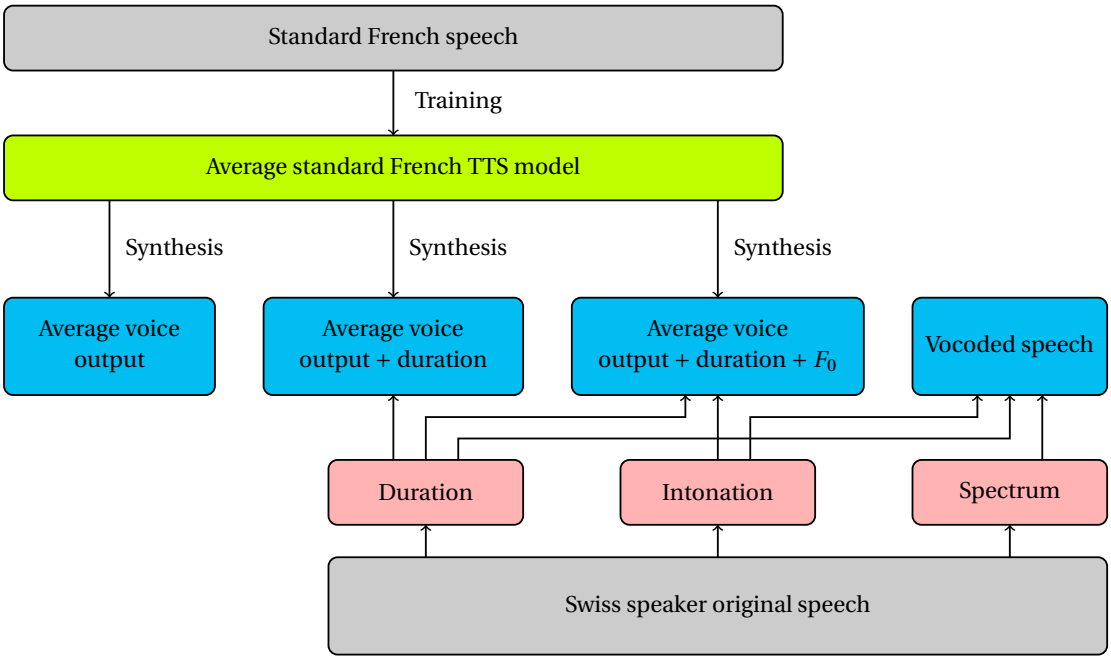

Figure 4.3 – Using original prosody to simulate Swiss accent in standard French TTS. Raw data is in grey, features in red, the models are in green and the final outputs are in blue.

Figure 4.3 shows the different outputs after altering the synthetic prosody, and the vocoded speech. The difference with the systems in Section 4.2 is that in this case, there is no adaptation of the TTS model to Swiss accents. As we use standard French pronunciation, we keep the average standard French model, and only the prosodic parameters are replaced with Swiss ones.

Data and Models

The data used for these experiments were the same as in Section 4.2 for training the TTS models. As there was no adaptation involved in these experiments, we only used the same test sentence for the same 12 speakers.

For the synthesis part, the baseline was the average standard French voice trained on BREF data. The acoustic features were extracted the same way as in the previous section.

Subjective Evaluation

The same common sentence as in the previous section was selected for the 10 Swiss and 2 French male speakers from our PFC dataset. The same (manually corrected) full-context labels were used for forced-aligned synthesis.

A listening test was conducted in order to evaluate the degree of accent of the file generated as described in Section 4.2.2. For this purpose, a webpage was built enabling subjects to listen to:

- 1 completely synthetic file (output of TTS model)
- 12 files with original duration (1 per speaker)
- 12 files with original duration and intonation (1 per speaker)
- 12 vocoded files from original speech (1 per speaker)

which sums up to 37 files in total. As in the previous experiment, the vocoded version allows simulating the best possible synthesis, with a lessened vocoder effect in the results. As in the previous experiment, for each file, the listeners had to give a degree of Swiss accent between 1 and 5, 1 being "no accent" and 5 "strong accent". The listeners could listen to the files as many times as they wanted and the test took approximately 10 minutes.

4.3.3 Results

28 subjects took the test. Among them, there were 17 males and 11 females, 23 were French and 5 were Swiss (from Vaud, Valais, Neuchâtel and St Gallen).

Figure 4.4 shows the mean and standard deviation of the three versions of the file for each speaker; the fourth version displayed in black, which is identical for all the speakers, corresponds to the average voice output. The means and variances show that when adding intonation and duration the values get closer to the vocoded version than just adding duration, and modifying only duration gives closer values than the average voice output, as we expected. For the speakers with highest degree of accent (based on the vocoded version), $NE75$, $NY31$, $NY32$, $NY59$ and $NY70$ ($PA86$ has different behaviour), the means of the *intonation + duration* version is still much lower than the vocoded one. $PA86$ is a 86 year old Parisian and although he does not have a Swiss accent, his accent was perceived as strong. The average voice being based on French accent and pronunciation, he has the same pronunciation as the average voice. Adding the prosody resulted in a degree of accent close to the original, explained by both correct prosody and pronunciation.

These results are confirmed by a Wilcoxon signed rank test which was performed for each speaker between each pair among the four versions presented (3, and the baseline average voice).

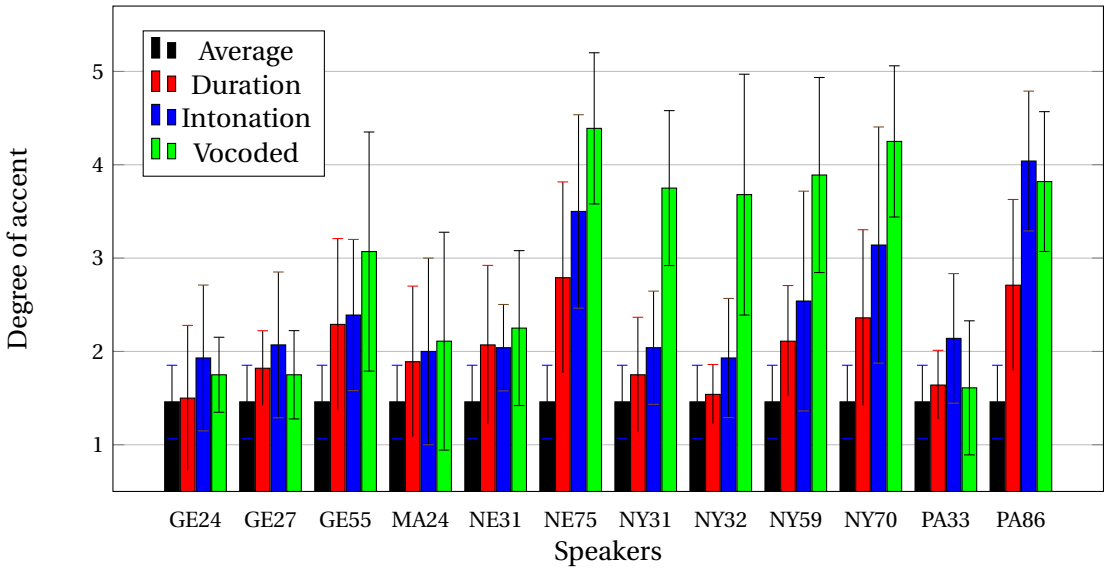

Figure 4.4 – Mean degree of accent for each version for the 12 speakers. Output of average TTS, with duration information, with duration and intonation, vocoded version.

In the case of average version versus vocoded version, 9 out of the 12 speakers have significantly different scores ($p-\text{value} < 0.01$); $GE24$, $GE27$ and $PA33$, corresponding to the least accented speakers, are not significantly different.

In the case of the version with duration information versus the vocoded version, 7 still have significantly different scores: $MA24$ and $NE31$ are not significantly different.

Finally, when adding original intonation, the 5 speakers mentioned earlier as *very different* ($NE75$, $NY31$, $NY32$, $NY59$ and $NY70$) from the vocoded version are still significantly different. In that case, the other 7 are not significantly different from the vocoded version (including $GE55$ and $PA86$).

Table 4.4 shows the means of absolute differences between scores per speaker. For each speaker, a comparison was made between 2 versions of the file among the average voice output (ave), the version including duration (dur), the version including duration and intonation (int) and the vocoded version which is the reference (voc). In 8 cases out of 12, the combination of duration and intonation is closer to the vocoded version (values in bold). The 4 other cases give the advantage to the version including only duration information.

Table 4.5 gives the global absolute difference between each system. The last column gives the distance between the vocoded speech and the other versions. We can see that between the average voice output and the version with duration information we reduce the distance to the vocoded version by 20%, between the version with duration and the version including duration and intonation, the reduction is 11% and the overall improvement from average to duration and intonation version gives 29% improvement. A Wilcoxon signed rank test confirmed that

Table 4.4 – Mean distances between configurations per speaker

Systems	*GE*24	*GE*27	*GE*55	*MA*24	*NE*31	*NE*75
ave-dur	1.04	0.82	0.96	0.96	0.96	0.86
ave-int	1.29	1.32	1.25	1.18	1.21	1.04
dur-int	0.61	0.79	1.07	1.14	0.75	0.68
ave-voc	1.68	1.93	1.54	1.25	1.57	1.54
dur-voc	**0.93**	1.54	**1.36**	1.64	**0.96**	1.04
int-voc	0.96	**1.25**	1.57	**1.07**	1.00	**1.00**

Systems	*NY*31	*NY*32	*NY*59	*NY*70	*PA*33	*PA*86
ave-dur	0.89	0.96	0.93	0.79	1.29	0.64
ave-int	1.29	1.14	1.61	1.39	1.50	1.29
dur-int	0.61	0.96	0.89	1.11	1.00	0.93
ave-voc	1.75	1.89	1.60	1.60	2.21	1.79
dur-voc	**1.29**	1.57	1.39	1.39	1.64	1.57
int-voc	1.39	**1.46**	**1.21**	**1.14**	**1.21**	**1.21**

the differences between score absolute differences were significant (p – value < 0.01 in the 3 cases).

Table 4.5 – Mean distances between configurations

	average	duration	intonation	vocoded
average	0	0.93	1.29	1.70
duration		0	0.88	1.36
intonation			0	1.21
vocoded				0

It demonstrates that prosody plays an important role in Swiss accent perception. However, for the most accented speakers, prosody alone is not enough to obtain the same degree of accent. In these cases, adequate pronunciation is required to perceive the Swiss accent. This is backed up by the fact that accented Parisian speech can be produced with standard French pronunciation and specific prosody.

The low number of Swiss subjects did not allow the evaluation of the difference in accent perception between French and Swiss listeners, but the numbers showed the same trends for both groups.

4.4 Conclusion

In this chapter we investigated Swiss French regional accent perception in the context of speech synthesis. The adaptation of TTS models from average standard accent to Swiss regional accents proved not to be sufficient to let native French listeners perceive the Swiss

accent. The experiments demonstrated our first hypothesis, that the standard speaker adaptation techniques could not adapt the prosodic characteristics of the Swiss speakers.

Our second hypothesis was that Swiss prosody mixed with standard French pronunciation would be perceived as Swiss accented. When analysing the perception of Swiss accent when using only Swiss prosody along with standard French pronunciation, we observed that using the real duration consistently increases the degree of accent perceived by the listeners, and that using the real intonation increases the perceived degree of accent even more. Therefore, this hypothesis was only partially demonstrated: using only prosody increased the perception of French accent, however prosody alone was not sufficient to emulate regional accents, especially when the speaker's accent was strong.

By combining speaker adaptation and prosody modification, we managed to synthesise speech with a degree of accent perceived as not significantly different from the real accented speech. This means that prosody should be dealt with in a different manner from the acoustic parameters which evolve at the segmental level. The incapacity of the models to produce adequate prosody prevents the variations observed in regional accents to be perceived when synthesising speech. For this reason, the remainder of this thesis concentrates on intonation modelling and in the next chapter, we propose a new intonation model to attempt to better model speech intonation in the context of S2ST.

5 Intonation Modelling

In this chapter, a novel physiologically based intonation model using perceptual relevance is introduced. As has been underlined in Chapter 4, correct prosody is needed to achieve the synthesis of regional accents. Furthermore, we are motivated by the possibility to translate prosodic events, e.g. make a (group of) word(s) prominent in the target language when its corresponding (group of) word(s) was emphasised in the source language.

In our approach, the matching pursuit (MP) algorithm is used to decompose the intonation contour. We introduce a perceptually relevant weighting function in the decomposition process, to extract perceptually relevant components. The components — high order damped system impulse responses — are physiologically plausible and can be compared with the components of the command-response model [Fujisaki and Nagashima, 1969]. In this chapter, along with the model, a simple automatic method is proposed to extract its parameters. The physiological aspect of the proposed model is interesting as it makes the model theoretically language independent. The model is evaluated on speech from three languages and multiple speakers.

The work presented in this chapter was a piece of collaborative work with Dr. Branislav Gerazov and Aleksandar Gjoreski, both based in the Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University, Skopje, Macedonia, and originally published in the following papers:

- Pierre-Edouard Honnet, Branislav Gerazov, and Philip N. Garner. Atom decomposition-based intonation modelling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4744–4748, Brisbane, Australia, April 2015. IEEE

- Branislav Gerazov, Pierre-Edouard Honnet, Aleksandar Gjoreski, and Philip N. Garner. Weighted correlation based atom decomposition intonation modelling. In *Proceedings of Interspeech*, pages 1601–1605, Dresden, Germany, September 2015

5.1 Background

The need for correct intonation in TTS systems as well as the more general study of intonation have motivated the creation of different intonation and / or prosody models. In the context of TTS, adaptive systems — almost exclusively statistical parametric speech synthesis (SPSS) — are of great interest in the research community. The current state of the art systems for SPSS are based on HMMs [Tokuda et al., 2002b; Zen et al., 2009] and DNNs [Zen et al., 2013]. HMM-based speech synthesis deals with intonation in a frame-wise manner: each frame from the training speech database has a value — or a null value in the case of an unvoiced frame — and HMM states are trained using these values. At synthesis time, F_0 is generated frame by frame, based on the HMM parameters, and its suprasegmental aspect is modelled using decision trees. This results in a speech often qualified as "flat" or lacking expressivity, which is due to the oversmoothing of HMMs [Toda and Tokuda, 2005].

There are three main ways of tackling the flatness of HMM-based speech synthesis at the intonation level: post-process the synthetic intonation coming from HMMs, use a different representation of F_0 in the HMMs or use an external prosody model that combines with other HMM parameters. A more detailed review of some of the most standard models is given in Chapter 2, Section 2.2. As a reminder, we give an overview of some of these models.

1. Within the framework of statistical paremetric speech synthesis, F_0 is typically handled in the same way as other acoustic features, that is frame-wise. In HMM-based TTS, multi-space probability distribution (MSD)-HMMs are used to take into account the fact that speech can be voiced or unvoiced [Tokuda et al., 2002a]. More recently, some work was done using continuous F_0 and it was shown that continuous F_0 improves the perceived naturalness of synthesis [Latorre et al., 2011; Yu and Young, 2011]. Another approach was proposed recently using continuous wavelet decomposition to separate the different levels of variation in F_0 [Suni et al., 2013]. In DNN-based synthesis, F_0 is modelled like the other acoustic features, and the input context given to the DNN contains similar linguistic information, and some additional features related to the position of the current frame (position in the phone, syllable, word, etc.).

2. In the second category, Hirose et al. [2011, 2012] proposed to use the command response (CR) model [Fujisaki and Nagashima, 1969] to estimate the F_0 model commands from linguistic information, and then optimise them according to the F_0 generated by HMMs. The goal was to increase the expressivity and to make some segments more prominent in the synthetic speech by altering the extracted intonation commands. Another attempt to integrate the CR model in HMM-based TTS was made by Hashimoto et al. [2012], where parameterised F_0, generated by the CR model — and therefore smoothed contour — was used for training the HMM intonation features, to avoid modelling noise.

3. The external prosody models, or intonation models are numerous. Among them, Hirst et al. [2000] model the intonation contour as a sequence of specific F_0 target points.

The *Tilt* model describes it as a sequence of events with specific shapes that can be automatically extracted with an obvious resynthesis step [Taylor, 2000]. Another model derived by Bailly and Holm [2005], called superposition of functional contours (SFC), is a data driven approach which is based on the superposition of elementary contours extracted with the use of neural networks. The first two models try to directly model intonation with no attempt to understand its underlying production process. SFC, on the other hand, mostly relies on metalinguistic information. Only a few models actually try to explain the intonation by investigating its production aspect. The most famous model in this category is the command response (CR) model of Fujisaki and Nagashima [1969]. This model decomposes the intonation into additive physiologically meaningful components.

5.2 Physiology of Intonation Production

In mimicking the abilities of humans in a machine, it is natural to try to mimic human physiological processes. Doing so is attractive as a technological advancement, and can be seen as an attempt to understand the underlying processes. Furthermore, such a model provides theoretical language independence. The vocal instrument of humans obviously does not depend on the language they speak, therefore, the way their muscles control vocal folds — consequently intonation — should not change from one language to another.

5.2.1 Cricothyroid Muscles and F_0

Fujisaki [2006] describes the F_0 contour as the superposition of multiple components in the log domain. By relating the tension of the vocal folds with their length, the author derives log F_0 as the sum of: a base component (related to the size and density of the membrane) which is assumed to be constant for a given speaker, speaking style and emotional state, and two time varying components related to the activation of the muscles controlling the vocal folds. The first is a global phrase component and the second is a sequence of local accent components. In the CR model, these two components are associated to the activation of two parts of the cricothyroid (CT), both generating a rise in the F_0 through a slow translatory movement and a fast rotary movement of the thyroid cartilage, respectively.

The CR model also allows for negative phrase and accent commands. Negative phrase commands are used to model the phrase final drops in F_0 in some languages [Fujisaki and Hirose, 1984; Hirose and Fujisaki, 1982]. Negative accent commands are necessary for the modelling of tonal languages such as Mandarin and Thai, as well as languages with pitch accents, such as Swedish and Bengali [Fujisaki, 2004, 2006; Fujisaki et al., 1993, 1998; Saha et al., 2011]. Both of these negative components are attributed to the opposite rotary movement of the thyroid, which in turn is credited to the thyrohyoid (TH) muscle [Fujisaki, 2006].

5.2.2 Other Muscles Related to Vocal Fold Control

Collier [1975] investigated relations between muscles and F_0 by analysing electromyographic (EMG) activity in laryngeal muscles and air pressure, by sampling them simultaneously. In accordance with Fujisaki's findings [Fujisaki, 2006], the CT muscle was found to be responsible for most of the major F_0 changes (both rising and falling), while subglottal pressure was found to control the gradually falling baseline of F_0. The sternohyoid (SH) and thyrohyoid (TH) were found to have no or negligible effect on F_0.

Later, a detailed analysis of intonation production was given by Strik [1994]. In this work, four physiological sources of F_0 change were identified by assessing their influence on pitch using measurements that included (EMG) recordings of the relevant laryngeal muscles:

1. Cricothyroid (CT) muscle – rotates the thyroid cartilage with respect to the cricoid, stretching the vocal folds and raising F_0,
2. Vocalis (VOC) muscle – found within the vocal folds, its contraction decreases vocal cord length, but increases their tensile stress, the net effect being a rise in F_0 [Titze and Martin, 1998],
3. Sternohyoid (SH) muscle – one of three strap muscles used to alter the position of the larynx; lowers the larynx decreasing vocal cord tension and F_0,
4. Subglottal pressure (P_{sb}) – increased P_{sb} is found to linearly correlate to increased F_0. The measurements presented by Strik [1994] show that the CT and VOC activations are correlated and cause a rise in F_0 , as do peaks in P_{sb}. By contrast, the activation of SH coincides with drops in F_0 . Another important observation to point out is that only the P_{sb} signal has a global component, while the others feature only local ones.

5.3 A Generalised Command-Response Model

5.3.1 The Command-Response Model

The command-response model, by Fujisaki and Nagashima [1969], is one of the most interesting models as it provides a physical meaning to pitch and has several components taking into account medium and short term variations in pitch. It therefore fits in our target model, which should account for pitch production and its understanding. It assumes that the pitch (more precisely, the logarithm of fundamental frequency) is the superposition of a base value (fixed for a speaker), phrase components (slow varying) and accent components (short term variations). These components are the response to phrase and accent commands, occurring at the beginning of a phrase (impulse) and during a syllable (step function) respectively. The CR model has been used successfully to model the intonation of multiple languages, e.g. Japanese [Fujisaki and Hirose, 1984], Swedish [Fujisaki et al., 1993], Chinese [Fujisaki et al., 2000], Bengali [Saha et al., 2011].

Its mathematical formulation is the following:

- One second-order, critically-damped linear filter in response to an *impulse-like phrase command*

$$G_{pi}(t) = \begin{cases} \alpha_i^2 t e^{-\alpha_i t} & \text{if } t \geq 0 \\ 0 & \text{if } t < 0 \end{cases} \tag{5.1}$$

  where α_i^2 is the natural angular frequency of the i^{th} phrase control mechanism component G_{pi}, and is assumed to be constant within an utterance.

- Another second-order, critically-damped linear filter in response to a *stepwise accent command*

$$G_{aj}(t) = \begin{cases} \min[1-(1+\beta_j t)e^{-\beta_j t}, \gamma] & \text{if } t \geq 0 \\ 0 & \text{if } t < 0 \end{cases} \tag{5.2}$$

  where β_j^2 is the natural angular frequency of the j^{th} accent control mechanism component, assumed to be constant within an utterance, while the maximum threshold γ is typically set to 0.9.

- The full logarithmic F_0 contour is given by:

$$\ln F_0(t) = \ln F_b + \sum_{i=1}^{I} A_{pi} G_{pi}(t - T_{0i}) + \sum_{j=1}^{J} A_{aj}\{G_{aj}(t - T_{1j}) - G_{aj}(t - T_{2j})\} \tag{5.3}$$

  where F_b is the bias level, I is the number of phrase components, J is the number of accent components, A_{pi} is the magnitude of i^{th} phrase command, A_{aj} is the amplitude of j^{th} accent command, T_{0i}, T_{1j} and T_{2j} are the time of the impulse for the phrase command i, onset and offset times of the current accent command j respectively.

Several methods have been proposed to extract the parameters, e.g. those of Agüero and Bonafonte [2005]; Agüero et al. [2004]; Kameoka et al. [2010]; Mixdorff [2000]; Narusawa et al. [2002], but the task proved to be difficult, and is still a matter for research [Torres and Gurlekian, 2016].

5.3.2 Generalised Components

The command-response model was successfully implemented in a TTS framework using specific topology HMMs by Kameoka et al. [2015]. Substates were introduced to model the duration of each state, and each state modelled prosodic events such as the step function for accent components and impulse for phrase components. By translating the CR model into a

probabilistic model, Kameoka et al. [2015] were able to successfully extract model parameters and to generate them in the context of TTS.

Based on the fact that substate HMMs can be used to model a step function, and that the step response function to the accent command is equivalent to the impulse response to a train of impulses, we can say that a step function is equal to a sequence of impulses if the impulses are separated by only one frame: for all the time steps between the beginning and the end of the step function, the value of the signal is one; it is zero elsewhere. Accent components could then simply be modelled using the same type of damped system and replace step functions by impulses as for phrase component. From a command-response point of view, signals are carried from the brain to muscles in nerves by means of impulses rather than by absolute levels. The typical response of a muscle to such impulses is a muscle twitch, as depicted in figure 5.1. This can be seen as the lowest muscular activity unit, then the contraction of a muscle would be attributed to sequences of impulses with a period shorter than that of the twitch.

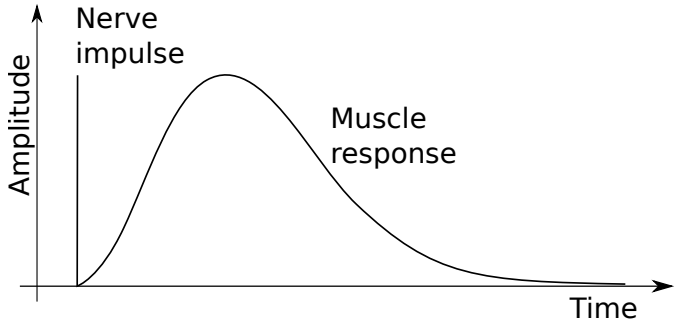

Figure 5.1 – Muscle twitch response to a nerve impulse.

The matching pursuit (MP) algorithm of Mallat and Zhang [1993] is a good candidate to decompose a signal in its basic elements, in our case muscle response to impulses. The next section introduces matching pursuit and its application to intonation modelling. Because of the way MP extract components from a signal, in our case the F_0 contour, a sequence of impulse responses to imitate the response to a step function is practically unlikely. To relax the constraints on the parameters introduced for the CR model, and as a way to generalise it in accordance with the possibility that more than 2 muscles control the vocal cords — and consequently the F_0 — we use the following damped systems:

$$G_{k,\theta}(t) = \frac{1}{\theta^k \Gamma(k)} t^{k-1} e^{-t/\theta} \quad \text{for} \quad t \geq 0 \tag{5.4}$$

Notice that this is the definition of a gamma distribution. The order k was assumed to be 2 in the CR model. Plamondon [1995] advocates the use of the log-normal distribution shaped response rather than the gamma distribution. This comes from the central limit theorem:

the log-normal arises as a limiting case of many impulses travelling some distance from the brain the muscle, the muscle itself being a compound of contractile fibers. Higher order gamma distribution shaped functions tend to log-normal, and it should be noted that they are indistinguishable down to a small level of detail. Prom-on et al. [2009] showed that higher orders better model the vocal fold tension control. Following these conclusions, the order of the impulse response we use is relaxed to be higher than the original order 2. Also, the flexibility introduced by using different scale parameters θ allows wider atoms. These atoms prove to be able to model flat accents, especially with several impulses. To relate with equation (5.1), $k = 2$ and $\theta = 1/\alpha$.

5.3.3 Matching Pursuit and Weighted RMS

Decomposition using Matching Pursuit

The matching pursuit algorithm [Mallat and Zhang, 1993] allows approximation of a signal as a linear combination of kernel functions — or atoms — taken from a dictionary. This dictionary is of fixed size, and contains predefined possible atoms. Figure 5.2 gives an example of dictionary, using the functions defined in equation (5.4), where $k = 6$ and θ can take the different values given in the legend. In an iterative manner, the algorithm finds the atom with the best correlation with the signal and then subtracts its weighted version until some desired accuracy is reached. The position of this best fitting atom is found by sliding all the possible atoms frame by frame, and its amplitude is directly given by the correlation with the signal to be decomposed. This process reduces the reconstruction error by local optimisations. For a given signal $s[n]$, and given a set of unit-normed kernel functions $\{\phi^m\}$, the signal would be decomposed as:

$$s[n] = \sum_{k=1}^{K} c_k \phi^k[n - n_k] + e[n] \tag{5.5}$$

where K is the number of atoms used to reach the desired accuracy, c_k is the correlation between the residual signal at step $k-1$ and $\phi^k[n - n_k]$; $e[n]$ is the residual. There can be different stopping criteria for the decomposition, such as the norm of the residual, the number of atoms, the correlation between the reconstruction and original signal, or the absolute amplitude of the last extracted atom.

The similarity with equation (5.3) resides in that the F_0 decomposition used in the CR model is a sum of kernel functions, where the functions are impulse and step function responses. We use the matching pursuit to decompose F_0 contours; the atoms used are of the form of equation (5.4). As it is based on correlation, the MP algorithm applied to intonation decomposition, using gamma distribution-shaped atoms, will result in modelling of both voiced and unvoiced regions of speech. It means that in unvoiced regions, where F_0 does not

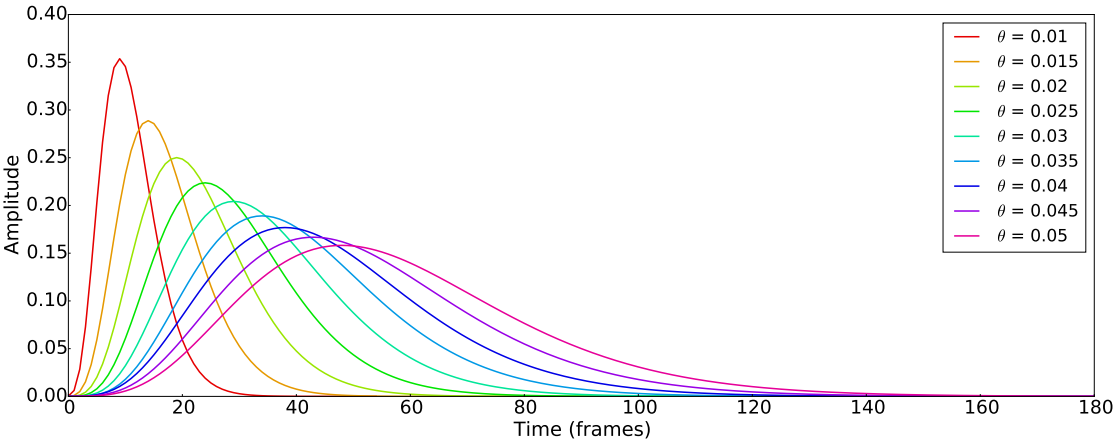

Figure 5.2 – Gamma distribution for $k = 6$, for various θ values.

(necessarily) exist, MP will try to match atoms to the interpolated values; it is hence likely to fit some errors from the pitch tracker. The (erroneous) atoms extracted in unvoiced regions will then have repercussions in surrounding voiced regions: small atoms will be extracted to correct the errors introduced by wrongly extracted atoms. The details are given in the following section in which we integrate perceptually relevant information in the decomposition process.

Selection of Atoms Based on Weighted RMSE

As discussed above, one obvious issue which arises when simply applying matching pursuit to a continuous F_0 contour is the unnecessary modelling of unvoiced parts of F_0. Most intonation models use discontinuous pitch trackers and then interpolate unvoiced regions using, for instance, spline interpolation [Yu and Young, 2011]. Then some methods concentrate the effort of modelling on the voiced parts, e.g. Mixdorff [2000]; Narusawa et al. [2002]; while others simply model everything equally, e.g. Hirst et al. [2000]; Taylor [2000]. To avoid the latter, one needs a way of assessing which parts of the F_0 contour are perceptually relevant, to focus on their modelling.

Two perceptually relevant objective measures of F_0 contour similarity — the weighted root-mean-square (WRMS) distance (5.6), and the weighted correlation (WCORR) coefficient (5.7) — were introduced by Hermes [1998].

$$WRMSE = \sqrt{\frac{\sum_i w(i)(f_1(i) - f_2(i))^2}{\sum_i w(i)}} \tag{5.6}$$

$$WCORR = \frac{\sum_i w(i)(f_1(i) - f_{1m})(f_2(i) - f_{2m})}{\sqrt{\sum_i (w(i)(f_1(i) - f_{1m})^2) \sum_i (w(i)(f_2(i) - f_{2m})^2)}} \tag{5.7}$$

Here f_1 and f_2 are the two F_0 contours that are being compared, f_{1m} and f_{2m} are their respective means, and $w(i)$ is the weighting function. The weighting function is defined as the maximum amplitude of the subharmonic sumspectrum (SHS), which is a weighted sum of the harmonics contributing to the pitch that was introduced by Hermes [1988].

The two proposed measures were aimed at automating the evaluation of student performance when teaching intonation [Hermes, 1998]. The results showed that the measures correlated well with the similarity categorisation done by five experienced phoneticians. Namely, the WRMSE was found to have a correlation of 0.679, to the experts' visual ratings, while the WCORR correlated better, at 0.67, with their auditory ratings. In both cases the inter-expert agreement was at 0.69 and 0.65, for the two tasks, showing that higher correlations cannot be obtained. Moreover, approximate thresholds were calculated for classifying the perceptual similarity of two intonation contours using the objective measures. The thresholds for WCORR are given in Table 5.1.

Table 5.1 – Weighted correlation thresholds for perceptual similarity of two F_0 contours found by Hermes [1998].

Category	WCORR	Perceptual F_0 similarity
1	> 0.978	no differences
2	> 0.946	differences audible
3	> 0.896	differences clearly audible
4	> 0.827	linguistic differences
5	< 0.827	completely different

In our work we modify the WRMSE and WCORR to assess the perceptual similarity of our modelled pitch contour compared to the originally extracted F_0. The modification are detailed below. The weighted RMS error (WRMSE) and the weighted correlation were calculated according to equations (5.8) and (5.9). Here f_0 is the reference F_0, $\hat{f}_0$ is the modelled F_0, i.e. its reconstruction, f_{0m} and $\hat{f}_{0m}$ are their respective means, and $w(i)$ again is the weighting function.

$$WRMSE = \sqrt{\frac{\sum_i w(i)(\hat{f}_0(i) - f_0(i))^2}{\sum_i w(i)}} \tag{5.8}$$

$$WCORR = \frac{\sum_i w(i)\hat{f}_0(i)f_0(i)}{\sqrt{\sum_i w(i)f_0(i)^2 \sum_i w(i)\hat{f}_0(i)^2}} \tag{5.9}$$

In our implementation, the measures are modified as follows:

1. We do not normalise the F_0 contours with their mean, as no offset is to be expected in our application scenario. In Hermes' work, the comparison was done between different speakers, so it is expected that the contours would have different means and variances related to speaker dependency.

2. We define the weighting function to be Equation (5.10), where $p(i)$ is an approximation of the probability of voicing (POV), as defined by Ghahremani et al. [2014], and $e(i)$ is the energy contour of the speech signal. This is in accordance with newer trends in perceptual intonation studies [d'Alessandro et al., 2011; Rilliard et al., 2011]. It makes sense as regions of speech with higher energy and higher probability of voicing will have more impact on the perception of intonation — and speech, more generally. The introduction of a continuous POV estimate in equation (5.10), allows the elimination of hard thresholds that were used to determine voicing by d'Alessandro et al. [2011] from our algorithm, making it more robust.

$$w(i) = p(i)e(i) \tag{5.10}$$

In a first attempt to reduce the erroneous atom extraction, we used the WRMSE to give increased importance to the modelling of perceptually relevant segments of the F_0 contour [Honnet et al., 2015]. To this end, we introduced an atom selection algorithm that uses the WRMSE to keep only the perceptually significant atoms from the set of atoms extracted using the traditional matching pursuit algorithm. A summary of the procedure is given in Algorithm 1.

The outlined algorithm proved to be apt at eliminating the extraneous atoms generated with the MP framework. This allowed for improved intonation modelling using the introduced gamma distribution-shaped atoms. The performance of the algorithm was verified across three different languages and both on male and female speakers (one of each gender for each language, so a total of six speakers) [Honnet et al., 2015].

Nonetheless, eliminating atoms at will from the set generated by the matching pursuit algorithm (MP) raised inconsistencies in the modelling process. Namely, sometimes when an atom which did not contribute significantly to the WRMSE was eliminated from the set, its influence in voiced regions was also eliminated. This means that the atoms that the MP algorithm fitted after it were then lacking in accuracy when modelling the F_0 contour. In other words, the subsequent atoms were compensating for, or taking into account, an atom that was not there anymore.

Algorithm 1 Atom decomposition with weighted RMSE based atom selection.

procedure ATOM DECOMPOSITION WITH WRMSE SELECTION
    Extract F_0, energy and POV from waveform.
    Subtract $F_b = F_{0\text{min}}$.
    Extract F_{0p} using matching pursuit and subtract it.
    Extract atoms using matching pursuit:
Loop:
    if WRMSE ≤ Threshold **then**
        goto *End*.
    else
        if Atom decreases WRMSE by more than 0.001 **then**
            Keep the atom and **goto** *Loop*.
        else
            Discard the atom and **goto** *Loop*.
End.

5.4 Weighted Matching Pursuit for Perceptually Relevant Decomposition

To improve our previous approach, we incorporated the perceptually relevant F_0 contour similarity measures, this time the weighted correlation defined in equation (5.9), as a cost function directly into the matching pursuit framework [Gerazov et al., 2015]. The introduced weighted correlation atom decomposition (WCAD) algorithm directly extracts the atoms which are perceptually relevant, eliminating the need for subsequent atom selection. Another improvement in the algorithm is the introduction of a novel phrase atom extraction algorithm. These two key modifications make WCAD a more consistent, integrated algorithm, with added physiological plausibility.

5.4.1 Introducing a New Correlation Measure in MP

A summary of the weighted correlation atom decomposition (WCAD) algorithm is given in Algorithm 2. The algorithm integrates the weighted correlation in the calculation of the cost function of the matching pursuit algorithm, and accommodates the peculiarities of the phrase atom extraction.

At the start of the algorithm, the energy e and POV p are calculated from the waveform. These are then used to calculate the weighting function w using equation (5.10). Next, the phrase atom is extracted from the utterance. In concordance with Strik's findings, we fit a single phrase atom per breath group. In the first step we estimate the start and end times of phonation, t_s and t_e, by thresholding the energy e with a starting threshold value Th_s and a terminal threshold value Th_e. This is done by selecting the first of consecutive frames for which the energy is higher than the threshold. This is simply done to avoid modelling the noise in the F_0 contour before and after the speech. The time instant t_s is used to align the position of the

Algorithm 2 Weighted Correlation Atom Decomposition algorithm.

procedure WCORR ATOM DECOMPOSITION
    Extract f_0, e (energy) and p (POV) from waveform.
    Calculate w from e and p.
    Extract t_s and t_e of phonation, based on Th_s and Th_e.
    Find θ_f for F_{0p} (phrase atom) at position t_s that maximises WCORR · CORR for $t_s \leq t \leq t_e - t_{\text{off}}$.
    Calculate F_{0p} amplitude using CORR.
    $f_{\text{diff}} = f_0 - F_{0p}$.
    $f_{\text{recon}} = F_{0p}$.
Loop:
    Find (local) $Atom$ with maximum WCORR · CORR with f_{diff} for $t > t_s$.
    Calculate $Atom$ amplitude using CORR.
    Increment $Counter_{\text{Atom}}$.
    $f_{\text{diff}} = f_{\text{diff}} - Atom$.
    $f_{\text{recon}} = f_{\text{recon}} + Atom$.
    if ($\text{WCORR}_{\text{norm}}(f_{\text{recon}}, f_0) > \text{threshold}(\text{WCORR}_{\text{norm}})$) **then**
        goto *End*.
    else
        goto *Loop*.
End.

maximum of the phrase atom t_{rm} with the start of phonation in the utterance.

In the next step, θ_f is chosen to maximise the cost function calculated as the product of WCORR, as defined in equation (5.9), and the standard correlation function CORR, between the phrase atom and the F_0 contour. This product is used instead of using the WCORR itself as a cost function. It was introduced to circumvent deadlocks due to zeros in the CORR function occurring at places where the WCORR has local maxima.

The cost function was calculated within the range of F_0 between t_s and $t_e - t_{\text{off}}$, where t_{off} is an offset time introduced to eliminate the phrase-final fall and rise in intonation from the phrase atom fitting. The extracted phrase atom amplitude is calculated using the standard correlation, after which the phrase atom is subtracted from f_0 to give the difference f_{diff}. The phrase atom is also used to initialise the F_0 reconstruction f_{recon}.

In the next part of the algorithm, local atoms are extracted from f_{diff} in a loop. At each iteration, the atom that maximises the cost function WCORR · CORR the most is selected, disregarding the parts of f_{diff} before t_s and after t_e. Each atom is subtracted from f_{diff} before the next iteration, and also added to f_{recon}. The loop is repeated until either 1) the reconstruction WCORR reaches the selected threshold value, or 2) the chosen maximum number of atoms is reached.

We have chosen our stopping criteria to be the WCORR over the SNR (signal to noise ratio between the signal and the residual) used in a standard matching pursuit implementation, the

matching pursuit toolkit (MPTK) [Krstulovic̀ and Gribonval, 2006], because of its determined perceptual significance. Since the thresholds determined by Hermes [1998] are based on the WCORR calculated between the zero-mean versions of both f_0 and f_{recon}, we follow suit and calculate the WCORR using equation (5.7). This weighted correlation of the zero-mean normalised F_0 contours is labelled $\text{WCORR}_{\text{norm}}$. The $\text{WCORR}_{\text{norm}}$ is calculated for the part of the F_0 contour that was actually modelled by our WCAD algorithm, as bounded by t_s and t_e.

5.4.2 A New Phrase Component

The introduced phrase atoms are based on the qualitative shape of the global component of the subglottal pressure P_{sb} seen in the plots of the results obtained by Strik [1994]. There, the global component starts with a peak at the start of phonation and then steadily decreases towards 0 with a time constant relative to the length of the utterance. The rise-time is much shorter than the fall-time, reflecting the nature of the physiological production of the P_{sb}, in which an initial pressure build-up that precedes speech, is followed by its gradual release that sustains phonation. This complex behaviour is provided by the interplay of the diaphragm and the rib cage muscles.

The phrase atoms are a modified version of the local atoms defined in equation (5.4), in that they follow one time constant θ_r during their rise, and another θ_f during their fall (5.11). Looking at Strik's plots, one can observe that the rise-time of the P_{sb} is consistent to a certain extent across the different utterances [Strik, 1994]. Since we lack objective measurements to properly model the rise time, but we still need the rising part when modelling consecutive utterances, we arbitrarily use a fixed θ_r to represent a fast rise time across the phrase atoms. On the other hand, θ_f is chosen to maximise the cost function in the matching pursuit framework, basically fitting the F_0 as well as possible taking into account our perceptual measure. In equation (5.11), t_{rm} refers to the time instant in which the rising portion of the atom reaches its maximum, calculated according to equation (5.12). In the descending portion, the phrase atom starts from this maximum value and decreases towards 0. In order to compensate for the difference between t_{rm} and the maximum time instant t_{fm} of the fall function defined in (5.13), the time index t' is introduced in equation (5.11), calculated using (5.14).

$$G_{k,\theta_r,\theta_f}(t) = \begin{cases} \dfrac{1}{\theta_r^k \Gamma(k)} t^{k-1} e^{-t/\theta_r} & \text{for } 0 \le t \le t_{rm} \\ \dfrac{1}{\theta_f^k \Gamma(k)} t'^{k-1} e^{-t'/\theta_f} & \text{for } t > t_{rm} \end{cases} \tag{5.11}$$

$$t_{rm} = (k-1)\theta_r \tag{5.12}$$

$$t_{fm} = (k-1)\theta_f \tag{5.13}$$

$$t' = t - (t_{rm} - t_{fm}) \tag{5.14}$$

5.5 Model Evaluation

To assess the plausibility and comparative performance of the introduced Weighted Correlation Atom Decomposition model we have designed 2 experiments. The first one (detailed in Section 5.5.3) analyses its ability to accurately capture the intonation dynamics, as well as the relative number of atoms required to reach a set modelling accuracy. The second one (Section 5.5.4) compares our generalised CR model with a state of the art implementation of the standard CR model.

5.5.1 Data Selection

The experiments were conducted on a large number of files, including speech in three different languages from both genders. This selection aims to demonstrate the language independent aspect of the model. Four databases were used: the SI84 set of the WSJ corpus [Paul and Baker, 1992] and CMU Arctic databases [Kominek and Black, 2004] for English, BREF [Lamel et al., 1991] for French and Phondat [Hess et al., 1995] for German. More detail on each of these dataset can be found in Chapter 2, Section 2.4. The data can be seen as two main datasets:

- The CMU Arctic data consisted of two speakers: a male speaker, *bdl*, and a female speaker, *clb*. This set is aimed at evaluating the performance on the algorithm on the intra speaker variability aspect, as a lot of data from each speaker is available.

- The second set, using speech from many speakers and three languages, aims at evaluating the algorithm on multi-lingual and multi-speaker aspects.

On the first dataset, from the utterances recorded for these 2 speakers, we manually selected the ones for which the used pitch extractor [Ghahremani et al., 2014][1] gave reliable results. The validity of the F_0 contours was assessed through comparison with two other pitch tracker outputs: STRAIGHT [Kawahara et al., 1999] and SSP from Garner et al. [2013][2]. The final dataset totals 1,729 utterances with a duration of 1.5 hours.

[1] See: http://kaldi.sourceforge.net/
[2] Available at: https://github.com/idiap/ssp

On the second dataset, a first random selection of the sentences was made, including 7,085 sentences from WSJ, 15,981 from BREF and 21,587 from Phondat. To avoid using files for which the pitch tracker yields unreliable contours, we performed a pitch comparison using the same 3 pitch trackers: SSP [Garner et al., 2013], the STRAIGHT vocoder [Kawahara et al., 1999] and the Kaldi pitch tracker [Ghahremani et al., 2014]. For all the files, the pitch was extracted with these 3 tools, and RMSE and correlation were calculated for each pair (Kaldi vs STRAIGHT, Kaldi vs SSP, SSP vs STRAIGHT). The files for which correlation was lower than 0.99 or RMSE was higher than 50Hz for at least one pair were discarded. As a result, 2,453 files were selected for WSJ, 6,387 for BREF, and 4,433 for Phondat. Finally, to balance the subsets for each language, 8,964 files (about 12.6 hours of speech) were kept by discarding the shortest (sometimes corresponding to single words) and longest files. Details are given in table 5.2

Table 5.2 – Data used for the experiments.

Database	# of sentences	# of speakers (male/female)	Hours of speech
Arctic	1729	2 (1/1)	1.5
WSJ	2453	76 (37/39)	5.1
BREF	2799	23 (10/13)	4.9
PhonDat	3712	164 (78/86)	2.6
Total	10693	263 (126/139)	14.1

5.5.2 WCAD Algorithm Parameters

The parameters used in the WCAD algorithm were determined through qualitative assessment of its performance on a set of randomly chosen utterances from the first dataset (the CMU Arctic database). It is reasonable to suppose that the optimal parameters are speaker dependent, but for convenience we pool them together and assume speaker independence.

To determine the optimal order of our model k that is used in generating the gamma shaped phrase and accent atoms (5.4), the difference in WCAD performance for the various k-s was analysed. Figure 5.3 shows the average WCORR versus the number of atoms per syllable for k-s in the range 2–7, for the French female speaker group. The curves were obtained by averaging the values over the whole French female speaker dataset. The curves are smooth because of the antialiasing of the plotting program [3], and the large amount of data. Figure 5.7 shows the same measure for all the values for $k = 6$ with the mean curve. We can see that $k = 4,5,6,7$ generally gives better performance than $k = 2,3$. The same trend is observable for all the speaker groups. This is in line with the findings of Prom-on et al. [2009]. However, the high variance across the utterances makes it difficult to clearly favour one k. Moreover there is no plausible reason to use several orders in our model, so we assume that using order 6 is reasonable, as it gives a slightly better average performance than the k of 4 used in our previous work [Honnet et al., 2015], and the improvement when going to order 7 is small. For further discussion on the choice of order, the reader can refer to the work of Prom-on et al.

[3] We used the Matplotlib library for python, http://matplotlib.org/

[2009].

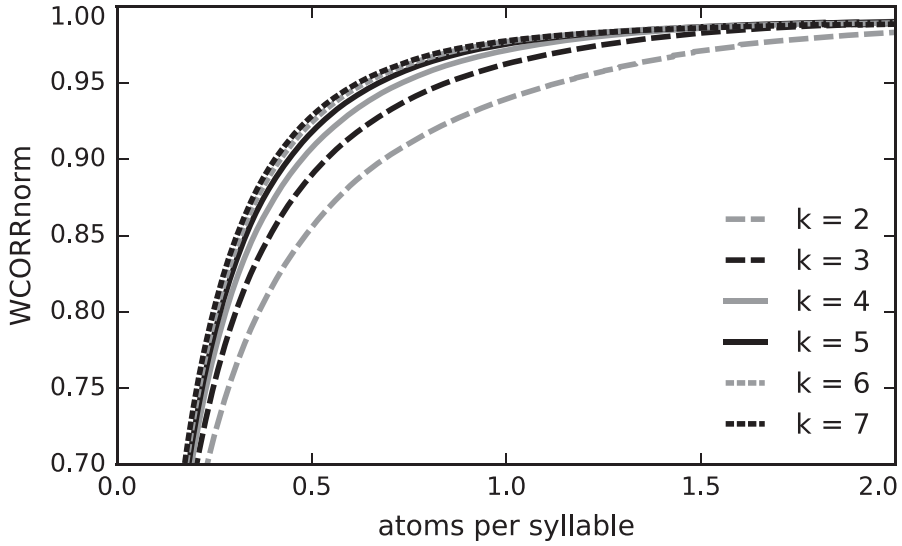

Figure 5.3 – WCORR vs number of atoms per syllable for the French female speakers for different values of k.

To determine the start and end of phonation, respectively t_s and t_e , we chose equal threshold values Th_s and Th_e of 0.01 for the normalised energy. The offset time t_{off} subtracted from t_e to leave out possible phrase-final falls and rises in F_0 was set to 150 ms. The θ_r for the rising part of the phrase atoms was fixed at 0.5. The range for the θ_f for the falling part of the phrase atoms was set to 0.1–10, and for the θ of local atoms to 0.01–0.05. This way, the constructed dictionaries provide an atom variability sufficient for the function of the WCAD algorithm. The maximum θ_f of 10 covers the long utterances with a slowly decreasing global P_{sb} component. And the θ range encompasses the area where the θ-s concentrate, as can be seen in the histogram of their distribution in Figure 5.4. The lower values of θ correspond to shorter atoms, which are mostly used for modelling sharper variations. The atoms using these low values have low amplitude; they help modelling the noise in intonation contours and getting higher accuracy in the reconstruction. Empirically, an alternative stopping criterion was set to a maximum of 10 atoms per second.

5.5.3 Model Performance

The plausibility of the WCAD algorithm is determined through assessing 1) how well it can model the F_0 contour, and 2) how many atoms it needs to do so. In order to determine this, we analyse the contribution of each of the atoms as they are added in each iteration of the modelling procedure. More specifically, we analyse how much does the addition of each atom increase the $\text{WCORR}_{\text{norm}}$ between the original and modelled F_0 contours. We use the $\text{WCORR}_{\text{norm}}$, in order to assess the perceptual quality of the modelled F_0 using the thresholds discussed in Section 5.5.2. To extract the continuous F_0 and POV estimates we use the pitch

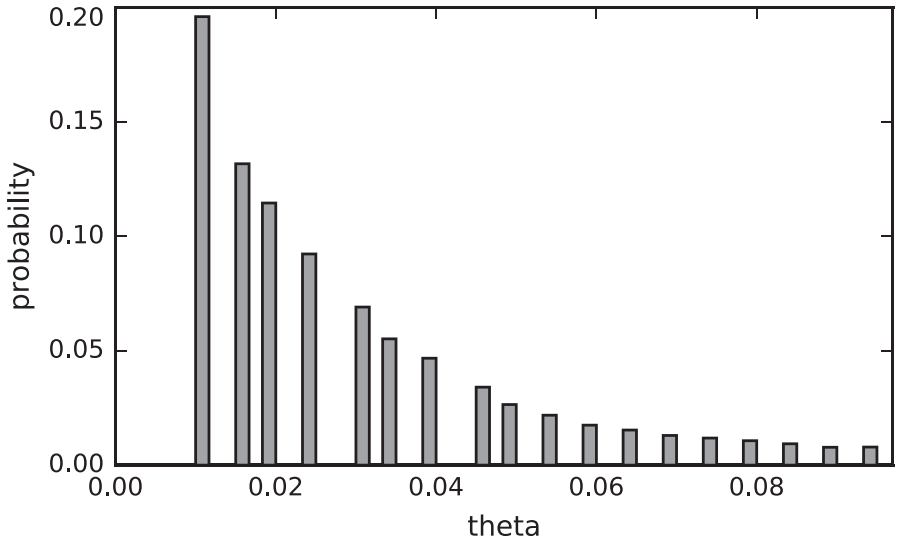

Figure 5.4 – Histogram of the distribution of θ of local atoms for the French female speakers.

tracker implemented in Kaldi by Ghahremani et al. [2014].

Our hypothesis is that, because of the nature of the matching pursuit algorithm on which WCAD is built, our algorithm will progressively increase the WCORR with the addition of each of the atoms, reaching a saturation point at the optimal number of atoms. We also hypothesise that relatively few atoms will be needed to construct a model perceptually close to the F_0 contour.

5.5.4 Comparison with Command-Response Model

In addition, we assess the comparative performance of our algorithm with the results obtained with Mixdorff's CR parameter extraction tool [Mixdorff, 2000]. We calculate the $\text{WCORR}_{\text{norm}}$ obtained with the CR model, and use it to assess the perceptual quality of the modelled contour, comparing it with our WCAD results.

Our hypothesis is that WCAD results would be comparable with those obtained with the CR model at a comparable number of atoms per syllable, and that the GCR model can reach higher accuracies than the CR model, due to possibility to change the threshold when decomposing a contour.

5.5.5 Results

Example of Decomposition

Example results of the Weighted Correlation based Atom Decomposition algorithm are given in Figure 5.5 for the utterance *arctic_a0112.wav* taken from speaker *bdl* in the CMU Arctic

dataset. The plots show the original F_0 contour, the extracted phrase atom and the extracted local atoms, and the reconstructed F_0. In order to obtain a clearer plot, only local atoms with amplitudes above 0.3 were used. As a comparison, the standard CR model extracted with Mixdorff's tool of the same example utterance is also given.

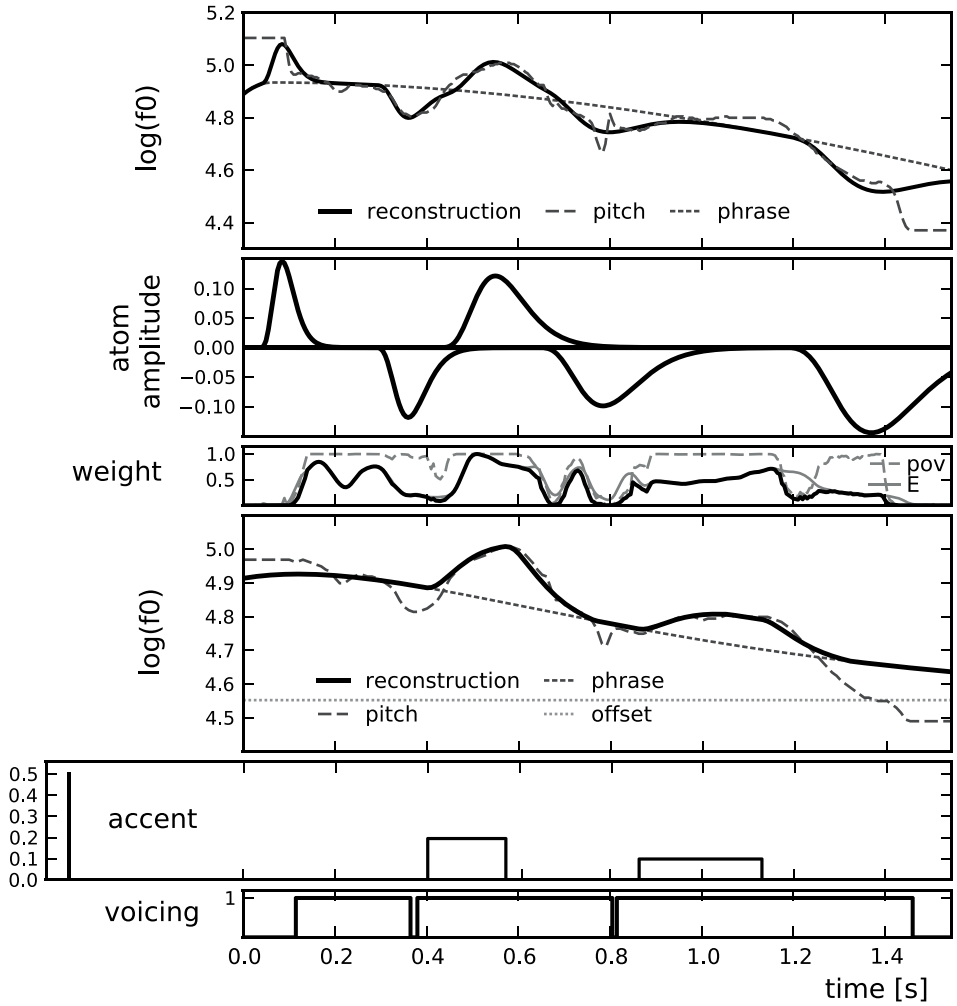

Figure 5.5 – Reconstruction of F_0 contour using the generalised CR model (1st panel), atoms extracted using the WCAD algorithm (2nd), and weighting function used (3rd); compared to the reconstruction using the standard CR model (4th), phrase and accent commands extracted using Mixdorff's tool (5th), and the voicing vector used (6th), for an utterance from *bdl*.

We can see from the plots that the WCAD algorithm, with the limit put on the atom amplitude, extracts 1 phrase atom and 5 local atoms to model relatively well the F_0. Mixdorff's tool extracts 1 phrase command and 3 accent commands to model the same utterance. The lower number of components is advantageous, but the standard CR model, however, fails to capture the phrase-final drop in F_0. Phrase-final drops were accounted for only later in the standard CR model, through the addition of negative phrase-final phrase commands [Fujisaki, 2004], and they are not automatically extracted by Mixdorff's tool.

On the other hand, the lack of negative accent commands for English in the CR model, precludes the proper placement of the phrase component. An example of this can be seen in Figure 5.6, in which the accent commands compensate for the wrongly placed phrase command. The WCAD algorithm, on the other hand, is not limited to using only positive local atoms, allowing it to do a better job at fitting the phrase atom, while at the same time being physiologically more plausible. For these two examples, we can see that the WCAD algorithm extracts a phrase component which looks like an average of F_0 movements. This smooth version of the F_0 curve makes sense physiologically as it would assure minimal activations, and thus conservation of energy. By contrast, the phrase component extracted by Mixdorff's algorithm is placed at the minimum of the F_0 curved. This can lead to incorrect accent commands: as one can see in Figure 5.6, to compensate for the offset between the extracted phrase component and what should be the actual phrase component, many erroneous step functions are used and stacked together to bring the contour up.

Results From Experiments

In the examples shown in Figures 5.5 and 5.6, we have limited WCAD to large amplitude atoms. The algorithm can, however, iteratively extract atoms to bring the modelled F_0 close to the original to an arbitrary degree, in terms of the cost function used. To analyse this performance we have calculated the $\text{WCORR}_{\text{norm}}$ at each iteration of the algorithm and plotted it as a point in the WCORR — atom/syllable plane, for all of the utterances for both speaker groups (male / female) from BREF. The results are shown in Figure 5.7 as grey dots. The figure also shows the average $\text{WCORR}_{\text{norm}}$ relative to the number of atoms/syllable, averaged across all the sentences for each speaker group, as a black curve.

The average $\text{WCORR}_{\text{norm}}$ plots obtained for the different speaker groups from the multilingual set are plotted for comparison in Figure 5.8. The curves represent the average performance of the GCR with $k = 6$ per speaker group, while the dots represent the performance of Mixdorff's tool at the sentence level for the same speakers and sentences. In Mixdorff's case, each sentence is represented by a dot as it only gives one decomposition result. In the GCR case, according to the number of local component we extract, we get different WCORR, hence the average curves. To calculate the WCORR for the standard model we only used the part of the F_0 contour that was between the start and end of voicing.

We can see that, as hypothesised, at the start the WCAD algorithm gives rapid improvements in the $\text{WCORR}_{\text{norm}}$ with the inclusion of the first (larger) atoms in the model. The improvement in WCORR then gradually decreases as more (smaller) atoms are introduced. The plots show that the improvements in WCORR reach a saturation point around 1 atom/syllable for all of the speaker groups of all databases, hinting at a deeper link between the syllable unit and elementary intonation atoms.

The results show that the WCAD algorithm performs equally well for speakers of different languages and gender. The female speakers show a slightly lower performance of the GCR

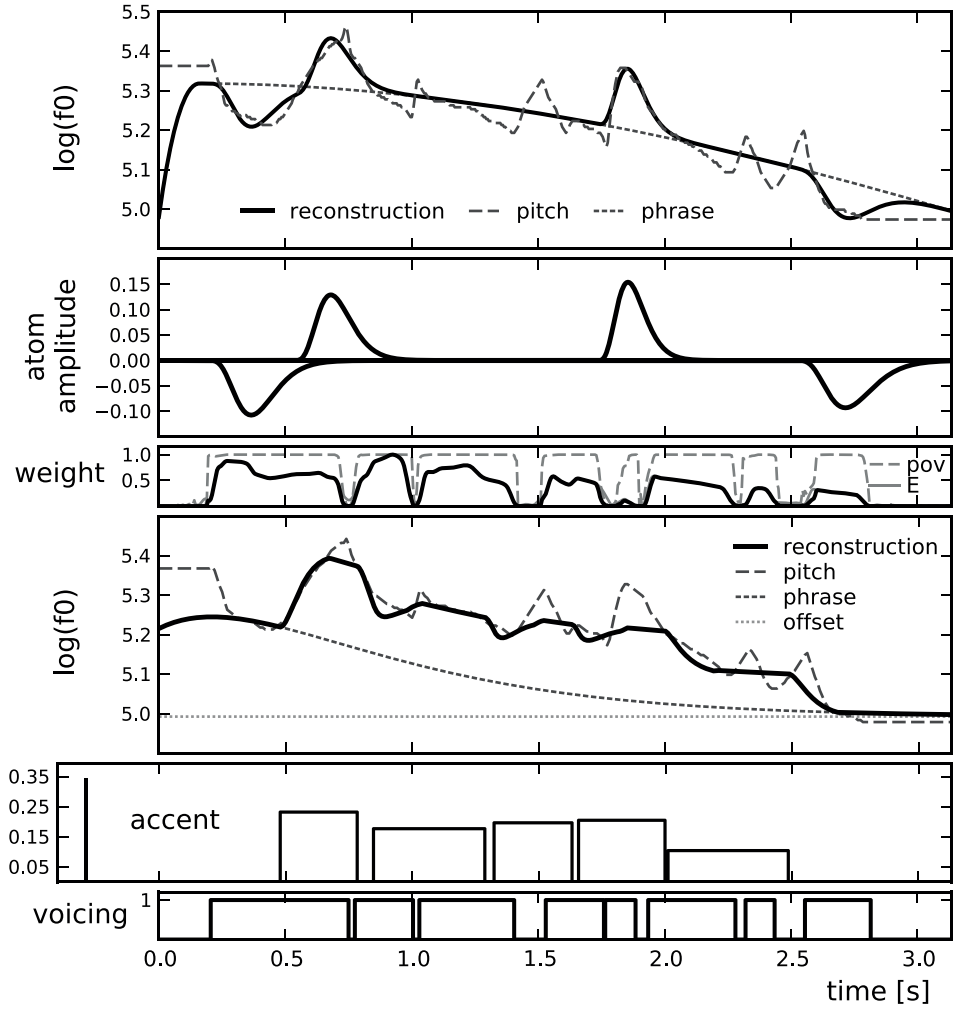

Figure 5.6 – Reconstruction of F_0 contour using the generalised CR model (1st plot), atoms extracted using the WCAD algorithm (2nd), and weighting function used (3rd); compared to the reconstruction using the standard CR model (4th), phrase and accent commands extracted using Mixdorff's tool (5th), and the voicing vector used (6th), for an utterance from *clb*.

model, as they often have more variations in their intonation, requiring more components to get the same precision. The hypothesis that speaker and language play a role in the complexity of the patterns comes naturally, however, the WCAD algorithm does not have an inconsistent behaviour across all of the data used, suggesting both speaker and language independence.

Compared to the standard CR model, the WCAD algorithm underperforms when using a smaller number of atoms in some cases (first dataset), but its accuracy reaches and goes beyond that of the CR model as more atoms are added. It is important to note that in the case of the plotted dots obtained from the CR model, "atoms/syllable" actually represents "commands/syllable", and that the commands in the CR model actually represent a response to a sequence of pulse excitations, as discussed in Section 5.3.2. On the other hand, the atoms in our generalised CR model correspond to single pulsed excitations, making straightforward

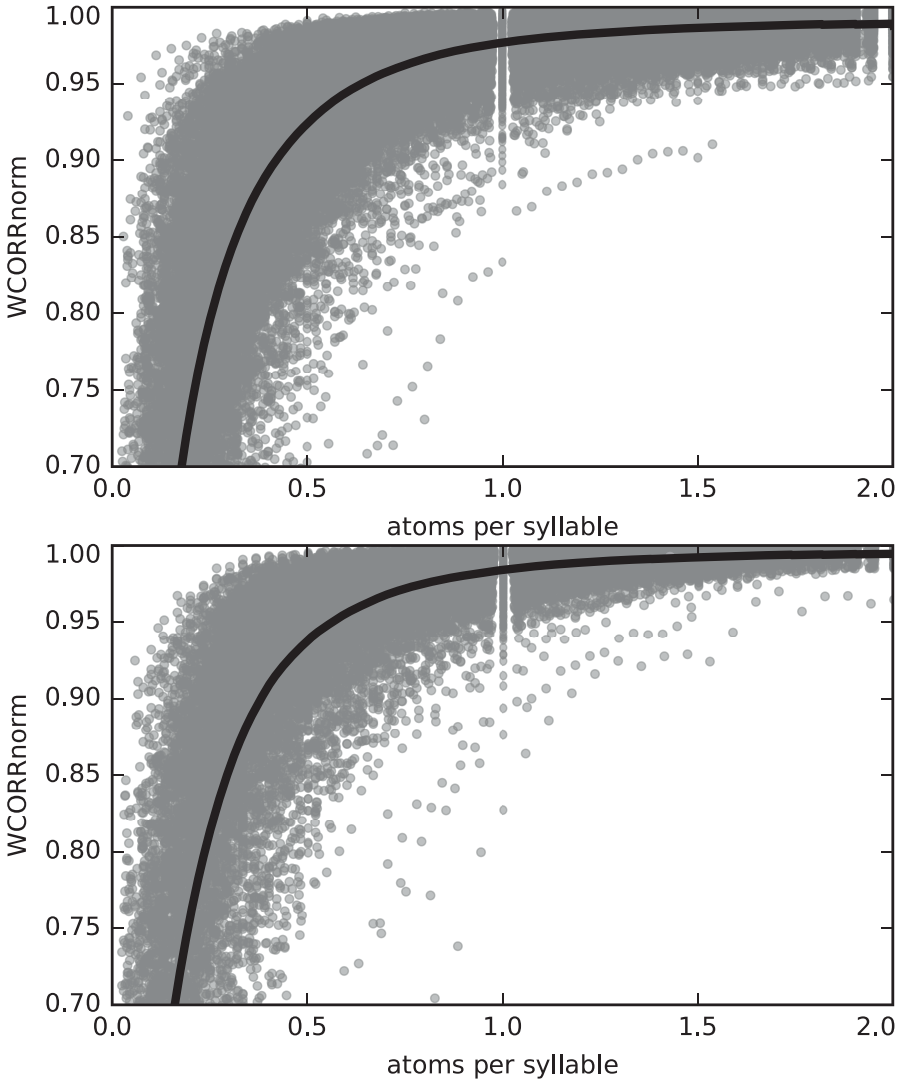

Figure 5.7 – Weighted correlation of the zero-mean normalised F_0 contours relative to the number of atoms per syllable for all of the utterances for the French female (top plot) and male (bottom) speakers (gray points), and the calculated average curve (black), for $k = 6$.

comparison on this plot slightly biased.

In order to get a sense of the number of atoms/syllable needed for the WCAD algorithm to reach a certain perceptual accuracy in modelling the F_0 contour, we used the different WCORR perceptual thresholds presented in Table 5.1 as stopping criteria. The results of this analysis are given in Table 5.3. The table lists the average number of atoms/syllable needed to reach the different perceptual WCORR thresholds, for each of the speakers or speaker groups. We can see that to reach perceptual indistinguishability (Category 1) WCAD uses on average 1 atom per syllable for the first dataset, as was also hinted by the WCORR plots in Figure 5.7. In the second dataset case, fewer atoms are needed to reach such perceptual quality. If we relax this accuracy condition and go with an F_0 model that permits some perceptual difference

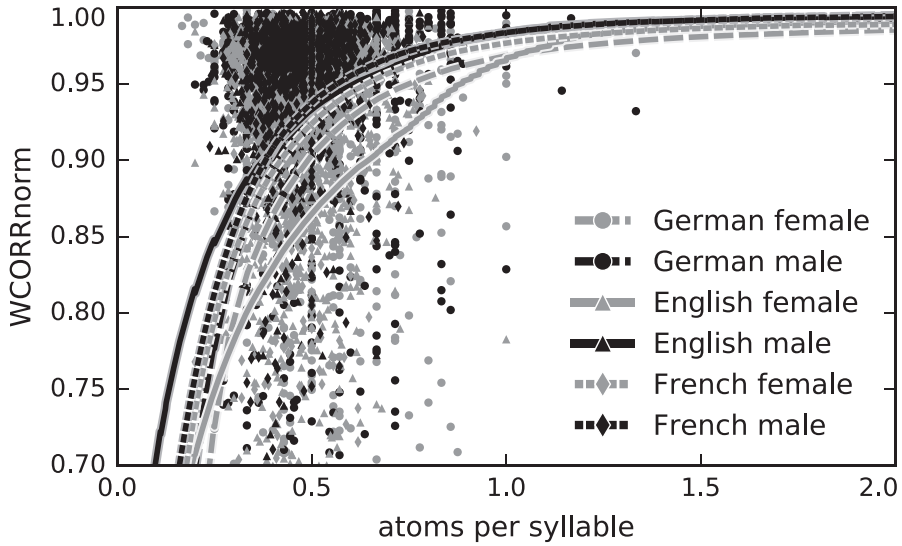

Figure 5.8 – Average weighted correlation of the zero-mean normalised F_0 contours relative to the number of atoms per syllable for the different speaker categories from the multilingual set, for $k = 6$ (curves). The WCORRs obtained with Mixdorff's implementation of the CR model are shown for comparison (dots, triangles and diamonds).

(Category 2), the generalised CR model needs on average a little more than half of this atom rate, i.e. 1 atom for every 2 syllables.

Table 5.3 – Number of atoms/syllable needed on average to reach a chosen perceptual WCORR threshold with the GCR model, for each speaker group from both datasets.

Speaker Group	Cat 1	Cat 2	Cat 3	Cat 4
Arctic *bdl*	0.75	0.48	0.34	0.24
Arctic *clb*	1.27	0.74	0.45	0.29
Group En M	0.69	0.53	0.34	0.26
Group En F	0.90	0.62	0.47	0.34
Group Fr M	0.71	0.47	0.32	0.22
Group Fr F	0.93	0.70	0.54	0.41
Group Ge M	0.81	0.52	0.41	0.29
Group Ge F	0.79	0.60	0.41	0.33
Arctic Average	1.01	0.61	0.39	0.26
Group Average	0.78	0.57	0.43	0.32

As a comparison to the performance obtained with the CR model, Table 5.4 gives the average WCORR, and the average total number of phrase and accent commands in the standard CR model for each speaker group. We can see that Mixdorff's tool gives a model with a WCORR of 0.96 on average for the first dataset, which corresponds to Category 2 from Table 5.1, and of 0.91 for the second dataset, which corresponds to Category 3. The average number of commands/syllable is 0.49 for the first dataset and 0.48 for the second. This number is to be compared with the results obtained with the WCAD algorithm at 0.61 for Category 2 (first

dataset), and 0.43 atoms/syllable for Category 3 (second dataset). In the first dataset case, a few more atoms are required for the GCR model compared to the standard CR model for reaching the same perceptual quality, while for the second dataset which has more variability, the GCR requires fewer atoms than the standard CR model. This affirms the comparable performance of our algorithm.

Table 5.4 – Average WCORR and number of atoms/syllable obtained by the standard CR model, for each speaker group

Speaker Group	WCORR	Cat	commands	com/syl
Arctic *bdl*	0.96	2	5.7	0.48
Arctic *clb*	0.96	2	6.1	0.51
Group En M	0.94	3	12	0.46
Group En F	0.91	3	14	0.47
Group Fr M	0.95	2	12	0.46
Group Fr F	0.94	3	12	0.48
Group Ge M	0.91	3	5	0.51
Group Ge F	0.83	4	5	0.50
Arctic Average	0.96	2	5.9	0.49
Group Average	0.91	3	10	0.48

Discussion

The example figures (5.5, 5.6) demonstrate the qualitative advantages of the more flexible GCR model over the standard CR model. The native allowance of negative atoms in the GCR model, as well as the design of the phrase atoms and the algorithm used to extract them, have allowed the extraction of an observably better phrase component. These two advantages result in better, physiologically more plausible modelling results overall.

The experiments confirmed the plausibility of the GCR model, and the WCAD algorithm as a means for the extraction of its parameters. The results show that the model can successfully capture the intonation dynamics for different speakers and languages to an arbitrary precision. The built-in WCORR measurement allows the user to set the perceptual quality of the modelled intonation patterns. The results show that high perceptual quality can be obtained with the model when using around 1 atom per syllable.

The results from the comparison showed that the WCAD algorithm gives comparable modelling performance to the standard CR model in terms of perceptual quality at a given atom/-syllable rate. It also accentuates the added flexibility of WCAD due to its iterative nature, which allows for an arbitrary modelling precision to be achieved. Namely, the results demonstrated that as more and more atoms are being added, the WCAD algorithm reaches WCORR-s that cannot be reached by the standard CR model. This is an inherent advantage of the introduced intonation model.

An additional point that we need to emphasise when we compare the GCR to the CR model

is that the parameters of the GCR model can be extracted fully automatically using the proposed WCAD algorithm. On the other hand, there is no automatic way to extract the "right" parameters for the CR model. Even advanced tools such as Mixdorff's that we used, are prone to erroneous output and need expert adjustment.

5.6 Conclusion

In this chapter, we presented a new intonation model that we call the generalised command-response model which decomposes F_0 into physiologically meaningful components. As its name indicates, it is a generalisation of the command-response model, in the sense that the constraints on the components are relaxed (we use higher order than the standard CR model). The GCR comes with a simple decomposition method based on the matching pursuit algorithm: the process will extract components in an iterative manner until it reaches a predefined stopping criterion. Compared to the CR model, different shapes are used for both global and local components, remaining damped system impulse responses. A modified version of the matching pursuit was proposed: instead of using correlation to find an atom at each iteration, a weighted correlation with a perceptual relevance is used as cost function, and allows use of a perceptually relevant stopping criterion.

The model was evaluated on speech from multiple languages and multiple speakers, and compared against an implementation of the standard CR model. It proved able to reach high accuracies with a reasonable number of components. Interestingly, it was found that about one atom per syllable was enough to model the F_0 curve reasonably well. Overall, we demonstrated that the model fits the curve as well or better than the CR model, with a similar number of components.

However, achieving a high reconstruction accuracy with the WCAD algorithm introduces atoms which model the noise inherent to the intonation curve. It is a difficult problem to automatically separate the prosodically meaningful events from the microprosodic noise with no linguistic and paralinguistic information. The following two chapters investigate applications of the GCR model: the possibility to learn the model parameters, in the context of synthesising intonation for TTS (Chapter 6); and the use of the model to transfer emphasis (Chapter 7). Finding linguistic meaning of the atoms has received some attention in the work of some project partners: Delić et al. [2016] found a high correlation between high and low tonal events in the ToBI system and positive and negative atoms respectively. Szaszák et al. [2016] also used atom decomposition for emphasis detection. In the same line of work, mutual information between linguistic labels and model parameters is investigated in Chapter 7, in the context of intonation-based emphasis transfer.

6 Intonation Synthesis

Generating prosody is a problem that is inherent to text-to-speech (TTS) synthesis. As the current TTS systems are able to synthesise reasonable quality speech, one of the criticisms which tends to come regularly is the lack of expressivity in the output speech. Prosody is the main vector for allowing expressivity and the standard TTS methods show their limits when it comes to model it.

We propose to use the GCR model (Chapter 5) to predict intonation. The model is physiologically based and aims at modelling intonation in a language independent manner. We assess the importance of the model parameters by human subjective evaluation. Then, using different statistical methods, we attempt to generate the model parameters, which could then be used for synthesising intonation. In this chapter, only local component prediction is investigated. The task of intonation prediction is evaluated on a large single speaker English speech corpus.

The work presented in this chapter was done in the context of an internship at the National Institute of Informatics (NII), Tokyo, Japan. Some research decisions were based on taking advantage of particular tools available at NII.

The work has not been published, mainly because the results are somewhat negative[1]. This in turn leads to the focus on transfer in the later chapters.

6.1 Background

Two types of approach are discussed in this section: intonation synthesis integrated in the TTS framework, and intonation synthesis in an external fashion.

[1] No reflection on NII

6.1.1 Integrated Modelling

The state of the art of intonation synthesis in the context of TTS was introduced in Chapter 2, Section 2.2.3. This is in the case of statistical parametric speech synthesis (SPSS). The two main approaches to the SPSS problem are HMMs and DNNs (and their variants). In both cases, intonation is treated like another acoustic parameter, such as spectral information. A relation is learned between linguistic labels and F_0 values (or a scaled version of it).

In the case of HMM, F_0 is modelled at the state level. This means that the emission probability density of each state will contain mean and variance (if normal distributions are used, as is generally the case) of F_0 for a specific sub-unit of a phone in a specific context. Decision trees handle the task of clustering the states which have similar distributions, and supra-segmental information is more likely to have an effect on F_0 than segmental details, e.g. the position of a word in the sentence, combined with the number of words in that sentence would probably bring more information about intonation than if the current phone is an “a” or an “o”. The voiced / unvoiced decision is traditionally handled by the use of multi-space probability distribution HMMs [Tokuda et al., 2002a]. More details are given in Chapter 2, Section 2.2.3. It is also possible to use continuous F_0, or a different continuous decomposition of it, and use different methods to model the frame voicing decision, e.g. [Ribeiro and Clark, 2015; Suni et al., 2013; Yu and Young, 2011].

In DNN-based speech synthesis, F_0 is modelled in the same way as other acoustic features. Each output frame is generated according to an input feature vector (linguistic context). In addition to the features which are used in HMM-based synthesis, some information about relative position within the phones is given, to model the evolution of the features more finely. A parallel can be made with the states of an HMM: a phone is not modelled with only 1 state but generally with 5 emitting states, which allow the modelling of articulation with neighbouring phonemes.

Recently, some work was done to try to separate the suprasegmental aspect from the segmental aspect in the training of deep networks. Ribeiro et al. [2016b] investigated how training separate networks and then combining their synthetic output was affecting the synthetic speech. It was found that a parallel structure was improving upon the baseline architecture.

6.1.2 External Modelling

Another type of approach to synthesise intonation consists of building models which are specifically designed for the task of intonation generation. It is known that synthetic speech is generally perceived as “over-neutral”, attributed to a “flat” intonation. One of the reasons for the flat intonation is the use of statistical models, which are learning averages (and ranges) of F_0 in specific contexts. Relevant models to this chapter are discussed in this section.

Bailly and Gorisch [2006] proposed applying the superposition of functional contours (SFC) of Bailly and Holm [2005] to German intonation generation. In this work, the authors manually

labelled a German speech corpus containing only declarative sentences for which phonetic and prosodic (borders and accents) segmentation was provided. Their annotation is on several levels: the first (highest) level is the sentence modality, in this case it is always *declarative sentence*. Then, the dependencies of each unit (where units can be words, groups, phrases, clauses) are considered, meaning that a unit can have a left dependency (linking the current unit with the preceding one), right dependency, interdependency (when the two units have the same "governor"), or independency (when none of the previous dependencies can be applied). As German is a compound language, some morphological decomposition was annotated (e.g. the word "Energiepolitisches" is parsed as "$[[\text{Energie}]_{\text{AM}}[\text{Politisches}]]$")[2]. Like English, German has lexical stress and the accents are annotated. Finally, emphasis was encoded at the word level. These annotations are used to train the prosodic models. This work was evaluated on 10 test sentences (with a relatively small training set of 70 sentences), and although the method resulted in mediocre objective and subjective measures, it is an interesting approach as the constraints on the different prosodic components are relatively relaxed.

Kameoka et al. [2015] proposed one of the most interesting approaches, in the context of intonation synthesis using a physiologically plausible model, the command-response model of Fujisaki and Nagashima [1969]. HMMs with a particular topology were employed to model the activations of global and local components. As a result, the system, constrained by some rules on the possible activation of components, can be in different states, corresponding to having an impulse for phrase command, no command, or an active command for accent (step function). It was extended by the use of substate HMMs, allowing modelling of the duration, related to the time spent in each state. By translating the CR model into a probabilistic model, Kameoka et al. [2015] were able to successfully extract model parameters and to generate them in the context of TTS for Japanese. As the CR model was originally developed by building on Japanese intonation theories, the link between its component and linguistic events is well established. This is one of the reason for the success of this method, which focuses only on declarative sentences as well.

6.2 Relation Between GCR Parameters and Perception

Before tackling the task of synthesising F_0 using the GCR model, a perceptive evaluation was carried out to measure the importance of the position of local atoms. Our hypothesis is that modifying slightly the position of local atoms should not bring perceivable differences in the speech signal, but that if the shift becomes important, the speech will be distorted.

6.2.1 Generating Test Material

To assess the importance of the position of the main impulses in our model, we attempt to modify the time at which the most important atoms — in a perceptually relevant way — occur.

[2] In English, "energy policy".

Here are the steps to generate our test files:

1. Extract parameters using WCAD algorithm (parameters are position, amplitude and θ)

2. Shift the most prominent atom by n frames, $n \in \{-5,-3,-2,-1,0,1,2,3,5\}$ where the frame shift is 5 milliseconds.

3. Reconstruct intonation using modified parameter atoms

4. Resynthesise speech using altered intonation.

The test data consisted of 20 sentences randomly selected from the Blizzard challenge 2011 corpus, provided by Wilhelms-Tricarico et al. [2011], described in Chapter 2, Section 2.4. As a reminder, this data consists of US English speech uttered by a female voice talent. The prompts were annotated for the speaker to read with target intonation patterns, making the data somewhat expressive.

For the atom extraction, we used two stopping criteria: a maximum number of atoms, related to the length of the utterance (there could be maximum 10 atoms per second), and a limit on the magnitude of the atoms (the atoms had to have an absolute amplitude greater than 0.3).

9 systems were then tested, with different shifts in the position of the most important atom (this atom could have a positive or negative amplitude). These were compared against a simple analysis-synthesis version of the same audio files.

6.2.2 Listening Tests

31 native English listeners, mostly in the age range 18–25 (students at the University of Edinburgh), took the subjective listening test in separate sound proof booths, with high quality headphones. The subjects were asked if pairs of samples sounded exactly the same or not. Each pair was composed of a vocoded version of the original file and a reconstructed version after modifying the position of the most prominent atom (or simple reconstruction after decomposition using our model in the case of the system $S0$).

The listeners had to judge 180 pairs of audio files: 20 reference files against each of the 9 versions described earlier.

6.2.3 Results

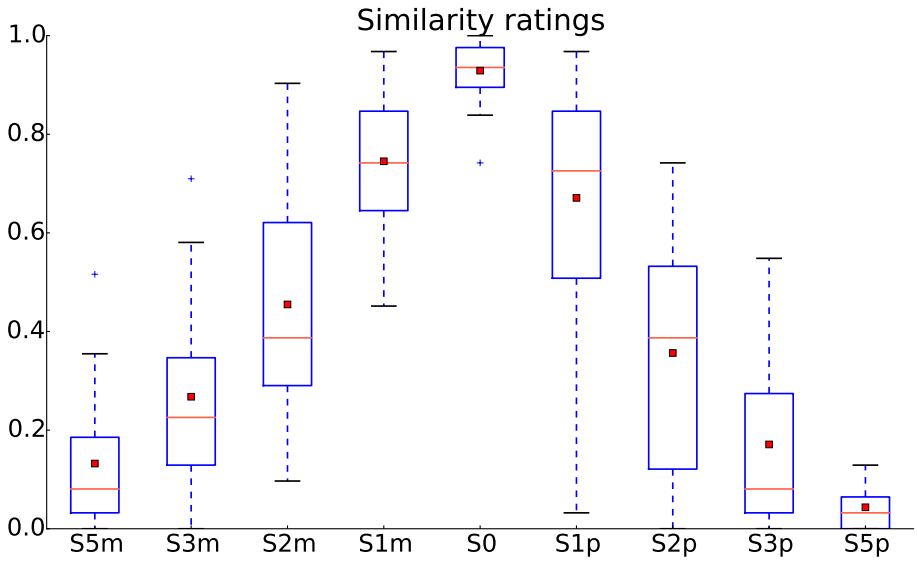

Figure 6.1 – Subjective listening test on the effect of atom position. 1 for identical, 0 for different. Red lines are medians, red dots are means.

The results of the listening tests are shown in figure 6.1. The value 1 corresponds to the cases where listeners perceived the speech as exactly the same as the original, 0 corresponds to different (with no information on how big the difference was). In the boxplot, the values are averaged at the sentence level, to account for the random choice of sentence. It can be seen that for a shift greater than 1 frame in the position of the most prominent atom, the perception of the reconstructed speech is different, due to some small distortion in the speech. The fact that we modified the position of the most important atom makes the modification have a bigger impact than if another random atom was shifted. A two-tailed paired t-test on the score of each sentence in the test set (average of scores given by the listeners) was performed on each pair of systems. This showed that all pairs of systems were significantly different at the level of $p < 0.01$, except the pairs (S3m / S3p), (S2m / S2p) and (S1m / S1p). This means that, from a perceptual point of view, shifting the most important atom, even by a small amount, is modifying significantly the perception of native listeners.

Our hypothesis, that small shifts in the position would not be perceivable by the listeners, was refuted, as even the smallest shift were perceived by some portion of the listeners. The intuition that the bigger the shift, the more perceivable difference was confirmed, which indicates higher degradation of the speech when the position of major atoms is altered. Ideally, when trying to predict atoms, the precision in the position should be a major concern; however, a shift of 1 frame seems acceptable to keep perceptual similarity as the means indicate that a majority of listeners do not distinguish it from the natural speech. The differences underlined in these results are when the most prominent atoms are modified, but it is expected and reasonable to assume that on other atoms, a small shift will be tolerable.

6.3 Synthesising GCR Parameters

Our motivation is to be able to predict intonation from linguistic context. Instead of following the frame by frame prediction used in state of the art system, we want to use our intonation model as a representation of the intonation. Predicting intonation then means predicting the model parameters. The reconstruction of intonation given the model parameters is straightforward. In the work presented here, only local component prediction is investigated. Our hypothesis is that statistical models can be trained to learn the relation between linguistic features and the model parameters, therefore enabling the synthesis of intonation contours.

Two statistical modelling approaches are investigated in this chapter: support vector machines (SVM) and deep neural networks (DNN). The task of the models is to output the parameters of our intonation model given a certain linguistic context as input.

6.3.1 Support Vector Machines

Support vector machines are classifiers. In that sense, they are generally not used to predict parameters, but rather separate distinct classes. To fit our task to SVMs, we therefore turn it into a classification problem: the classifier should output a binary answer to the question "Does this frame correspond to an atom?". In other words, the system would predict the position of *positive* frames, where we define a positive frame as a frame with an atom.

SVM Generalities

First introduced by Boser et al. [1992], SVMs are binary classifiers, whose goal is to separate two classes in some feature space by a hyperplane. The distance between the two classes should be maximised, and support vectors are the data points from each class which are the closest to the hyperplane. In most real case scenarios, a linear separation in the space where the data lies is not possible. Two methods allow reducing the problem: the introduction of a soft margin, and the projection of the data into a different space. The former, introduced by Cortes and Vapnik [1995], consists of allowing the classifier to ignore some data points, or place them on the wrong side of the margin, which in the best case can disregard wrongly labelled data. The latter, which is a projection into a higher dimension space, is performed to try to separate better the classes, as in the example depicted in Figure 6.2, where the data is projected from input space (in 2 dimensions) to the *feature space* (in 3 dimensions). For this, a non-linear kernel function is typically used.

SVM Specifications

To fit our task, we choose to predict the presence or absence of an atom at the frame level. The presence of an atom as we define it here consists of the presence of an impulse which triggers

[3] Source: http://www.imtech.res.in/raghava/rbpred/svm.jpg

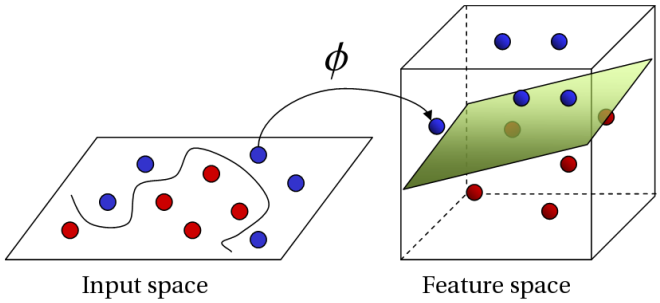

Figure 6.2 – SVM: projection in higher dimension space.[3]

a response — an atom. To this end, the input features will consist of a linguistic feature vector for each frame. To be consistent with the speech synthesis framework, we use almost the same linguistic features as in the case of a standard TTS system. In addition to the features commonly used[4], some information about the relative position of the frame in the phoneme, in the utterance and on the length of the utterance. As we are interested mostly in high level linguistic features, we reduce the feature set by removing segmental information such as the phone identity which, because it is encoded in a one hot manner, requires many dimensions.

Balancing the Data

In our case, as can be seen from the atom decomposition results in Chapter 5, the impulses corresponding to the presence of an atom are sparse. It means that, from a frame-level modelling point of view, only few frames contain an atom command: in the data used in the experiments (see Section 6.4), the presence of atom in a frame occurs less than 2% of the time only. It is known that SVMs do not perform well with unbalanced data. To overcome the under-representation of the *positive* class, we propose to "add noise" on the atom position. To do so, the following procedure was used: if t is the position of a positive frame, the surrounding frames were added in the positive class by duplicating them as positive frames, and keeping their negative frame version. Then, for a positive frame t, we would have:

- frame $t-4$ is added once as a +ve sample and remains once as a -ve sample (1+1-)
- frame $t-3$ is added twice as a +ve sample and remains once as a -ve sample (2+1-)
- frame $t-2$ is added 3 times as a +ve sample and remains once as a -ve sample (3+1-)
- frame $t-1$ is added 4 times as a +ve sample and remains once as a -ve sample (4+1-)
- frame t is added 4 times as a +ve sample in addition to the existing +ve sample (5+)
- frame $t+1$ is added 4 times as a +ve sample and remains once as a -ve sample (4+1-)

[4] For instance in the HTS demo feature set, see http://hts.sp.nitech.ac.jp/

- frame $t+2$ is added 3 times as a +ve sample and remains once as a -ve sample (3+1-)
- frame $t+3$ is added twice as a +ve sample and remains once as a -ve sample (2+1-)
- frame $t+4$ is added once as a +ve sample and remains once as a -ve sample (1+1-)

The noise added on the position should increase the robustness of the learnt models, and be more adaptive to unseen data. This noise introduction may however reduce the precision of the model to identify which frames are *positive*. We did not add noise on other contextual features as it would be difficult to modify them and remain consistent (for instance, it is not easy to modify randomly the position of the syllable in the word and keep the other features consistent with this new position). In our case, the features are still consistent — as accurately as the automatic text analyser and the automatic time alignment allow — and only nearby frames are duplicated as positive examples. This addition of noise on the position may seem contradictory with the findings of Section 6.2, where it was shown that position had to be very precise. However, if predicting atoms in a region around the real position is possible, with more importance on this position, we expect that the output of the models can be processed to retrieve the wanted position.

6.3.2 Deep Neural Networks

Deep neural networks have become a standard data-driven approach in many research fields. They have shown a great ability to improve the state of the art in speech applications such as ASR with e.g. the work of Veselỳ et al. [2013], TTS starting with the work of Zen et al. [2013], in various tasks of natural language processing [Collobert and Weston, 2008], or in image processing for classification, e.g. Krizhevsky et al. [2012]. Following the trends in the speech synthesis community, we investigate the use of DNNs in the prediction of our model parameters.

DNN Generalities

DNNs are introduced in Chapter 2 as an inherent part of statistical parametric speech synthesis research. A deep neural network is essentially a multiple layer network composed of hidden units, themselves composed of nodes (or neurons). Depending on the architecture of the network, the nodes are connected as desired. In a simple feed-forward architecture, neurons are connected from one layer to the next one, and the outputs of one layer become the inputs of the next layer. Then for each node in a layer, a weight is applied on each output of the previous layer, and the sum of the weighted values of each node connected to the current node with a bias term is passed through some activation function. The outputs will then be used as inputs for the next layers or as posterior features, in the last (output) layer.

Synthesising All Parameters

As we are working in the framework of DNN-based TTS, we decided to adopt a frame-by-frame strategy. One possible way to present the task of atom parameter generation to the DNN is to make it learn the relation between linguistic features and output parameters at the frame level. In the input it means using the same features as in the SVM case. In the output, as the frame indicates the timing (or position), we can generate parameters, and only the output frames which have a value for amplitude and θ should be considered, the others indicating the absence of an atom. For this setting, we used one of the standard activation functions, the *sigmoid* function.

An additional feature was investigated for atom parameter prediction: the phrase component of the model. The values of the phrase contour for each frame were normalised and then used as input features of the DNN. This was done following the assumption that the phrase can be predicted, and that we are investigating local component synthesis only.

Synthesising Parameters Separately

As we anticipate that it may be difficult to generate all the parameters at once, we investigate the prediction of atom parameters in a separate manner:

- Predicting atom position: As a first task, we train a network to predict only the presence or absence of an atom. This is the same task as for the SVM-based approach. With this knowledge, we can than predict other parameters. As we expect a binary output, a *softmax* layer was used as output layer, because of its ability to classify distinct classes, here presence or absence of atom. The *softmax* activation function is discussed in Section 6.4.

- Synthesis of amplitude and θ: training a network for *positive* frames only, meaning that in that case, we assume the presence of an atom. The prediction will then focus on the amplitude and θ of atoms.

The combination of positions and other parameters should allow the reconstruction of all the local components. In that case, we do not change the input features compared to the "synthesising all at once" approach.

6.4 Experiments

In this section, we present experimental evaluation of the proposed methods, namely predicting atom positions with SVM, predicting all parameters with one DNN, and predicting position and other parameters separately using different DNNs.

6.4.1 Experimental Setup and Evaluation Method

Data

As for the evaluation of atom position in the listening test in Section 6.2, we used the Blizzard Challenge 2011 data. The corpus contains 16.6 hours of speech, for about 12000 sentences.

Training / Testing Sets For the DNN training, we used 9,000 sentences. For testing the models, 1000 held-out sentences were used. In the SVM case, due to the high computational cost of training, we used about 40 minutes of speech, or about 500 sentences corresponding to 500,000 frames. The test set in that case consisted of about 72,000 frames (80 sentences, approximately 10 minutes of speech), and in the case of artificially augmented data about 100,000 frames). In the case of separate training, we used the same 9,000 sentences, which corresponded to about 160,000 frames.

Features

The features used for the different experiments are the same, simply varying in normalisation or size of the feature set.

Input Features There were two sets of linguistic features. The full feature set was based on the feature set defined by Takaki et al. [2015]. The original vector dimension was 897: 858 binary features for categorical linguistic contexts, 36 numerical features for numerical linguistic contexts, and 3 numerical features for the position of the current frame and duration of the current phoneme. Two other features were added: the length of the utterance and the relative position of the current frame in the utterance.

A reduced feature set was created by deleting some segmental information, as their influence on the intonation contour is small. As described earlier, features like the phone identity were left out of the reduced set as they require many dimensions for little long term dependency relevance. In that case, the final feature vector size was 169. In the DNN training, the normalised phrase component values were added as an input feature, to provide more context about the long term intonation component. The experimental results presented here were based on this reduced feature set, as little difference was observed between the full set and the reduced set in terms of system performance. Note that in the DNN case, we added the normalised phrase component values, then having a 170 dimension input feature vector.

Output Features In the case of SVM, the output consisted of a single value, which could be positive or negative according to the class of the frame. For the DNNs, in the case of *positive frame only* training and in the case where we trained all the features, there were 2 outputs: amplitude and θ. The training for position prediction had 2 outputs to model the 2 classes

presence and absence of atom.

The atom parameters were extracted using the weighted correlation based atom decomposition, with the same settings as for the listening test material generation: two stopping criteria were used, a maximum of 10 atoms per second and a minimum absolute amplitude of 0.3 for each atom.

In all the cases, the input features were normalised to have zero-mean and unit-variance. For the DNNs, output features were normalised to be within the range 0.1–1.0, while the SVM classes were labelled as 1 or -1, for presence or absence of atom, respectively, and consistently with the chosen tool.

Models

SVM In the SVM case, the Gaussian *radial basis function* (RBF) was used as kernel function:

$$k(\mathbf{x_i}, \mathbf{x_j}) = \exp(-\gamma \left|\mathbf{x_i} - \mathbf{x_j}\right|^2), \gamma > 0 \tag{6.1}$$

where $\mathbf{x_i}$ and $\mathbf{x_j}$ are feature vectors in the input space, and $\gamma = \frac{1}{2\sigma^2}$ where σ^2 is the variance.

A first grid search was performed on two parameters: the soft margin C, and the free parameter of the Gaussian radial basis function γ, within the ranges $C \in \{2^{-1}, 1, 2, 2^2, .., 2^{16}\}$ and $\gamma \in \{2^{-10}, 2^{-9}, .., 2^3\}$. A second grid search was conducted on a reduced range to try to refine the parameters: $C \in \{2048, 3072, 4096, 8192, 12288, 16384, 32768\}$ and $\gamma \in \{2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 1, 2\}$. Results are presented on the second set. For the experiments , the libsvm library[5], implementation of Fan et al. [2005], was used.

DNN for Modelling All Parameters Using the same framework as for DNN-based speech synthesis, we investigated various numbers and sizes of layers in this work: between 1 and 5 layers containing between 128 and 2048 units. The *sigmoid* function was used, with RMSE as minimisation criterion. DNNs were trained using back propagation. The implementation of the neural networks was the same as Takaki et al. [2015], and provided by the first author of this work.

Modelling Position of Atoms In this case, as we were interested in classifying the frames like in the SVM case, the *softmax* function was used in the output layer.

[5] Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/

The *softmax* function is defined as:

$$f(y_j) = \frac{e^{-y_j}}{\sum_{k=1}^{K} e^{-y_k}} \tag{6.2}$$

where y_j is the input of node i, and the $\{y_1, y_2, .., y_K\}$ are the inputs of the K nodes of the same layer. It is commonly used as an output layer in classification tasks. As can be seen from its formulation, when the input is high its value will be increased with respect to other nodes, the sum of all of them remaining equal to one. The cross entropy was used as a cost function. In the context of this work, only a smaller architecture was investigated: the network had two hidden layers containing 128 units and using *sigmoid* function, and one output layer using *softmax* function.

Modelling Positive Frame Parameters The architecture of the DNNs in this case was the same as in the *modelling all the parameters* case. The number of layers was between 1 and 4, and the number of units in each layer between 128 and 1024. The difference was in the training and testing sets, which only contained the positive frames. In the results, the architectures with 1024 units per layer are not presented, as they performed significantly worse than other systems and sometimes could not be trained successfully. This was probably due to a too small training set with respect to the number of parameters to be trained in the network. The case 4 layers with 512 units could not be trained either and is therefore not in the results.

Evaluation Metrics

In order to measure the accuracy of the proposed systems, several aspects are evaluated. In accordance with the fact that the position of prominent atoms is very important in the perception of the speech, the first factor that we are interested in is the position of the atoms. In the case where we try to predict amplitude and position only for *positive* frames, we want to evaluate how close these parameters are from the ground truth ones. Finally, we are interested in the reconstructed F_0 contour, which is the final output of the full intonation prediction system.

Evaluation of Position Prediction using The Gamma Factor One of the peculiarities of the model is that it consists of a series of discrete events, called “commands”, which are basically impulses — or spikes — that would then be passed through some filters. As underlined earlier, the position of the impulses – or position at which the atoms are triggered – is a key element in the prediction of the intonation contour. To evaluate the position of these atoms, a measure that takes into account the nature of such time sequences is needed. The gamma factor, or coincidence measure, is a measure that makes it possible to compare two spike sequences

and was introduced by Kistler et al. [1997]:

$$\Gamma = \frac{N_c - 2N_d\Delta\nu}{N_d + N_m}\frac{2}{1-2\nu} \tag{6.3}$$

where N_c is the number of coincident spikes, N_m the number of spikes generated by the model for this utterance, N_d the number of spikes in the data for this utterance, Δ the time interval to search for coincident atom (half window size) and ν is the spike generation frequency of the model (spikes/second generated by the model in average).

This measure should help us to compare the position of the generated commands with the original ones (extracted from the intonation contour with our decomposition method). The Δ parameter is the interval in which we estimate that the error is acceptable for localization, and the information from the listening tests should allow us to set it up accordingly. A value of 1 means that both spike sequences are identical, while a value of 0 means that the model spike sequence is random, a value lower than 0 then means that the model is performing worse than random.

Evaluation of Parameters for *Positive* Frames In the case of *positive* frames, for the separate training strategy, we measure the correlation and RMSE of the amplitudes of the atoms. This is done only on a few frames per sentence, but as the positive frames are in a chronological order, we measure if the network could learn the temporal evolution of the parameters, even though the input features are disconnected, as the frames are not actual neighbours.

Evaluation of F_0 Curves Finally, in the case where we have all the parameters, i.e. the DNN predicting all the features, and the DNN predicting only other parameters given the position, we can reconstruct the F_0 contour and compare it with the original. For this, we use standard measures: root mean square error (RMSE) and correlation. This will give an idea of how the end-to-end local atom prediction performs.

6.4.2 Results

We start by presenting some raw results from SVM and DNN systems, and follow with some post-processing of the system outputs. Overall results are then presented and discussed.

Example Results

Figure 6.3 shows a typical output of an SVM system tested on a sentence with standard frequency of atoms. Figure 6.4 shows the output of the same system for the same sentence, tested with the "noisy" version, where artificial positive frames are added around the real

positive frames. The second version of the sentence is more representative of the training data provided to the system. Figure 6.5 shows the raw output of a DNN for a test sentence (scaled to the actual test file), with the sequence of amplitudes. It differs from the SVM case as we are trying to predict amplitudes jointly with the atom position. Figure 6.6 shows an example of output from the neural network trained with a *softmax* output layer.

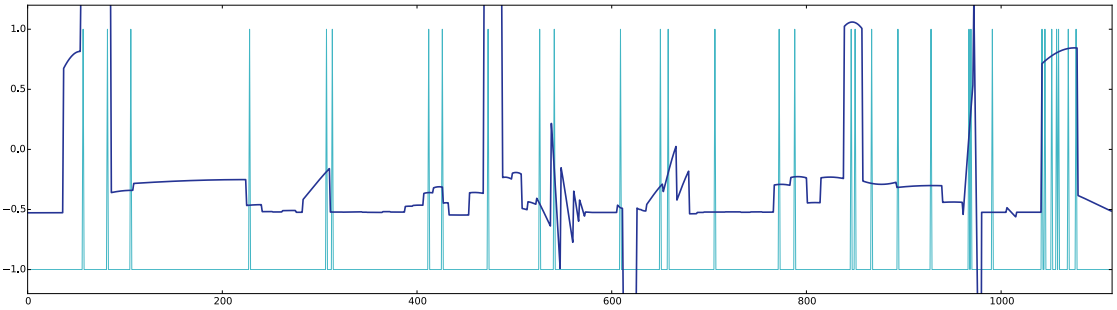

Figure 6.3 – SVM output example. The parameters are $C = 2^{15}$ and $\gamma = 2^{-5}$. Dark is synthetic, light is extracted impulse sequence.

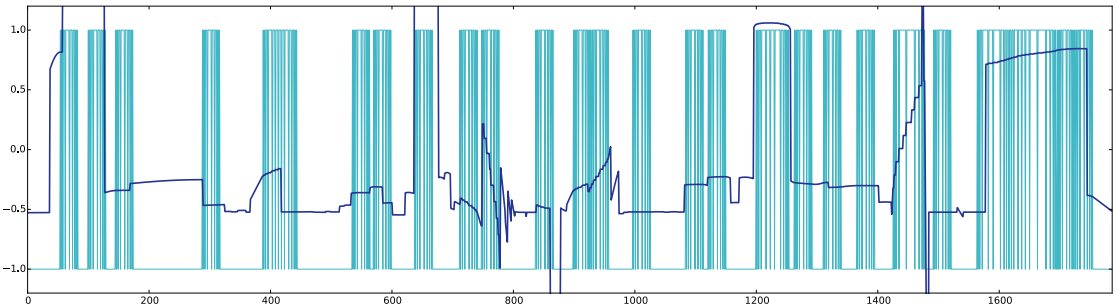

Figure 6.4 – SVM output example on noisy test file. The parameters are $C = 2^{15}$ and $\gamma = 2^{-5}$. Dark is synthetic, light is extracted impulse sequence.

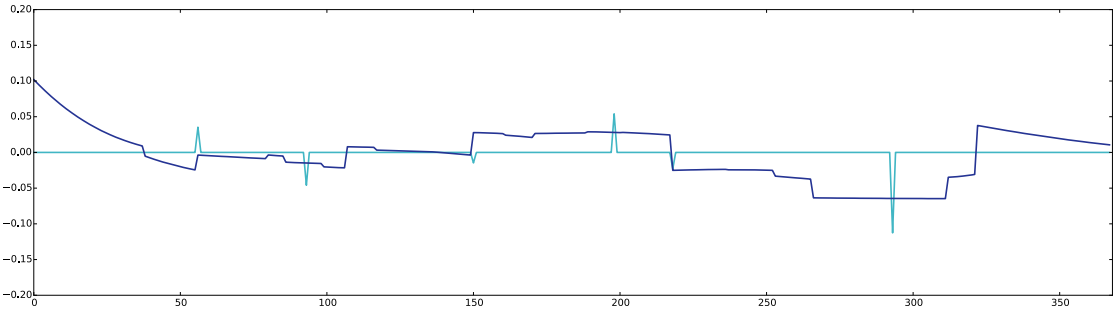

Figure 6.5 – DNN output example. The DNN has 5 layers and 256 units per layer. Dark is synthetic, light is extracted impulse sequence.

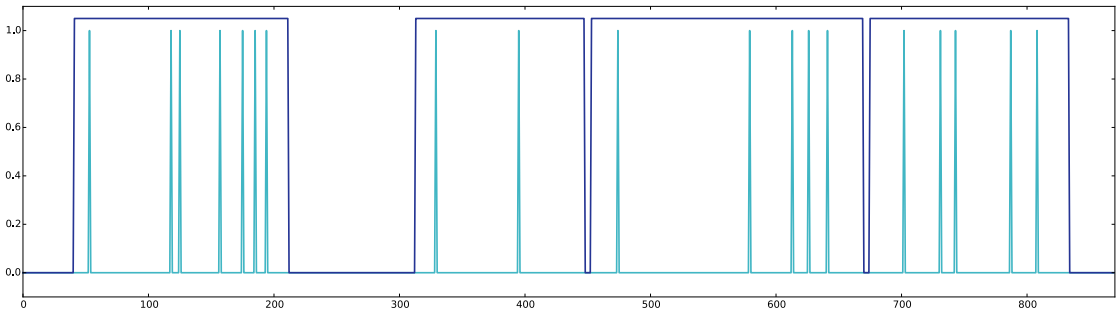

Figure 6.6 – Neural network with softmax layer output example. The curves are scaled for displaying purposes. The network had 2 layers of 128 units and a softmax output layer. Dark is synthetic, light is extracted impulse sequence.

One obvious observation from these plots is that the output in both SVM and DNN systems should be processed to be compared with the original spike sequences. In the case of the network with a *softmax* layer, no processing is needed as the decision is already made binary by the *softmax* layer. However, it seems that the network is not able to learn the position of the impulses, but rather regions were atoms are likely to exist. In this example, the regions where the function is not activated correspond to silences or phrase breaks. This means that the network was able to learn simple "probability of producing an atom" relation with the labels. Overall, this architecture does not seem suited to the precise prediction of atom position.

Need for Output Post-Processing

In the case of SVM, one simple way to transform the output of the system to a format which enables comparison with the original spike sequence is to put a threshold on the output stream, yielding a binary classification. Figure 6.7 shows the same sentence as Figure 6.3 and 6.4, after applying a threshold at 0 on the output, in both cases. For DNN, another approach was taken: the local maxima which are greater than the median and the local minima which are lower than the median are turned into impulses, conserving their value at that time step (the median value is assumed to be 0 in a non normalised vector). In the case where everything is modelled at once, this is first done on amplitude output values. Then the position of the impulses in the amplitude sequence is used as impulse positions for the θ values, this way impulses are aligned in the two output streams. Figure 6.8 shows the same sentence as in Figure 6.5, post-processed using this strategy.

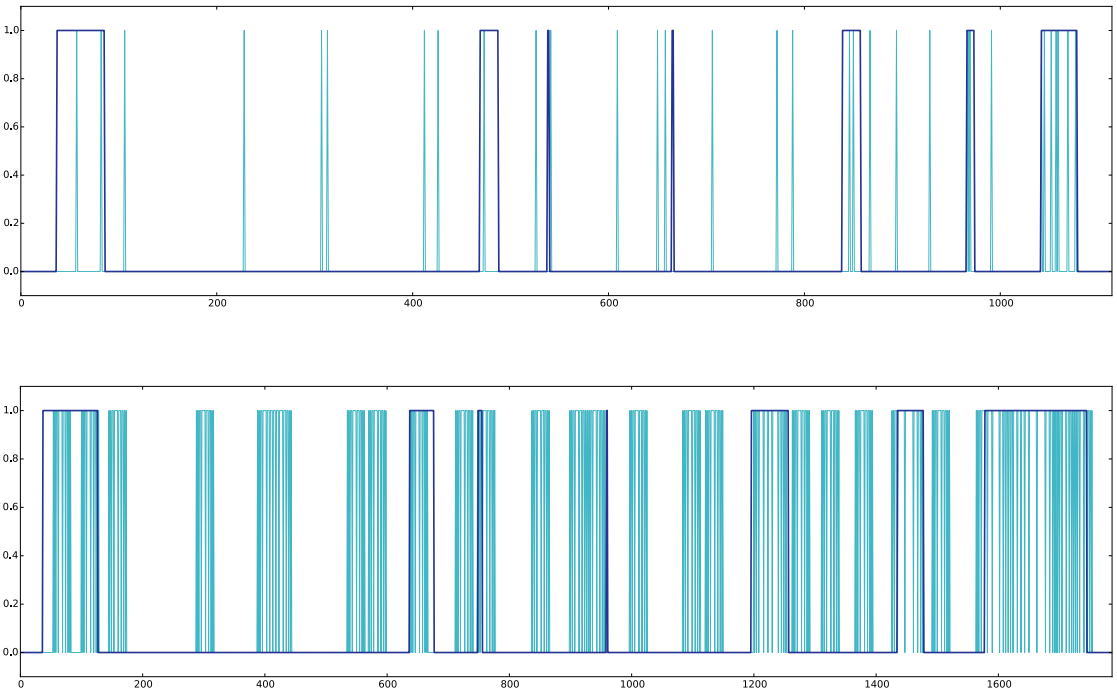

Figure 6.7 – Post-processed SVM output amplitude example. Top: original test file, bottom: test file with noise on position. The parameters are $C = 2^{15}$ and $\gamma = 2^{-5}$. Dark curves are synthetic, light are extracted impulse sequences (with added noise in second case).

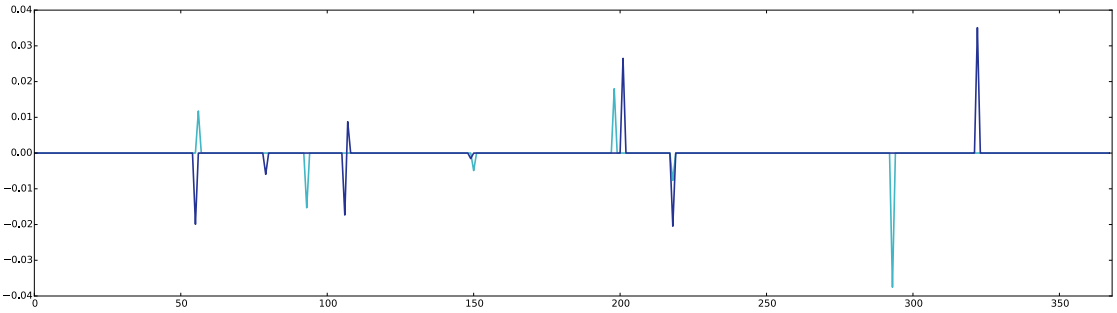

Figure 6.8 – Post-processed DNN output amplitude example. The DNN has 5 layers and 256 units per layer. Dark is synthetic, light is extracted impulse sequence.

Gamma Factor Analysis

Figure 6.9 presents the gamma factor means and variances obtained for the SVM systems, with 80 sentences as test data in the first case, and the same 80 sentences augmented with noise on the position in the second case. The mean performance and variance of each system are given. The Δ parameter was set to 0.005s, which corresponds to one frame shift, the coincident spikes can then be either one frame earlier or one frame later than the ground truth ones.

Figure 6.10 presents the mean and variance of the gamma factor obtained for the DNN systems,

calculated on 1000 test sentences.

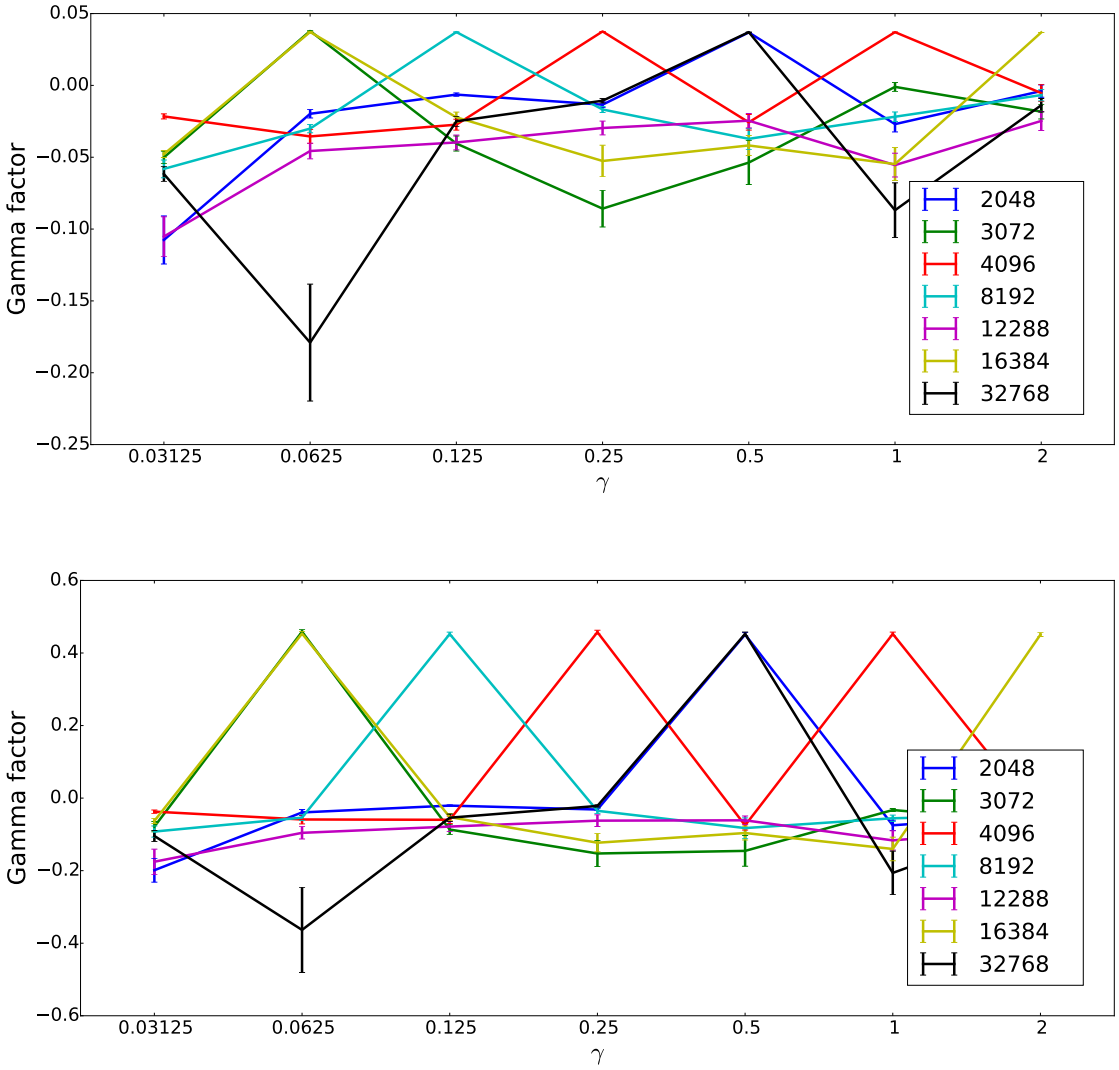

Figure 6.9 – Γ factor for SVM systems as a function of γ, for different C values. Top: on real test file, bottom: same test files with noise on position.

In the SVM case, we see that the results are slightly better when testing on data with noise on the position, which is expected as this data is more similar to the training set. This is insufficient to estimate precise positions. SVMs seem to be able to learn regions where atom commands are likely to fire, however the precision on the atom positions is not good enough. Overall, the inter-system variance is very small and located around chance level, with an average performance between $\Gamma = -0.18$ and $\Gamma = 0.04$ in the clean case. When testing on noisy data, the values span between $\Gamma = -0.36$ and $\Gamma = 0.46$. On clean data, the best performance is obtained with the system for $\gamma = 2^{-5}$ and $C = 3072$, with an average of $\Gamma = 0.04$. Consistently, the same system gives the best results with noisy data with an average of $\Gamma = 0.46$. However this performance is not significantly different from other systems. Furthermore, considering

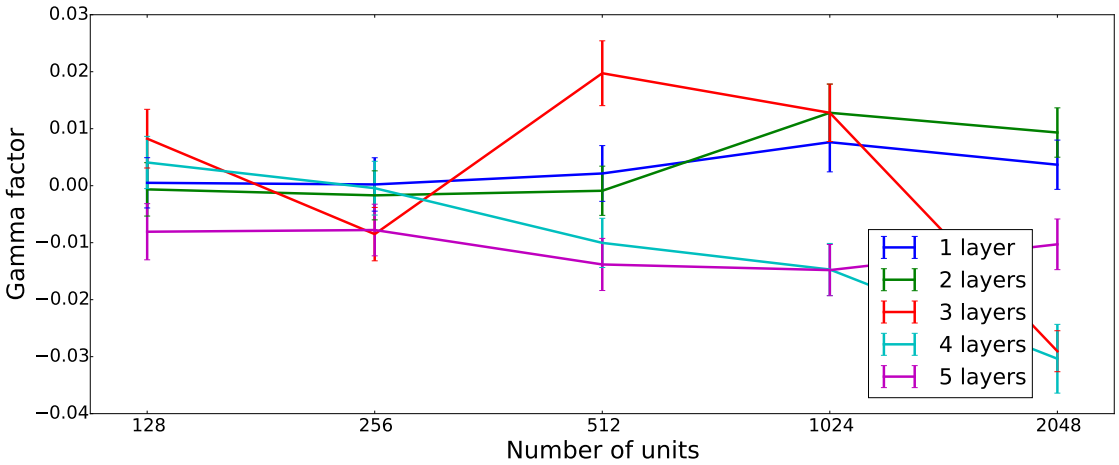

Figure 6.10 – Γ factor for DNN systems as a function of the number of units per layer, for different number of layers.

the shape of the data, the use of gamma factor for evaluation may not be relevant in that case.

For the DNN, the results are somehow similar to the SVM case, with a precision which is very low. The results are between $\Gamma = -0.03$ and $\Gamma = 0.02$, close to the chance level. The best performance was obtained with 3 layers of 512 units, but is still very poor ($\Gamma = 0.02$). The fact that the training is done at the sentence level, implying some continuity on the output, shows not to be suited to the task. In the experiments using a softmax output layer, the gamma factor was still low with an average of $\Gamma = 0.04$. In that case, when looking at the results at the sentence level, it seems that for most sentences, the number of coincident spikes was equal to the number of spikes in the data, but the number of spikes produced by the model and its frequency were very high, as the output was activated continuously in large regions, as illustrated in Figure 6.6. One of the reason for this continuous activation is the continuity in the labels: the difference in the input features varies smoothly in time. For instance, the features related to the position of the frame in the various contexts (phone, word, sentence), are simply incremented of a time step between two consecutive frames.

The results obtained with respect to the coincidence measure showed the inadequacy of both approaches to predict accurately atom positions from mere linguistic features. To assess the overall performance in the prediction of the local component contours, we investigate the parameters generated when the network is trained with only amplitude and θ parameters, and the reconstruction of F_0 given these system output parameters.

Analysis of Generated Parameters and Reconstructed F_0

A subset of 100 sentences was randomly selected among the 1000 sentences, to calculate the reconstruction of the full F_0 curve using output amplitude and θ generated by the neural networks.

Table 6.1 – Average RMSE between generated and extracted parameters. Left table: amplitude; right table: θ.

Layers \ Units	128	256	512
1	0.03	0.03	0.03
2	0.03	0.15	0.71
3	0.03	0.03	0.03
4	0.03	0.03	—

Layers \ Units	128	256	512
1	0.30	0.30	0.30
2	0.29	0.30	0.30
3	0.29	0.29	0.30
4	0.30	0.30	—

Table 6.2 – Average correlation between generated and extracted parameters. Left table: amplitude; right table: θ.

Layers \ Units	128	256	512
1	0.50	0.51	0.52
2	0.56	0.07	0.08
3	0.51	0.55	0.53
4	0.50	0.51	—

Layers \ Units	128	256	512
1	0.27	0.29	0.27
2	0.30	0.28	0.26
3	0.30	0.30	0.26
4	0.29	0.27	—

First, a rescaling of the data was done, at a global level: the minimum and maximum values of the features had been lost, because of the normalisation (between 0 and 1) before training. Therefore, the data had to be rescaled.

Before reconstruction, we analyse how close the generated parameters are to the parameters extracted from the natural speech. Table 6.1 gives the RMSE between the extracted parameters and the parameters generated by the multiple systems trained only on *positive* frames, while Table 6.2 gives the correlation. There is no significant difference in the performance of the various architectures, except for the systems using 2 layers and more than 128 units per layer which perform worse. The RMSE is very high for the θ parameter, and although it may seem low for amplitudes, considering that the measured values were normalised between 0 and 1, they are also higher than acceptable, because the range of amplitudes is very high and many amplitudes which are relatively small are compressed in the normalised version. The correlation coefficients, which measure how similarly parameters evolve from one atom to another, are also low.

The reconstruction step given the rescaled generated parameters is trivial: from the position, amplitude and θ of each atom, the atom curve can be reconstructed, then the final contour is simply the sum of all local atoms with the phrase component in the log domain.

Table 6.3 gives the median RMSE and correlation between the reconstructed curve and the F_0 extracted from natural speech. Both RMSE and correlation were calculated only for voiced frames, according to the voicing obtained from the TEMPO pitch tracker of Kawahara et al. [1999]. The variance in the results for each system is very large, and the mean performances are affected by large errors for some sentences. The values provided in these tables are to be

Table 6.3 – Median measures between reconstructed and extracted F_0. Left: RMSE (Hz); right: correlation

Layers \ Units	128	256	512
1	42.6	43.8	42.6
2	42.2	65.1	7520
3	40.6	42.6	42.6
4	43.1	41.2	—

Layers \ Units	128	256	512
1	0.918	0.915	0.918
2	0.918	0.813	0.228
3	0.922	0.914	0.921
4	0.915	0.921	—

compared with the RMSE and correlation obtained between the phrase component alone and extracted F_0, and with the same measures between the reconstructed F_0 using extracted atoms and the extracted contour. The median modelling accuracy of the phrase component alone reaches an RMSE of 36.5Hz and a correlation of 0.933. The reconstruction given atoms gives an RMSE of 6.6Hz and a correlation of 0.998. From these numbers, the results are in line with the results on generated parameter errors, and none of the systems seem able to generate coherent parameters. The result is then that the phrase component, which gives a first approximation of the contour deteriorates when adding generated atoms. This means that the statistical modelling techniques employed in this work are not able to predict atoms in this scenario.

General Discussion

From all the results presented before, the main observations are that both SVM and DNN with a softmax layer seem able to learn that certain segments of the signal have a higher probability to feature atoms than others. In the examples shown in Figure 6.6 and 6.7, we can see some activated regions around actual impulses. In the case of the DNN, the regions where the output is 0 are related to silences. It is unlikely to have atoms in the silent regions, from the way they are extracted (based on a probability of voicing and energy weight, see Chapter 5). Consequently, from the silence labels, the model seems able to learn that there is a very low probability of having atoms.

Although the proposed methods are able, in most cases, to find regions with high probability of having atoms, they both fail in the precise localisation of these atoms. A few reasons could be the cause of this difficulty: no weight related to the importance of the atom (e.g. amplitude) is given as an information for the system to know which samples are more important than others. Consequently, an atom with an amplitude of 0.05 will have the same effect as an atom with an amplitude of 3 in the training, although these atoms do not have the same role in the modelling of the intonation contour. Another factor is the absence of information regarding other atoms: considering the whole F_0 contour, it is obvious that atoms are not totally independent from each other. Therefore, some information about other atoms should be provided to the model in order to predict atoms at a specific frame.

In the case where the training was aimed at synthesising parameters only for *positive* frames, the architectures investigated were not able to learn from the data. This may be partly due to the construction of the networks, which rely on continuity in the input and output, and is absent in this scenario. The task of predicting discrete events may not be realisable using this type of network. Even though the parameters generated are different from the ones extracted from the signal, the final reconstruction of the F_0 contour could still be somehow natural. Analyses showed that it was not the case, and that the synthetic atoms tended to cause the base component to deteriorate rather than improve. Informal listening tests confirmed that the resulting output was not suited to speech synthesis.

6.5 Conclusion

In this chapter, we investigated the use of the generalised command-response model proposed earlier for generating intonation, in the context of speech synthesis. A listening test showed that the position of high magnitude atoms was crucial to perceive the speech as identical to the original samples. This implies that the most prominent atoms are located in key regions of the intonation contour and that disturbing their position, even from a very small shift, is perceptible by native listeners.

Several standard statistical methods were explored to try to predict the parameters of the model, in the TTS framework. SVM and DNN proved able to learn regions of the sentence where atoms were likely to exist, but failed to predict accurately the position of these atoms. Several reason may explain this inability, including the dependency between atoms which is not modelled and their relative importance, which is also ignored and therefore gives an equal weight to all atoms in training.

In trying to divide the task of predicting position and other parameters, DNNs were trained using only frames which contained atoms. The architectures investigated did not allow synthesis of acceptable contours, and the synthesised parameters showed low correlation and large error with the extracted ones. The fact that the parameters correspond to discrete event description and that the neural networks model some continuity in this framework may be partly responsible for the low performance observed.

The use of our GCR model to model of all the local intonation components for synthesis was revealed to be a challenging task, but there are several possible alternatives that could be investigated for intonation event position prediction. Modifying the cost function in the DNN training to optimise the gamma factor at the sentence level could be interesting. Another idea could be to adapt the method proposed by Kameoka et al. [2015] to the GCR model, removing some constraints on the components and their non negativity. Another approach, inspired by the way we model intonation, would be to investigate spiking neuron-based methods. Finally, the way the output of the systems was post-processed might not be optimal, and a precise location of maxima could reduce the imprecision of the models.

Although these lines of research seem appropriate to aim the synthesis of intonation, this thesis further exploits some properties of the model in a different way. By construction and its parameterisation, the model lends itself to studying specific prosodic events. In this direction, in the next chapter, we focus on the analysis of events in the intonation related to prominence.

7 Intonation-based Emphasis Transfer

As was underlined in the introduction (Chapter 1), speech-to-speech translation systems are a reality, at least for some of the most resource-rich languages. However, at this stage, the synthesis part of such systems outputs a generic synthetic voice, which can fast turn monotonous in use, but also neglect the transmission of the implicit meaning of a sentence.

In this chapter, we propose to use the GCR model described in Chapter 5 for the synthesis of emphasis. As a complementary task to speech synthesis and intonation modelling, emphasis can add value to synthetic speech, especially in the S2ST context as it would allow translating more than just textual information.

After an analysis of model parameters in the neutral and emphasised contexts, we propose two approaches exploiting atoms to produce emphasis in synthetic speech. The first one simply consists of the transfer of local components from an emphasised word to a neutral word with the same context. The second approach consists of training predictors to generate atoms in the desired linguistic context, for an emphasised word.

The contributions presented in this chapter contain some unpublished work, and some extension of the work published in the following papers:

- Pierre-Edouard Honnet and Philip N. Garner. Emphasis recreation for TTS using intonation atoms. In *Proceedings of the 9th ISCA Speech Synthesis Workshop*, pages 14–20, Sunnyvale, CA, USA, September 2016a
- Pierre-Edouard Honnet and Philip N. Garner. Intonation atom-based emphasis transfer. Idiap-RR Idiap-RR-14-2016, Idiap, 5 2016b

This work is closely related to some collaborative work on emphasis detection published in the two following papers:

- Milos Cernak and Pierre-Edouard Honnet. An empirical model of emphatic word

detection. In *Proceedings of Interspeech*, pages 573–577, Dresden, Germany, September 2015

– Milos Cernak, Afsaneh Asaei, Pierre-Edouard Honnet, Philip N. Garner, and Hervé Bourlard. Sound pattern matching for automatic prosodic event detection. In *Proceedings of Interspeech*, pages 170–174, September 2016

7.1 Background

7.1.1 Prosody in S2ST

To improve S2ST systems, speech researchers have tried to personalise them, using for instance cross-lingual speaker adaptation, which is the adaptation of TTS models in the target language (L2), given some data in a different language, the source language (L1). This is typically achieved by mapping input and output language TTS models to evaluate transforms at the model or at the feature level, e.g. the work of Liang and Dines [2011]; Wu et al. [2009]; Yoshimura et al. [2013].

On prosodic aspects, modelling around the S2ST context has gained interest in the last decade, with first approaches proposed by Agüero et al. [2005]; Agüero et al. [2006]. This work consisted of the exploitation of information in the intonation of input speech in the source language to improve the naturalness of the synthetic speech in the output language. By using a modified version of the CR model of Fujisaki and Nagashima [1969] to characterise the intonation of the source speech, this approach relies on a mapping of accent groups between languages. This method allows feeding source language information to annotate text in the target language in the S2ST framework, if a parallel corpus is provided to train the annotation model, and the languages are close: the clustering algorithm works under the assumption that some pitch movements in one language have a correspondence with pitch movements in the other language.

7.1.2 Emphasis in S2ST

In this chapter, we consider emphasis for an isolated word or group of words, from a synthesis point of view, meaning that we are interested in generating some target (group of) word(s) in a more prominent way than the rest of the utterance. Emphasis is introduced in Chapter 2, Section 2.3 in the context of automatic speech processing, including automatic emphasis detection, synthesis and emphasis transfer in S2ST.

In an S2ST scenario, one can easily understand that the users may be interested in translating not only their speech, but also their intentions or underlying meaning, through prominence. A parallel can be made with instant messaging, or social networks, in which users are inserting “smileys”, “emoticons” or “emojis” to make sure that the other users understand the intentions of the message. Tsiartas et al. [2013] conducted a large-scale human evaluation on the per-

ception of S2ST quality and showed that the perceived quality of S2ST was correlated with cross-lingual prosodic emphatic transfer. In other words, emphasising the correct words in the output language in TTS based on the emphasised words in the input language helps in the S2ST task. It is straightforward to imagine a system that can retrieve emphasis, tag the words that were emphasised in the translation, and synthesise speech taking this prominence into account, like in the example provided in Figure 7.1.

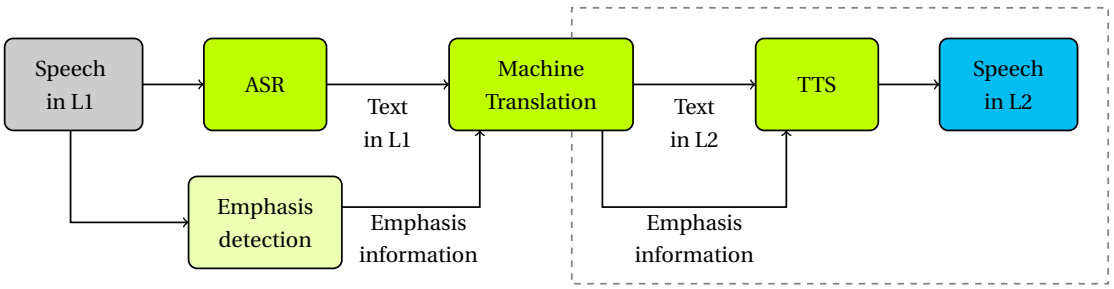

Figure 7.1 – Integrating emphasis in speech-to-speech translation.

In such a setting, emphasis detection can be achieved by any method, purely based on signal processing, such as the work of Kennedy and Ellis [2003], Arons [1994], Heldner et al. [1999] or Cernak and Honnet [2015]; or using data-driven models which focus on the difference between neutral speech and emphasised speech, like the work of Liang [2016] or Cernak et al. [2016]. In this chapter, we are concerned with the part in the grey dashed frame: emphasis synthesis.

The task of emphasis synthesis has been approached in unit-selection-based TTS by Strom et al. [2007] using recordings of emphasised speech covering the phonetic space of diphones. Another data-driven approach proposed by Yu et al. [2010] exploits decision trees to distinguish emphasis from neutral speech in the training of HMM-based TTS systems. Other methods applying post-processing for instance to the intonation were proposed by Hirose et al. [2012]; Ochi et al. [2009].

Alternative approaches trying to directly exploit the input data to reproduce emphasis in the output language have been proposed by Anumanchipalli et al. [2012], with a mapping between input and output intonational feature vectors; or more recently, Do et al. [2015a] who proposed a more complete and integrated framework with the use of linear regression HSMMs to preserve word-level emphasis in S2ST. The addition of pause "transfer" was proposed later [Do et al., 2015b], and the recent trends in speech processing made previous methods evolve toward deep learning [Do et al., 2016a]. Pause transfer was investigated earlier in this context by Agüero et al. [2008].

7.2 Emphasis Analysis using the GCR Model

To assess the relevance of the parameters extracted from our model, we examined the mutual information shared between the parameters and some linguistic features. Our hypothesis is that the statistics on parameters will differ between the neutral and emphasised case.

7.2.1 Data and Model Settings

Analyses are conducted on several datasets: a subset of the *SIWIS* database described in the previous chapter, in Section 3.2, which contains sentences with emphasised words, and their neutral version. The emphasised part of the French female speaker database, described in the same chapter, in Section 3.3, was used in the same fashion. A subset of the Blizzard 2011 database [Wilhelms-Tricarico et al., 2011] was used to evaluate mutual information between parameters and accent and stress. Finally, the part of the Blizzard 2008 [Karaiskos et al., 2008] database containing emphasis was also used in the analyses.

The analyses consist of finding the relevance of our model parameters with respect to linguistic cues. For this reason, in the remainder of the chapter, atom parameters are the features employed. They were extracted using the weighted correlation algorithm described in Chapter 5, with the following parameters fixed:

- the order of the components was fixed to $k = 6$, following results obtained in Chapter 5.
- the range of possible θ for the local components was 0.01–0.05.
- the two stopping criteria used were: a weighted correlation threshold of 0.99 in the speech segment, i.e. excluding the starting and ending silences, and a maximum of 10 atoms per second.

7.2.2 Atom Frequency

For this first analysis, we selected three speakers numbered *26*, *28* and *29* from the *SIWIS* database. For speaker *29* (female), we used both English and French data, for speaker *26* (male) English data and for speaker *28* (male) French data. These speakers were selected because the vocoder used — the STRAIGHT vocoder [Kawahara et al., 1999] — worked well for them. For each language and each speaker, we used 25 neutral sentences and the 25 same sentences with emphasis on a word. Thus, 100 neutral sentences were compared with their emphatic versions.

We first look at the number of atoms needed to model the local behaviour of F_0 in the emphasised word. We do not investigate duration modifications in this work. However, to compare the number of commands in the neutral and focused case, we measure the duration of the word under investigation for each sentence in table 7.1. The average difference between the

Table 7.1 – Number of atoms and additional duration (in sec.) needed on average for target word per speaker.

Speaker	Neutral	Emphasised	Difference (norm)	Duration difference (sec.)
29 (EN)	5.28	7.48	2.20 (-1.98)	0.28
26 (EN)	5.36	8.80	2.03 (2.58)	0.20
29 (FR)	5.08	7.96	2.88 (-2.56)	0.33
28 (FR)	7.20	10.56	3.36 (-1.05)	0.27
ALL	5.73	8.70	2.97 (-0.75)	0.27

duration of the emphasised word and the neutral word is calculated and given with the average number of atoms and their difference (emphasised - neutral) in the two contexts for each speaker. The difference between the number of atoms required for emphasised and neutral cases is also given when normalised over time inside brackets in the 4^{th} column.

As we might expect, more atoms are needed to model the target word in the emphasised case. We might think that one of the reasons for this is the fact that the words have a longer duration, but looking at the difference in number of atoms normalised over the duration of the words (inside brackets in the 4^{th} column), we can see that in average, there are fewer atoms per second in the emphasised version of the word. This is interesting as it shows that the way the atoms are distributed in the emphatic word is not only related to the duration of the word, as compared to the neutral case.

By comparison, the regions outside the target word typically have 30 atoms, and require just 3 more on average in the emphatic case. The ratio of numbers of atoms between emphatic and neutral is 1.1 ± 0.04 on average, which can be explained by a slightly slower speaking rate, used for increasing the emphasis on the target word (for duration, the ratio is 1.19 ± 0.02).

To evaluate further the relation between parameters and linguistics, we analyse mutual information in the following section.

7.2.3 Mutual Information

By looking at mutual information, we expect to find some clues on how atom parameters relate with linguistic features. We first measure the mutual information between atom parameters (amplitude, position and θ) and classical contextual features, and then look at the differences between neutral and emphatic data. Our hypothesis is that when a word is emphasised, the model will extract different types of atoms, then we should observe some difference in mutual information between emphasis and atom parameters. If we denote the labels as L and the model features as F_i, the mutual information was calculated the following way:

$$I(L,F_i) = \sum_{l} \sum_{f \in F_i} p(l,f) \log_2 \left(\frac{p(l,f)}{p(l)p(f)} \right) \tag{7.1}$$

with $p(l, f)$ the joint probability of L and F_i, and $p(l)$ and $p(f)$ their respective marginal probabilities. These probabilities are maximum likelihood estimates based on occurrences in the data. The model parameters were quantised in different ways for the different databases, according to their distributions. For the relative position of the atom in the word, or in the syllable, a simple linear quantisation was done between 0 and 10 for all the datasets. For θ values, the original values were in the range 0.01–0.05, and were simply scaled to be in the range 1–9, for all the datasets. Amplitudes were quantised differently as different distributions were found for the different datasets. Appendix A gives the parameter distributions for the French database, on which the rescaling was based. For the *SIWIS* data, amplitude was rescaled between 0 and 10; for the *Roger* data, it was rescaled in the range -4–5, and for the French data, between -10 and 10. The labels l are binary. The results presented are normalised by dividing the mutual information by the entropy of the labels, defined as:

$$H(L) = -\sum_{l} p(l)\log_2(p(l)) \tag{7.2}$$

Table 7.2 shows normalised mutual information in the case of the single English female speaker of the Blizzard 2011 database, for about 300 neutral read sentences. The labels are syllable level labels.

Table 7.2 – Normalised mutual information between atoms and linguistic features for neutral speech, at the syllable level.

	Pos	Amp	θ	Amp & θ	Pos & Amp	Pos & θ	Pos & Amp & θ
Accent	11.1	11.1	8.7	13	23.3	22.6	23.3
Stress	8.1	8.2	6.8	9.5	22.3	21.8	22.2
Acc. & Stress	13.9	13.4	11	16	27.1	26.6	27.3
Acc. or Stress	7.7	7.9	6.2	9.1	23.8	23.2	23.7

The results indicate that mutual information between amplitude and accent and between relative position in the syllable and accent are the highest for single feature and single context. As can be expected, the syllables which are both accented and stressed have higher mutual information with atom parameters. We also notice that using all atom parameters (amplitude, position and θ) does not bring more information than using position and amplitude only.

Given these initial observations, we looked at the mutual information between amplitude, position and number of atoms per syllable, and accent, stress and emphasis. In that case, the data consisted of about 300 sentences from several English speakers, in two scenarios: neutral sentence and sentence with one emphasised word or group of words. This data comes from the multilingual *SIWIS* database. To compare both cases, the same "*target words*" were used: the word emphasised in the emphasised case was selected as *target word*, and in the neutral case it was also tagged as emphasised, to see its effect on the parameters. The results are presented in table 7.3.

Table 7.3 – Normalised mutual information between atoms and linguistic features [neutral / emphasised].

	Position		Amplitude		N_{atoms} in Syllable	
	Neut	Emph	Neut	Emph	Neut	Emph
Accent	14.1	14.9	12.4	13.0	8.4	8.8
Stress	11.4	11.5	10.3	10.4	7.3	7.8
Emphasis	24.0	20.6	20.8	17.4	11.9	18.9
Acc. & Stress	18.0	18.8	15.6	16.0	10.3	10.8
Emph. & Stress	48.3	35.2	40.5	29.8	26.6	44.4
Emph. & Acc.	60.4	55.8	53.8	47.5	38.2	56.1

In this table, we can see that the highest difference in mutual information between emphasised and neutral data is observable between emphasis and the number of atoms in the syllable. This contradicts our hypothesis that the atoms in an emphasised word are different from the ones in a neutral word. One possible explanation for this finding is that the F_0 curve presents more variations in the region of emphasised word, resulting in a need for more atoms to fit the curve. To verify further the difference between the "principal" atoms in each word, we looked at the same measures as in table 7.3, but with a constraint on the number of atoms: we selected only the first n atoms — ranked by amplitude — in the emphasised word, where n is the number of atoms in the same target word in the neutral case. In the cases where the neutral version had more atoms, the number of atoms kept was restricted in the same way, to always have the same number of atoms. Table 7.4 gives the results for mutual information with the same number of atoms.

Table 7.4 – Normalised mutual information between atom parameters and linguistic features [neutral / emphasised] using the same number of atoms.

	Position		Amplitude		θ	
	Neut	Emph	Neut	Emph	Neut	Emph
Accent	14.3	14.8	12.5	13	9.9	10.6
Stress	11.3	11.8	10.2	10.3	8.5	8.9
Emphasis	28.5	29.8	24.6	25.2	21.7	23.7
Acc. & Stress	18.3	18.9	15.7	16.4	12.1	12.9
Emph. & Stress	57.8	60	48	51.9	41.7	46.9
Emph. & Acc.	74.9	78.1	65.6	69.9	60	62.8

We can see that when the number of atoms is the same for the neutral and emphasised case, mutual information between both amplitude and position and accent, stress and emphasis is higher in the emphasised case. This is particularly true for emphasis. Then, in addition to the fact that emphasis manifests itself with more atoms, emphasis seems to be expressed through different patterns for the components: different positions, amplitudes and θ. It validates our intuition, that when a word is emphasised, the components resulting from the decomposition are distinguishable from the neutral case.

Table 7.5 gives the same results as table 7.4, for *Roger* data, on a bigger number of sentences, but with fewer different contexts for the emphasised words. In this case, there is no neutral version of the sentences, and there are only 3 different sentence schemes, as described in Chapter 2, Section 2.3. We also investigate the weighted correlation obtained during decomposition, which is directly linked to amplitude, but carries additional information about voicing and energy. More details on the weighted correlation are given in Chapter 5, Section 5.3.3.

Table 7.5 – Normalised mutual information between emphasis and atom parameters on Roger data.

	Position	Amplitude	θ	wcorr
Emphasis	11.0	9.3	7.6	10.1
Accent	11.6	9.0	8.3	10.1
Stress	16.3	10.8	9.6	11.5

All the features seem to behave in a similar fashion, and the results indicate that in this case, stress is the feature which shares most information with atom parameters. It can be explained by the fact that the sentences are short, and the number of emphasised words is not so unbalanced compared to the number of neutral words. In that case, we can expect that stress is more important. The results can also be explained by the expressivity of the data. The speaker delivered rather expressive speech, and would have naturally emphasised stressed syllables in all words more than an average speaker.

We also performed similar analyses on the French female speaker database, but the only contextual label investigated was the emphasis of the word. French being a language lacking lexical stress, it is not straightforward to obtain reliable annotations for stress and accent from text analysis. Table 7.6 provides normalised mutual information between the emphasis binary label and atom parameters, for the sentences containing emphasis and their neutral versions, so a set of 1575 sentences in each case. The labels and relative positions were calculated at the word level in this case.

Table 7.6 – Normalised mutual information between emphasis and atom parameters for French.

	Position		Amplitude		θ		wcorr	
	Neut	Emph	Neut	Emph	Neut	Emph	Neut	Emph
Emphasis	56.7	57.3	98.4	101.1	23.9	27.4	27.4	34.7

As was done for the results presented in table 7.4, we present analyses with restricted number of atoms, to compare the same numbers. The results present the same behaviour as for the English data analyses. In this case, θ seems to present larger differences than amplitude and relative position, hinting larger atoms rather than higher amplitude ones. The weighted correlation seems to be especially important to distinguish emphasis from neutral words, which is logical as emphasis would also manifest itself in energy.

From all these analyses, the main conclusion that can be drawn is that the atoms extracted in

emphasised words have more pronounced features than in neutral words. However, it is not obvious how intonation is affected by the presence of emphasis on a particular word.

Preliminary observations revealed that using the GCR approach to model the F_0, and simply modifying the amplitude or θ of the atoms, i.e transforming them into bigger atoms, is not resulting in emphasis. The next sections present two approaches to simulate emphasis in speech, based on atom transfer and on atom prediction.

7.3 Atom Transfer

As discussed earlier (Chapter 6, Section 6.5), our intonation model lends itself to emphasis transfer in the sense that it has local components shaping the F_0 contour, and the combination of these components to shape word contours indicates specific prosodic events, such as emphasis. We showed that the model parameters share mutual information with the emphasis label, where the emphasis label is an annotation of known word emphasis. In the same line of research, it was recently demonstrated by Delić et al. [2016] that the model atoms correlate with ToBI markers [Silverman et al., 1992]. Another use of the model was found in the detection of emphasis or stress [Szaszák et al., 2016]. Based on these results, we believe that atoms can convey emphasis. Our hypothesis is then that using atoms from an emphasised word can elicit emphasis in a neutral word with a similar context.

7.3.1 Transfer of Prominent Atoms

In a first attempt at generating emphasis in neutral speech, we investigate the simple addition of some local components from an emphasised word to a neutral version of the same word in the same context. Figure 7.2 gives an overview of the approach.

The idea here is to identify the most prominent local atoms in the emphasised word, and as a proof of concept to transfer them to a neutral version of the same sentence. Doing so on natural speech will allow evaluating the capabilities of prominent atoms in different versions of the same word in the same context.

The procedure is then the following:

1. The atoms are first extracted in both scenarios (neutral / emphasised).
2. Given the position of the words in the sentence through automatic label alignement, and the knowledge of which word is emphasised, the atoms in the target word are identified.
3. Only the most prominent atoms are selected.
4. As the two versions of the sentence have different durations, for the target word and the other words, we calculate the relative position of the atoms in each syllable.

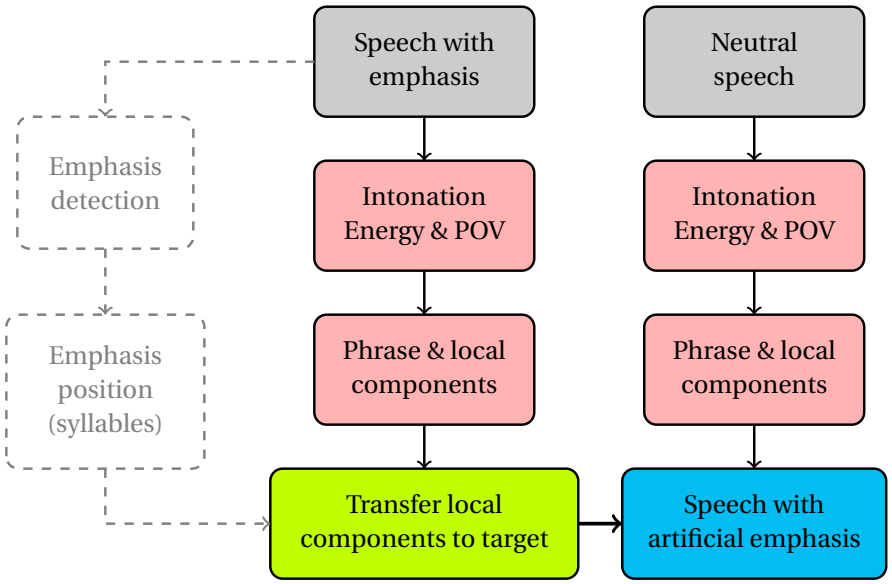

Figure 7.2 – Emphasis transfer in a neutral sentence. POV stands for probability of voicing. Inputs are grey, features light red, output is blue.

5. Selected atoms are transferred to the corresponding syllable in the neutral sentence target word.
6. The F_0 contour is reconstructed
7. The speech is resynthesised.

To show that most prominent atoms are important for emphasis perception, we select only the most prominent atoms — atoms with the highest absolute amplitude — and transfer them to the approximate position in the neutral sentence. In this particular context, it is easy to find the corresponding position, because the words are the same and thus have the same number of syllables at the same position; moreover we can assume that their relative durations are extended in a similar fashion.

7.3.2 Evaluation

Data

Our goal is to investigate local emphasis on some words in full sentences. For that, we use a part of the multilingual *SIWIS* database, described in Chapter 3 Section 3.2. The experiments were conducted on the parallel set of sentences with emphasis: each sentence was uttered once in a neutral way, and once with specific focus on a predefined word. The speakers were told which word to emphasise before reading the focused version.

The data used for the emphasis transfer experiment is a subset of the dataset described above:

speaker *29* was selected, and the evalution was carried out only on the English sentences, to ease subjective listening tests.

Experimental Setup

Initial experiments showed that adding atoms from the syllable which directly precedes the emphasised word did not bring perceivable difference, therefore only the main atoms inside the target word were transferred. It was found empirically that transferring more than 3 components did not improve the perception of emphasis, thus only 3 atoms — or fewer in the case where there were fewer atoms in the word — were given to the neutral sentence. It is also in line with the average additional number of atoms needed to model the target word found in Section 7.2.2. In particular, for the 20 sentences selected for this speaker, we found that 2.4 additional atoms were needed on average in the emphatic case. The procedure described earlier is used to extract, select and transfer atoms to the neutral word. The STRAIGHT vocoder [Kawahara et al., 1999] was used to extract spectral parameters and for resynthesis.

Listening Tests

A subjective listening test was conducted to evaluate the validity of our approach. The listeners were asked to listen to the samples in a random order and identify which word in the sentence sounded the most emphasised, and for this word give a level of emphasis with a 3-level choice: *clear*, *moderate* or *slight* emphasis. Each subject had to listen to 3 versions of 20 sentences, S1–S20, for a total of 60 audio samples. One version consisted of the original neutral sentence, another one the original sentence with emphasis, and the last one the neutral sentence with artificial emphasis. An example for each level of emphasis was given in the instructions, to understand how to rate the degree of emphasis. The listeners always had to identify a "most emphasised" word in order to control that emphasis transfer had an effect compared to the neutral sentences. We expected listeners to rate the neutral sentences as *slightly* emphasised, the original emphatic version as *clearly* or *moderately* emphasised, and the artifically emphasised version closer to the emphatic version, as the aim is to increase the impression of emphasis on the target word.

30 subjects participated in the test, with a high majority of non native fluent English speakers, most of them being in the age range 26–35.

7.3.3 Results

Figure 7.3 shows an example of transfer for a sentence with the two $\log F_0$ contours of the same sentence in the two different contexts, and the resulting contour (*S6* in the results). For the $\log F_0$ curves, the green indicates a high probability of voicing, while the blue indicates a high probability of being in an unvoiced region. The syllable boundaries are displayed, with the coloured region being the target word, *"somewhat"*. In the bottom panel, we can see the

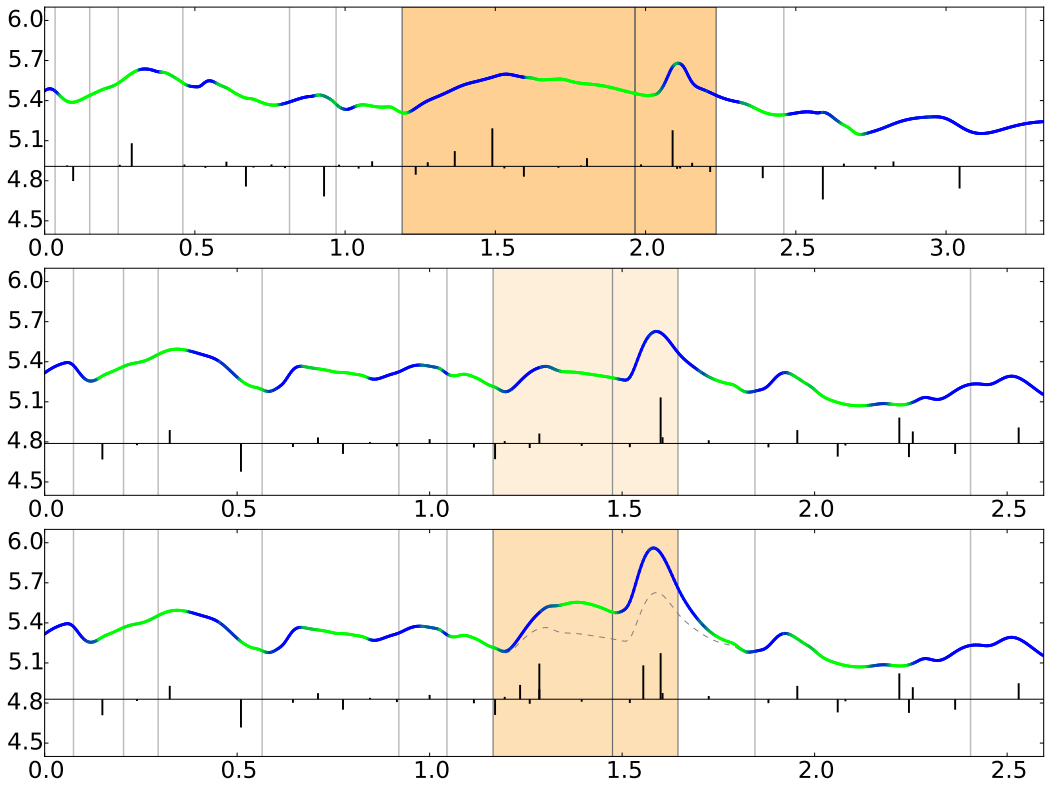

Figure 7.3 – Example of $\log F_0$ contour and local commands for the sentence *"The matter seems to be somewhat confused."*. Top panel: sentence with emphasis on the word *"somewhat"*. Middle: neutral sentence. Bottom: neutral sentence with transferred emphasis on the word *"somewhat"*.

original neutral contour in dashed grey, while the modified curve presents an increased F_0. The atom commands are displayed in black, and we can see that 3 components were added to the neutral sentence.

Figure 7.4 shows the number of people identifying the target word as most emphasised for each sentence. For each triplet of bars, the left-most one corresponds to the neutral version of the sentence, the middle one is the neutral with emphasis transfer and the right-most the original emphasised version. The height of the full bars corresponds to the number of votes for the target word independently of the strength of the emphasis. The different colours account for the level of emphasis that the voters chose when they chose the target word. The darkest (bottom) colour stands for *clear* emphasis, medium for *moderate* emphasis and the lightest one (top) for *slight* emphasis.

We observe 2 main trends in the results:

- In 8 cases — *S1, S3, S4, S6, S7, S17, S20* — the number of people perceiving the target

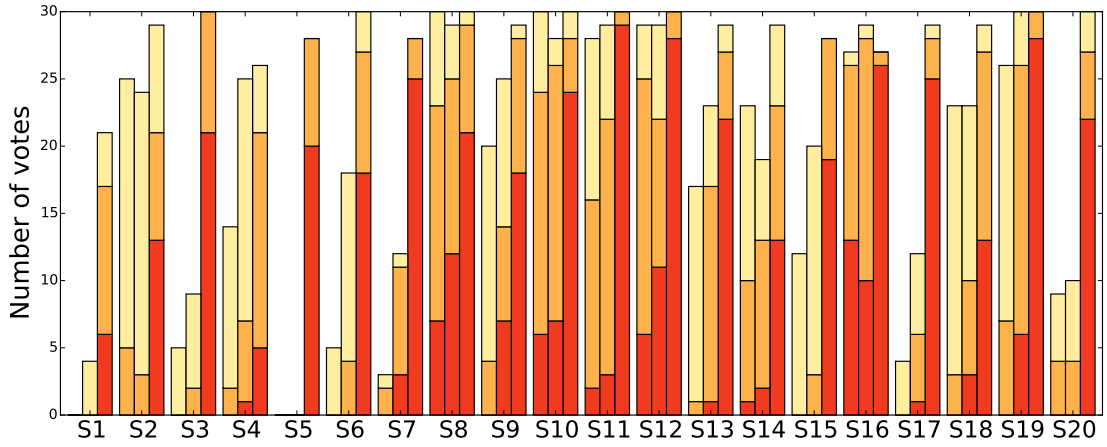

Figure 7.4 – Atom transfer subjective listening test results. Red = clear, orange = moderate, yellow = slight emphasis. A missing bar indicates 0 vote for the target word.

word as emphasised increased. For 3 of these cases, a majority of people voted for the target word when intonation was modified. For the other 5 cases, the perception of emphasis increased significantly when modifying the intonation, but did not reach the majority of votes. These 8 cases showed that the emphasis is consistently shifted towards the target word, with a higher level of emphasis.

- In 11 other cases out of the 20, the majority of the listeners voted for the target word in the neutral case, even though the speaker did not have any particular instructions. For 4 of these cases, adding atoms decreased the number of votes for the target word, however in all these cases, the number of subjects choosing a *clear* emphasis increased, and the number of *moderate* emphasis also increased. In 2 cases the total number stayed the same, but there was an increase in the number of *clear*, and in *moderate* emphasis. In the 5 other cases, the total number always increased and the level of emphasis was also rated higher. These 11 cases showed that when the emphasis is already perceived on the target word, its strength is increased when adding emphasis atoms.

- In the last case (*S5*), adding local components from the emphasised word intonation was not enough to make the perception of emphasis change for the listeners, the target word being a non content word. Most of the listeners kept the main content word as most emphasised.

Discussion

The global trend in the results confirms the hypothesis: transferring local components from an emphasised word to a neutral sentence increases the impression of emphasis in the target word in most of the cases. We can also see that the way emphasis is perceived — in other words how strong the emphasis is — is affected by adding local positive or negative components. The modification of the resulting intonation contour increases the strength of the emphasis.

In some cases, emphasis was not perceived on the target word mainly because of the *reset* at the beginning of the sentence — sentences start with a raising intonation before gradually decreasing. It may have been confusing for the listeners — and this was feedback from some of the listeners — to choose between a slightly emphasised word in the middle of an utterance and the natural higher pitch that occurs at the start of speech.

We cannot expect the intonation alone to help the listeners to perfectly perceive the emphasis on the target words, however the results indicate that it consistently increases the perception of the emphasis and its strength. The initial hypothesis, that adding atoms extracted in an emphasised word in natural speech to a neutral version of the same word can elicit emphasis perception, is then demonstrated to some extent. In the next section we go on to try modelling the emphasis in different contexts, enabling the synthesis of emphasis on a target word without having an emphasised version of the same word.

7.4 Atom Generation for Word-level Emphasis

In this section, we investigate the use of clustering methods to predict the local intonation components of emphasised words. It fits in the *emphasis synthesis* part of the diagram in Figure 7.1, in the grey dashed frame. Using emphasised word F_0 decomposition in context, we attempt to predict the model parameters for an emphasised word in some specific context. The frame-wise prediction of the model parameters attempts in Chapter 6 were revealed to be challenging. To understand it better, we modify the parameterisation of the model characteristics and try to model jointly word-level components. Our hypothesis is that the local model components can be used as word-level intonation to synthesise emphasised words. We restrict ourselves to the intra-lingual case, but due to the language independence of the intonation model used, it seems reasonable to assume that this method can work for any given language.

7.4.1 Atom Generation using Random Forests

The general idea of the proposed approach is illustrated in Figure 7.5. Having an external model which takes linguistic features as input allows generating emphasis related parameters in the same context as for traditional speech synthesis.

Regression Trees

Decision trees were briefly introduced in Chapter 2, in the context of state clustering for HMM-based speech synthesis. A decision tree is a clustering method which splits a training dataset based on some criterion at each node of the tree except the leaf nodes. In a classification decision tree, the leaf nodes contain class values. In the case of a regression tree, the value is the mean of the training samples observed in the branch. If an unseen context is given, the

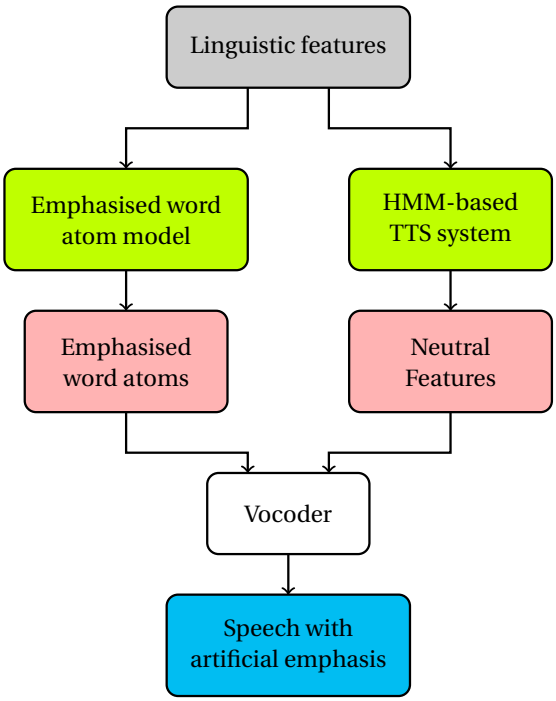

Figure 7.5 – Emphasis synthesis using external atom generation. Input is grey, models green, features light red, output is blue.

output will be the average of the leaf nodes which are reachable for the most similar context.

Random Forests

Random forests, introduced by Breiman [2001], are a combination of tree predictors. The idea behind random forests is to train multiple decision trees on a subset of the training data, to avoid overfitting of single trees to the training set. These subsets are generally created by randomly selecting data from the full training set. At test time, the output is the mode of the classes of the individual trees. For these reasons, random forests are known to work better than single trees. A random forest can be used for regression as well, on the exact same principle. In this case, at test time, the mean of the trees is the output. Because of their demonstrated better modelling capabilities, random forests were investigated for prediction of atom parameters given word-level linguistic features.

Predicting Word-level Parameters In order to exploit random forests, we propose the following strategy: contextual factors are used to cluster atom parameters, where the parameters are described in a single output vector. In each emphasised word, some preliminary observations to characterise the system showed that most variations were captured using 5 or fewer atoms per word. This leads to a heuristic but reasonable upper limit on the number used in the experiments. It results in an output vector of dimension 15. In the cases where the number

of atoms is lower than 5, the vector is filled with zeros to meet the desired size. The first 5 extracted atoms — the 5 atoms with the highest weighted correlation — are selected, and then ordered by position, to keep a consistency among all emphasised words.

Two strategies were investigated to integrate the generated atoms in synthetic speech:

- Replacing local atoms in the target word by the generated atoms. This is done by extracting model components on the synthetic F_0 contour, identifying the atoms inside the target word, removing them and adding the generated components to the contour.
- Adding generated atoms to the synthetic F_0. In a similar fashion to the approach described in Section 7.3, the output of the system was simply added to the F_0 contour in the target word.

7.4.2 Evaluation

Data

We used two English datasets containing emphasis in these experiments: a subset of the *SIWIS* emphasis data, and the emphasis data from the Blizzard challenge 2008 database, both described earlier. Additionally, a TTS system trained on the WSJ corpus was used.

Training / Testing Sets 100 neutral sentences from speaker 29 of the *SIWIS* database were used for adaptation of our HMM-based TTS average voice. Then a ten-fold cross-validation strategy was used for emphasis experiments: 5 combinations of 20 adaptation sentences and 5 test sentences were tested, to be able to synthesise all the test files without having them in training / adaptation sets. The same strategy was adopted for the training of random forests. For the experiments on *Roger* data, a single speaker HMM-based TTS system was trained on about 4400 sentences which did not contain specific emphasis. We used a total of 1671 sentences with one or two emphasised words (2211 emphasised words). This set was divided into a training set of 1879 emphasised words, corresponding to 1476 sentences, and a testing set of 332 words, from 321 sentences. As the words were processed independently, for the sentences with two emphasised words, one could be in the training set and the other in the testing set.

Features

The linguistic features investigated are word-level features, including syllable-level information concerning stress and accent. For each word, the position of the stressed and or accented syllables was used with a maximum of 3 stressed syllables and 2 accented syllables[1] The

[1] using 3 accented syllables was investigated as some words in the data contained 3 accented syllables, but the position of the 3rd accent was found not to be informative.

selected set of features is a subset of the standard linguistic features of HMM-based speech synthesis:

- Number of syllables in the word
- gPOS (guess part of speech) of the word
- Stress position(s) in the word [0-3]
- Accent position(s) in the word [0-2]
- Word position in the phrase
- TOBI endtone of the phrase
- Phrase position in the utterance
- Number of phrases in the utterance
- Number of words in the utterance
- Number of syllables in the utterance

Systems

Baselines For the experiments on *SIWIS* data, an HMM-based TTS average voice was trained using WSJ corpus SI84 dataset (described in Chapter 2, Section 2.4. The voice was adapted to the target speaker characteristics by using 100 neutral sentences. Test sentences were synthesised using these models, to generate a neutral synthetic version. Then a second adaptation was done using 20 emphasised sentences to synthesise the 5 remaining sentences[2]. The second adaptation was done 5 different times to cover all the possible test sentences. In the end, two versions were available for each sentence: one from the neutral HMMs, one from the emphasis-adapted HMMs. For the second version, time alignment was used at synthesis time, as the duration models deteriorated severely during adaptation.

For the experiments on *Roger* data, an HMM-based TTS system was built, and the test sentences were synthesised in two fashions: one without giving any other information than the standard contextual features obtained from text analysis, and one with timing information, obtained from automatic time alignment done using the original sentences.

[2] An attempt was made to build an HMM system using automatically labelled emphasis on the same neutral speech, followed by adaptation using manual emphasis annotation on emphasised data, and then synthesise emphasised words, but it did not work, most probably because the automatic labelling of the neutral data performed poorly.

Random Forests For the *SIWIS* data, two strategies were investigated: training random forests to predict word-level atom parameters, which were later adapted using emphasised data, and training random forests using only emphasised words. In the first case (denoted RF1 later), about 60,000 words from the WSJ corpus were used to train 25 trees. Then the emphasised words from 303 sentences from the SIWIS database were used to refine the model with 3 more trees, and finally 20 of the 25 sentences from speaker 29 were used to train 2 more trees, while the remaining 5 were kept for testing. This was repeated 5 times for the different sets to cover all the test sentences. In the second case (denoted RF2 later), the initial 15 trees were trained using the 303 sentences, and then, as in the first case, 20 sentences were used to train 3 more trees. For the *Roger* set, multiple numbers of trees were investigated, in the range 15–30. All the training data was used in this case. The random forests were trained in both cases using mean square error as a training criterion, with no limitation on the size of the trees.

Addition and Replacement The addition of atoms was investigated only for the *SIWIS* data, and ought to be compared with the approach proposed in Section 7.3. For that reason, the addition of atoms was performed on the neutral HMM output F_0, that we expect to be similar to the neutral real speech. For each test sentence, the five generated atoms were placed at their respective predicted positions in the target word.

The replacement of atoms was performed on both *SIWIS* and *Roger* data. After generation of the atoms, the model parameters were extracted on the synthetic F_0, the atoms in the target word dismissed, and the five predicted atoms were added to the global contour in this word.

Listening Tests

A listening test was conducted on systems from both databases. Table 7.7 gives the description of each system presented to the listeners.

Table 7.7 – System description.

System name	Description
roger-voc	Vocoded emphasised version of the sentence (*Roger*). **Reference**.
roger-hmm-neut	HMM-based output (*Roger*). **Baseline**.
roger-hmm-align	HMM-based output using time alignment (*Roger*).
roger-replace-rf	Local component replacement with random forest output (*Roger*).
siwis-voc-neut	Vocoded neutral version of the sentence (*SIWIS*).
siwis-voc-emph	Vocoded emphasised version of the sentence (*SIWIS*). **Reference**.
siwis-hmm-neut	Neutral HMM-based output (*SIWIS*). **Baseline**.
siwis-hmm-emph	Adapted HMM-based output using time alignment (*SIWIS*).
siwis-replace-rf	Local component replacement with random forest output (*SIWIS*).
siwis-add-rf	Addition of random forest output to full contour (*SIWIS*).
siwis-transfer	Transfer of atoms from emphasised speech to neutral speech (*SIWIS*).

The test was designed as a MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor)

test. The subjects were asked to rate the emphasis of the target word for each sentence, where the target word was capitalised. They could use a slider to rate each sample comparatively with others. The slider had 5 indications: “No emphasis”, “Slight emphasis”, “Noticeable emphasis”, “Normal emphasis”, and “Strong emphasis”, in that order. In the *SIWIS* data case, each of the 20 test file had 7 versions, and in the *Roger* case, there were 4 versions of 9 sentences (3 times each of the 3 different patterns). In both cases, the reference was the system with emphasised vocoded speech (**roger-voc** and **siwis-voc-emph**), and the baseline the neutral HMM output (**roger-hmm-neut** and **siwis-hmm-neut**).

The vocoded version of the neutral (**siwis-voc-neut**) and emphasised (**siwis-voc-emph**) natural sentences are the ground truth files which should give the least and most emphasis, respectively. One of the systems, **siwis-transfer**, is the system used in Section 7.3. The atoms generated for **siwis-add-rf** were the output of a random forest trained with 18 estimators, from the strategy using only emphasised words for training (RF2). In this case, the generated atoms were added in the target word in the **siwis-hmm-neut** output F_0. The **siwis-hmm-neut** sentences were generated using automatic time alignment on the neutral version of the sentences. Thus, this method is directly compared with the original neutral vocoded speech, and with the method proposed in the previous section, where atoms from emphasised speech are transferred. The same atoms were used for the **siwis-replace-rf** approach, however in that case the local atoms in the contour synthesised by the **siwis-hmm-emph** system were replaced by the generated atoms. This way, we want to evaluate the replacement of atoms on the adapted HMMs.

In the *Roger* case, we did not have a neutral version of the original speech, so we expect the **roger-hmm-neut** system to produce the least emphasised speech. The reference will be, as for the *SIWIS* case, the vocoded original speech with emphasis, **siwis-voc-emph**. The **roger-hmm-align** method is expected to be perceived as more emphasised than **roger-hmm-neut**, because the duration information of original speech is present. Finally, the **roger-replace-rf** approach is expected to be the closest to the reference, as atoms generated by the random forest will replace local components of the **roger-hmm-emph** output F_0.

The BeaqleJS (Browser based Evaluation of Audio Quality and comparative Listening Environment) framework of Kraft and Zölzer [2014][3] was used to create the web-based listening test.

7.4.3 Results

F_0 Reconstruction

Figure 7.6 shows an example of local component replacement on the *SIWIS* data, on the baseline HMM output F_0.

[3] Available at https://github.com/HSU-ANT/beaqlejs.

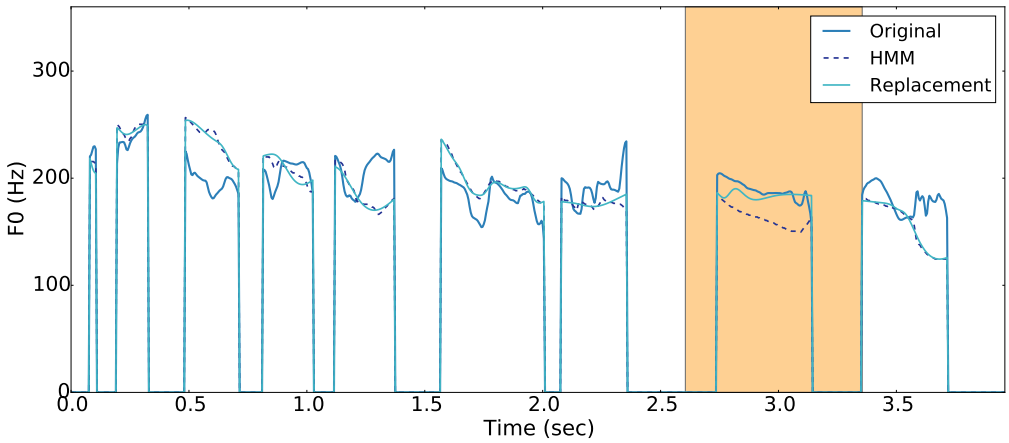

Figure 7.6 – Example of reconstructed F_0 contour after replacing atoms for the sentence *"The Commission has debated the action plan for the next* **five** *years."* The continuous darker blue curve is the original F_0, the dashed and darkest one is the baseline synthetic F_0, and the lightest continuous one is the proposed one (atom replacement). The coloured region delimits the target word.

In this case, we can see that in the target word, the replacement of local atoms by the ones generated by the random forest reduces the difference in F_0 with the real contour in the target word region.

Table 7.8 gives the average RMSE and correlation of three systems at the word level for the emphasised word in each case, and at the utterance level, for the *SIWIS* dataset. RF1 is the system trained using neutral and emphasised data, RF2 is the system trained using only emphasised data. The word-level values are calculated on a reconstruction of the local contour only using local atoms in the word, for both real speech and the three systems.

Table 7.9 gives RMSE and correlation for different number of trees in the random forests, on the *Roger* dataset, at the sentence level, and inside the target word only. We also give the same measures for the F_0 generated by the HMMs, and the same contour after extraction of the model parameters and reconstruction. Here the word-level values are simply calculated by restricting the calculation window to the word boundaries, so they include global component. For all the measures, RMSE and correlation are calculated on voiced frames only, based on the STRAIGHT voicing extracted from the natural reference speech.

Table 7.8 – Average correlation and RMSE at the word level, and utterance level, for SIWIS data.

Measure \ System	HMM	RF1	RF2
RMSE word ($\log F_0$)	0.14	0.12	0.11
RMSE sent (F_0 in Hz)	40	37	37
Correlation word ($\log F_0$)	0.01	0.12	0.12
Correlation sent (F_0)	0.96	0.92	0.92

Table 7.9 – Average correlation and RMSE at the word level, and utterance level, for Roger data, for different sizes of random forest.

Measure \ N_{trees}	HMM	HMM_{recons}	15	18	21	24	27	30
RMSE word (Hz)	16.16	16.17	14.86	14.93	15.0	15.0	14.97	15.0
RMSE sent (Hz)	19.32	45.83	39.34	39.37	39.39	39.38	39.38	39.39
Correlation word	0.898	0.959	0.960	0.960	0.960	0.959	0.959	0.959
Correlation sent	0.894	0.914	0.927	0.926	0.926	0.926	0.926	0.926

SIWIS set, table 7.8 At the word level, the results are showing a low correlation, especially for the baseline system. In that case, we can expect that the way parameters were extracted has an impact on the local decomposition. Because the phrase component was imposed to be the same as in the original log F_0 contour, the algorithm may extract atoms in a different way to compensate the fact that this phrase component is not fitting optimally the synthetic log F_0, e.g. in some cases where the contour is actually lower than the phrase component, negative atoms would be extracted, which may lead to negative correlation for some word-level contours. There is no significant difference between the two other systems for this measure.

The RMSE at the word level is showing similar trend, with similar results for the baseline and the RF models. The RF models show slightly lower RMSE, but with no significant difference. When looking at the whole sentence, the baseline shows worse correlation when using the log F_0, but higher correlation when calculating it only on voiced frames. On the other hand, the RMSE is slightly lower in the RF cases compared to the HMM. The fact that we use a parametric version of the synthetic curve along with the atoms generated by the RF models results in a smoother version, which may allow to reduce some error, hence the lower RMSE. At the same time, it can explain that correlation is a bit higher in the HMM case, because the model may smooth some patterns which should actually be modelled. One thing that should be underlined is that the HMM models have been adapted using emphasised data, and that the synthetic speech sounds generally more pronounced than before emphasis-specific adaptation. However, in the case where we did not use time aligned labels, the duration prediction output extremely slow speech, compared to the neutral model.

Roger set, table 7.9 The results do not seem very informative as all the systems perform very similarly. At the word level, the proposed systems slightly decrease the RMS distance with the reference compared to both HMM output and decomposed-reconstructed HMM output. Similarly, the correlation is increasing while replacing atoms but does not show significant difference among the different systems. At the sentence level, we observe a degradation in the RMS measure, which is increased by the extraction of parameters followed by reconstruction. When replacing local components with the random forest outputs, the error is slightly decreased but remains much higher than for the baseline. This observation is a logical conse-

quence of the reduction of RMSE in a segment of the sentence (the target word). On the other hand, the correlation is increased after extraction-reconstruction of the HMM output, and the replacement of local components further increases the correlation, with no significant difference among the different settings. These results contrast with the trends observed on the *SIWIS* data. This could be explained by two reasons: first, in the *SIWIS* case, the phrase component on the HMM output was imposed for the decomposition, while in the *Roger* case, it was extracted in the standard manner; second, in the *Roger* case, a lot more data was available to train the random forests, and the context was very limited, thus the task was simpler.

It is difficult to conclude from the results as the objective measures do not show significant differences among the proposed systems, although some slight improvement over the baseline is observable. Also, observing only the F_0 contour without looking at the repercussions on the final waveform may not be enough to conclude on the emphasis of words. The listening tests are expected to reflect the perception of emphasis from a human perspective.

Listening Test Results

32 listeners participated in the subjective test, with their majority being non native but fluent English speakers, mostly aged in the range 25–36 years old. The completion of the test took an average of about 25 minutes. Figure 7.7 shows boxplots of the ratings obtained for each approach, for both datasets. Detailed boxplots for each sentence are given in Appendix B, Figures B.1 and B.2.

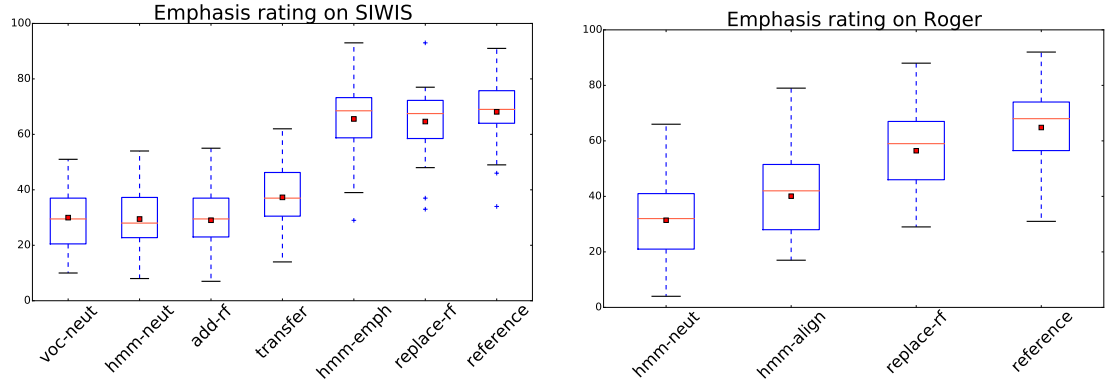

Figure 7.7 – MUSHRA listening test results per dataset. Left: SIWIS data, right: Roger data. The red dots are the means, the read lines are the medians, the boxes indicate

For the *SIWIS* case, we can see two main clusters for the systems. In the first group, we observe that the addition of atoms to the output has little effect on the perception of emphasis on the target word. The transfer of atoms using real atoms, however, seems to slightly increase the perceived strength of emphasis on the target word. The emphasis in the second group, which represents the adapted HMM output, the system with atom replacement and the reference, is perceived as stronger than the first group. To analyse further these observations, a two-tailed

paired t-test on the absolute values given by the listeners was performed on each pair, with a Holm-Bonferroni correction to account for multiple comparisons. All the pairs of systems were significantly different at the level of $p < 0.01$, except the pairs: (voc-neut / hmm-neut), (voc-neut / add-rf), (hmm-neut / add-rf) and (hmm-emph / replace-rf). This implies that in this case, the replacement of atoms in emphasis-adapted HMM output does not change the emphasis perceived on the target word, and both remain signficantly different from the reference. The transfer is significantly different from the first three systems, but it is reasonable to think that this is due to the fact that the added atoms in that case come from real emphasised speech.

The *Roger* boxplot gives more explicit differences between the methods under evaluation. In this case, giving the duration of each phoneme to the synthesiser results in an increase of the perceived emphasis strength, which is further increased by the replacement of local components with the ones generated by the random forest. Finally, the emphasis in the reference file, which is produced naturally, is perceived as the strongest. Again, to validate the results, we conducted a two-tailed paired t-test with a Holm-Bonferroni correction on each system pair. All pairs of systems were significantly different at the level of $p < 0.01$. This confirms the visible differences between each system, and demonstrates that the proposed method significantly improves a simple time aligned based synthesis, in terms of emphasis production. The proposed method is still significantly different from the original emphasised speech.

General Discussion

The objective results on F_0 presented on the task of emphasis synthesis using the GCR model showed that when having enough data, and in similar context, the replacement of local components in the target word F_0 contour helped in reducing the error at the word level, and in increasing the correlation both at the word and sentence level.

The listening tests demonstrated that listeners perceive significantly differently HMM output, HMM output when duration information is provided, HMM output with duration with atom replacement and the original emphasised speech. The replacement of local components significantly increased the perceived strength of emphasis in the *Roger* case, where a lot of training data was available, and needed to be modelled for only a limited context. Adding atoms did not change the perceived emphasis compared with natural and synthetic neutral speech in the *SIWIS* scenario. The replacement of local components did not differ significantly from emphasis-specific adapted HMM output, which was rated close to natural emphasis, although significantly different.

To improve the emphasis generation, distribution parameters could be used instead of values in the tree leaves, if enough data is available to train reasonable models. On the synthesis aspect, the proposed method could be applied with some complementary method, such as duration alteration, intensity modification, or model adaptation, if data is available.

Emphasis Transfer Scenarios Following the same idea as existing emphasis translation system, we propose two possible ways of exploiting the atoms in the input language to affect the prominence of words in the output language. In the first case, one system per language could be used, and the knowledge of which word should be emphasised would enable the synthesis of emphasis through F_0 modification. Another method could consist of using atom information from the input sentence together with linguistic context from both the input language and output language sentences, and try to predict the atoms for the output sentence. This second approach could benefit from approaches such as sequence to sequence learning, introduced by Sutskever et al. [2014] and recently applied to machine translation by Luong et al. [2015]. However, these methods imply having a lot of parallel bilingual data.

7.5 Conclusion

In this chapter, we exploited the capabilities of the GCR model, introduced in previous chapters, in terms of emphasis modelling for speech synthesis.

In analyses of the linguistic relevance of atom parameters, it was shown that these parameters share mutual information with stress, accent and emphasis. With a consistently higher number of components required to model emphasised words, the model also seems to use different combinations of parameters to describe these words.

Two approaches were proposed to simulate emphasis in synthetic speech in an external fashion, i.e. outside of the TTS framework. The first one, demonstrated on natural speech, consisted of using intonation components from natural emphasised speech, and directly adding them to the intonation contour of a neutral word. Listening tests demonstrated that in most cases, this could lead subjects to perceive emphasis on the target word. This method, proposed only in the scenario where both emphasised and neutral versions of the sentence are available, was investigated in the context of emphasis specific word-level intonation modelling. Random forests were employed for the task of intonation component prediction, and failed when these components were only transferred to the target word on synthetic speech. However, when replacing local components of the synthetic intonation contour by the same generated components, we demonstrated that the emphasis strength perception was improved, compared to a simple duration alteration, performing similarly to an HMM-based TTS system adapted using emphasised speech. This validated our hypothesis that local components can be used as word-level intonation to synthesise emphasised word. This indicates that intonation manifests itself with different patterns in emphasised words, rather than existing on top of background prosody.

Overall, the best approach proved to combine well with the exploitation of duration information from actual emphasised speech: when synthesising from neutral HMM-based speech synthesis with given phone durations, replacing local components by synthetic ones generated by random forests significantly increased the perceived perception. This work aims to be further investigated in the context of emphasis translation, for speech-to-speech translation.

8 Conclusions and Future Directions

8.1 Conclusions

In this thesis, we investigated prosody modelling for speech synthesis and emphasis generation. These aspects of speech synthesis are concerned with the lack of nuance in the output speech in a traditional speech-to-speech translation system, introduced by the sequential processing of information, with independent building blocks.

With the confirmation of the need for better prosody modelling in speech synthesis in mind, an initial investigation of regional accents of French was conducted. Through listening tests, we evaluated how native French listeners perceived Swiss regional accents, on synthetic speech adapted with standard methods, and when partial original prosody was provided. Simple adaptation of TTS models using accented speech was shown to be insufficient for the listeners to perceive accentedness. Mixing standard French pronunciation, i.e. segmental level characteristics, with Swiss prosody including rhythm and intonation showed some ability to increase the perceived accentedness, but not to the extent of naturally accented speech. Combining these two methods — standard adaptation, and replacing prosody with natural one — resulted in speech with a degree of accent perceived as not significantly different from real accented speech.

In an attempt at modelling intonation in a theoretically language independent manner, a physiologically plausible model was proposed. The model, inspired by the command-response model, and by muscle response to nerve impulses, can be viewed as a generalised command-response model, in that some of the constraints on the basic components of intonation are relaxed. The model was proposed with an extraction algorithm, which automatically decomposes intonation in components which are assumed to correlate with muscle control of the vocal folds. This algorithm, based on the matching pursuit algorithm, integrates perceptually relevant measures based on the energy and probability of voicing of the speech segments, and should therefore extract perceptually relevant components.

This intonation model was applied to intonation synthesis, using statistical modelling methods. We evaluated the importance of the intonation component positions in the perception of the full modelled intonation, and found that the position of the most prominent components was crucial and needed high precision. Two statistical models were investigated in the task of

model component prediction: support vector machines and deep neural networks. Although these approaches seem able to learn regions where components are likely to exist, their precision is not sufficient to predict a full coherent intonation contour.

In the context of intonation-based emphasis synthesis, two databases were designed and recorded for the *SIWIS* project. Both corpora contain emphasis on specific words. The first one is multi-lingual and with mutliple speakers, while the second one was recorded by a French speaking voice talent and consists of high quality speech.

Linguistic relevance of the atom parameters was found in measuring mutual information between our intonation model parameters and accent, stress and emphasis. A first approach to emphasis replication was proposed by simply extracting intonation components in an emphasised word and adding them to the contour of the same word, initially uttered in a neutral way. Then, using random forests, intonation atoms were predicted from linguistic context, and the knowledge that they were emphasised. Replacing local components of intonation in a neutral sentence by the predicted components proved to elicit emphasis perception, when enough data was available to train the models.

8.2 Perspectives

Along with the work presented throughout the thesis, limitations in some proposed approaches exist, and could be considered as future directions for research.

One of the issues partially addressed in this work concerns the characterisation of the model parameters. While it was shown that atom parameters share mutual information with accent, stress and emphasis, it would be interesting to investigate and understand their relation with other linguistic features.

The prediction of full intonation contours using standard statistical models was revealed to be a challenging task. Other approaches are proposed in the conclusion of Chapter 6, including changing the optimisation criterion, or investigating different types of networks. The way output features are parameterised in the emphasis synthesis scenario, i.e. the structure of the output vector, could also be investigated, for a sentence level prediction.

The work presented on emphasis synthesis, limited to the alteration of intonation, could be extended in multiple ways. One possible line of research consists of combining the method with duration and intensity modification, for a stronger perceptual effect. Moving to cross-lingual transfer is another interesting direction, for which possible approaches are introduced in Chapter 7. They include mapping words in the translation framework and synthesis of local components as proposed in this thesis, or using linguistic context from both languages with components extracted from input speech to predict components of the target language. An interesting line of future work would be to integrate the prediction of the intonation events in the translation, using for instance a neural based approach.

A Atom Parameters in Emphasised Speech

In this appendix, we give some statistics about the French database recorded within the *SIWIS* project. These statistics are the results of analyses on the decomposition of the intonation contour of the **emph** set, and the same sentences in the neutral case.

Figure A.1 shows the distribution for each of the parameters inside the emphasised words, and inside the same words in the neutral version. We can see that in the neutral case, atoms tend to be distributed slightly more often at the beginning of the word, with larger atoms (larger θs) and with higher absolute amplitudes (more atoms in the tails, fewer in the low amplitude region around 0). To try to balance the values when doing a quantisation on the amplitudes, the following mapping was done:

Table A.1 – Amplitude quantisation.

Quantised to	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1
Upper limit	-1.4	-1	-0.7	-0.5	-0.4	-0.33	-0.26	-0.2	-0.12	0.12

Quantised to	1	2	3	4	5	6	7	8	9	10
Upper limit	0.18	0.23	0.29	0.38	0.5	0.7	1	1.4	2	

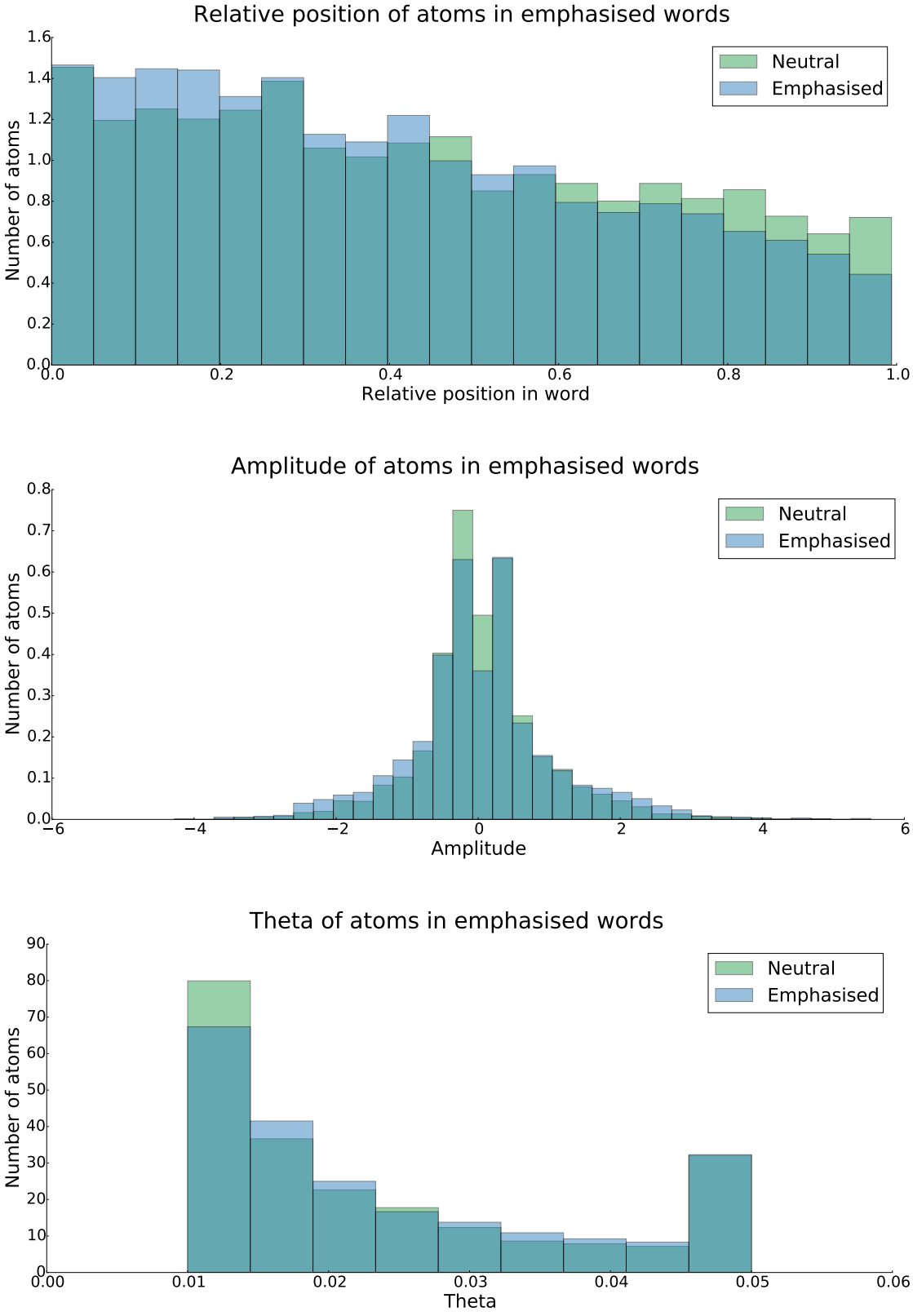

Figure A.1 – Comparison of atom parameters between neutral and emphasised case.

B Emphasis Synthesis Listening Test Results

This appendix provides more detailed result plots of the listening tests conducted in Chapter 7, where the subjects were asked to rate the strength of the emphasis on a target word. Figure B.1 shows the sentence by sentence results for the 4 methods tested on the *Roger* set, while Figure B.2 shows the ratings of the 7 methods used in the *SIWIS* data case. The trends observed on these plots are the same as in the general case, presented in Chapter 7, Section 7.4.

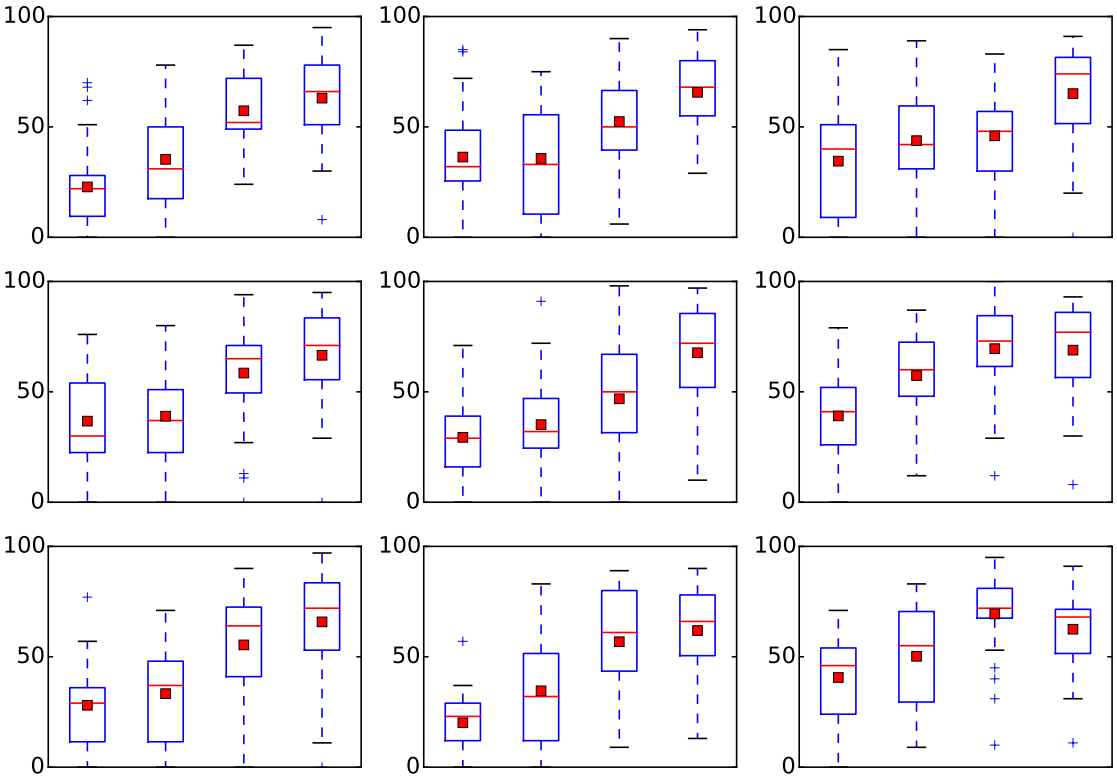

Figure B.1 – MUSHRA results per sentence for Roger data. The systems are from left to right: hmm-neut, hmm-align, replace-rf, reference.

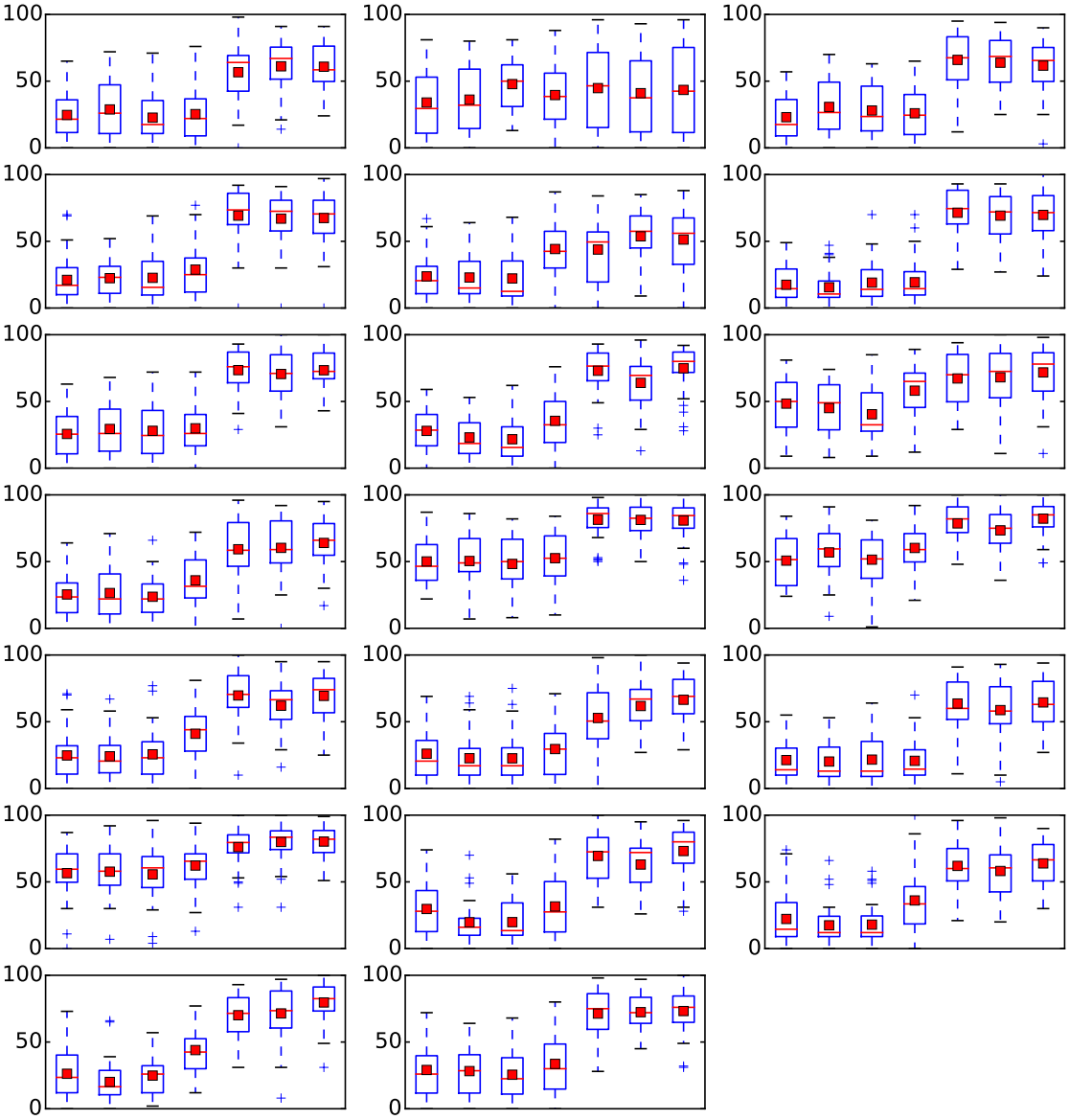

Figure B.2 – MUSHRA results per sentence for SIWIS data. The systems are from left to right: voc-neut, hmm-neut, replace-rf, transfer, hmm-emph, add-rf, reference.

Bibliography

Stefanie Aalburg and Harald Hoege. Foreign-accented speaker-independent speech recognition. In *Proceedings of the International Conference on Spoken Language Processing*, pages 1465–1468, 2004.

Pablo Daniel Agüero and Antonio Bonafonte. Consistent estimation of Fujisaki's intonation model parameters. In *SPECOM*. Citeseer, 2005.

Pablo Daniel Agüero, Klaus Wimmer, and Antonio Bonafonte. Automatic analysis and synthesis of Fujisaki's intonation model for TTS. In *Speech Prosody*, 2004.

Pablo Daniel Agüero, Jordi Adell, and Antonio Bonafonte. Improving TTS quality using pitch contour information of source speaker in S2ST framework. In *Proceedings of the 12th International Workshop "Advances in Speech Technology 2005"*, 2005.

Pablo Daniel Agüero, Jordi Adell, and Antonio Bonafonte. Prosody generation for speech-to-speech translation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 557–560, 2006.

Pablo Daniel Agüero, Juan Carlos Tulli, and Antonio Bonafonte. Pause transfer in the speech-to-speech domain. In *Speech Prosody*, pages 87–90, Campinas, Brazil, 2008.

Helene N Andreassen and Chantal Lyche. Le français du canton de Vaud: une variété autonome. *J. Durand, B. Laks & C. Lyche (éds), Phonologie, variation et accents du français, Paris: Hermès*, pages 63–93, 2009.

Gopala Krishna Anumanchipalli, Luís C. Oliveira, and Alan W. Black. Intent transfer in speech-to-speech machine translation. In *Proceedings of the fourth IEEE Workshop on Spoken Language Technology*, pages 153–158, 2012.

Barry Arons. Pitch-based emphasis detection for segmenting speech recordings. In *ICSLP*, 1994.

Maria Astrinaki, Junichi Yamagishi, Simon King, Nicolas d'Alessandro, and Thierry Dutoit. Reactive accent interpolation through an interactive map application. In *Proceedings of the 8th ISCA Speech Synthesis Workshop*, page 265, Barcelona, Spain, August 2013.

M. Avanzi, G. Christodoulides, Schwab S., Bardiaux A., and Goldman J.-Ph. La variation prosodique régionale et stylistique en français – analyse de neuf points d'enquête PFC. In *Journées PFC*, Paris, 2013.

Mathieu Avanzi. A corpus-based approach to French regional prosodic variation. In *The third Swiss Workshop on Prosody*, Geneva, 2014.

Mathieu Avanzi, Sandra Schwab, Pauline Dubosson, and Jean-Philippe Goldman. Chapitre 5: La prosodie de quelques variétés de français parlées en Suisse romande. *Champs linguistiques*, pages 89–118, 2012.

Gérard Bailly and Ian Gorisch. Generating German intonation with a trainable prosodic model. In *Proceedings of Interspeech*, pages 2366–2369, 2006.

Gérard Bailly and Bleicke Holm. SFC: a trainable prosodic model. *Speech Communication*, 46 (3):348–364, 2005.

Alan Black, Paul Taylor, and Richard Caley. The Festival Speech Synthesis System. Technical report, Human Communication Research Centre, University of Edinburgh, 1997.

Alan W Black and Keiichi Tokuda. The Blizzard challenge 2005: Evaluating corpus-based speech synthesis on common datasets. In *Proceedings of Interspeech*, pages 77–80, Lisbon, Portugal, September 2005.

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

Hervé Bourlard and Christian J Wellekens. Multilayer perceptrons and automatic speech recognition. In *Proc. of the First Intl. Conf. on Neural Networks*, volume 4, pages 407–416, 1987.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Milos Cernak and Pierre-Edouard Honnet. An empirical model of emphatic word detection. In *Proceedings of Interspeech*, pages 573–577, Dresden, Germany, September 2015.

Milos Cernak, Afsaneh Asaei, Pierre-Edouard Honnet, Philip N. Garner, and Hervé Bourlard. Sound pattern matching for automatic prosodic event detection. In *Proceedings of Interspeech*, pages 170–174, September 2016.

Yi-Ning Chen, Yao Qian, and Frank K. Soong. State mapping for cross-language speaker adaptation in TTS. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4273–4276, April 2009.

Robert AJ Clark, Monika Podsiadlo, Mark Fraser, Catherine Mayo, and Simon King. Statistical analysis of the Blizzard challenge 2007 listening test results. In *Proceedings of Blizzard Challenge Workshop*, 2007.

René Collier. Physiological correlates of intonation patterns. *Journal of the Acoustical Society of America*, 58(1):249–255, July 1975.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3): 273–297, 1995.

Marie-Héène Côté. Laurentian French (Québec): extra vowels, missing schwas and surprising liaison consonants. In Randall Gess, Chantal Lyche, and Trudel Meisenburg, editors, *Phonological variation in French: illustrations from three continents*. John Benjamins, Amsterdam, 2012.

Christophe d'Alessandro, Albert Rilliard, and Sylvain Le Beux. Chironomic stylization of intonation. *Journal of the Acoustical Society of America*, 129(3):1594–1604, March 2011.

Tijana Delić, Branislav Gerazov, Branislav Popović, and Milan Sečujski. A linguistic interpretation of the atom decomposition of fundamental frequency contour for American English. In *International Conference on Speech and Computer*, pages 59–66, Budapest, Hungary, 2016. Springer.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

Quoc Truong Do, Shinnosuke Takamichi, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. Preserving word-level emphasis in speech-to-speech translation using linear regression HSMMs. In *Proceedings of Interspeech*, pages 3665–3669, Dresden, Germany, September 2015a.

Quoc Truong Do, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. Improving translation of emphasis with pause prediction in speech-to-speech translation systems. In *International Workshop on Spoken Language Translation*, Seattle, USA, December 2015b.

Quoc Truong Do, Sakriani Sakti, Graham Neubig, and Satoshi Nakamura. Transferring emphasis in speech translation using hard-attentional neural network models. In *Proceedings of Interspeech*, pages 2533–2537, San Francisco, CA, USA, September 2016a.

Quoc Truong Do, Tomoki Toda, Graham Neubig, Sakriani Sakti, and Satoshi Nakamura. A hybrid system for continuous word-level emphasis modeling based on HMM state clustering and adaptive training. In *Proceedings of Interspeech*, pages 3196–3200, 2016b.

Jacques Durand, Bernard Laks, and Chantal Lyche. *Phonologie, variation et accents du français*. Paris, Hermès, 2009.

Rong-En Fan, Pai-Hsuen Chen, and Chih-Jen Lin. Working set selection using second order information for training support vector machines. *Journal of Machine Learning Research*, 6: 1889–1918, 2005.

Yuchen Fan, Frank K. Yao Qian, Soong, and Lei He. Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4475–4479, Brisbane, Autralia, April 2015. IEEE.

Hiroya Fujisaki. Information, prosody, and modeling—with emphasis on tonal features of speech. In *Speech Prosody 2004, International Conference*, 2004.

Hiroya Fujisaki. The roles of physiology, physics and mathematics in modeling prosodic features of speech. In *Speech Prosody*, Dresden, Germany, May 2006.

Hiroya Fujisaki and Keikichi Hirose. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan (E)*, 5(4):233–242, 1984.

Hiroya Fujisaki and Shigeo Nagashima. A model for the synthesis of pitch contours of connected speech. Technical report, Engineering Research Institute, University of Tokyo, 1969.

Hiroya Fujisaki, Mats Ljungqvist, and Hiroshi Murata. Analysis and modeling of word accent and sentence intonation in Swedish. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 211–214. IEEE, 1993.

Hiroya Fujisaki, Sumio Ohno, and Changfu Wang. A command-response model for F_0 contour generation in multilingual speech synthesis. In *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, 1998.

Hiroya Fujisaki, Ryou Tomana, Shuichi Narusawa, Sumio Ohno, and Changfu Wang. Physiological mechanisms for fundamental frequency control in standard Chinese. In *Proceedings of the International Conference on Spoken Language Processing*, pages 9–12, 2000.

Philip N. Garner, Milos Cernak, and Petr Motlicek. A simple continuous pitch estimation algorithm. *IEEE Signal Processing Letters*, 20(1):102–105, 2013.

Branislav Gerazov, Pierre-Edouard Honnet, Aleksandar Gjoreski, and Philip N. Garner. Weighted correlation based atom decomposition intonation modelling. In *Proceedings of Interspeech*, pages 1601–1605, Dresden, Germany, September 2015.

Branislav Gerazov, Aleksandar Gjoreski, Aleksandar Melov, Pierre-Edouard Honnet, Zoran Ivanovski, and Philip N. Garner. Unified prosody model based on atom decomposition for emphasis detection. In *Proceedings of ETAI*, Struga, Macedonia, 2016.

Pegah Ghahremani, Bagher BabaAli, Daniel Povey, Korbinian Riedhammer, Jan Trmal, and Sanjeev Khudanpur. A pitch extraction algorithm tuned for automatic speech recognition. In

Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pages 2513–2517. IEEE, 2014.

Matthew Gibson, Teemu Hirsimäki, Reima Karhila, Mikko Kurimo, and William Byrne. Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis using two-pass decision tree construction. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4642–4645, 2010.

Jean-Philippe Goldman, Pierre-Edouard Honnet, Rob Clark, Philip N. Garner, Maria Ivanova, Alexandros Lazaridis, Hui Liang, Tiago Macedo, Beat Pfister, Manuel Sam Ribeiro, Eric Wehrli, and Junichi Yamagishi. The SIWIS database: a multilingual speech database with acted emphasis. In *Proceedings of Interspeech*, pages 1532–1535, San Francisco, CA, USA, September 2016.

Ricardo Gutierrez-Osuna and Daniel Felps. Foreign accent conversion through voice morphing. Technical report, Department of Computer Science and Engineering, Texas A&M University, 2010.

Abualsoud Hanani, Martin J Russell, and Michael J Carey. Human and computer recognition of regional accents and ethnic groups from British English speech. *Computer Speech & Language*, 27(1):59–74, 2013.

Hiroya Hashimoto, Keikichi Hirose, and Nobuaki Minematsu. Improved automatic extraction of generation process model commands and its use for generating fundamental frequency contours for training HMM-based speech synthesis. In *Proceedings of Interspeech*, 2012.

Xiaodong He and Yunxin Zhao. Fast model selection based speaker adaptation for nonnative speech. *IEEE Transactions on Speech and Audio Processing*, 11(4):298–307, 2003.

Mattias Heldner, Eva Strangert, and Thierry Deschamps. A focus detector using overall intensity and high frequency emphasis. In *Proc. of ICPhS*, volume 99, pages 1491–1494, 1999.

Dik J. Hermes. Measurement of pitch by subharmonic summation. *Journal of the Acoustical Society of America*, 83(1):257–264, 1988.

Dik J. Hermes. Measuring the perceptual similarity of pitch contours. *Journal of Speech, Language, and Hearing Research*, 41(1):73–82, February 1998.

Wolfgang J Hess, Klaus J Kohler, and Hans-Günther Tillmann. The Phondat-verbmobil speech corpus. In *Proceedings of EUROSPEECH*, 1995.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012 2012.

Keikichi Hirose and Hiroya Fujisaki. Analysis and synthesis of voice fundamental frequency contours of spoken sentences. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 7, pages 950–953. IEEE, 1982.

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

Keikichi Hirose, Keiko Ochi, Ryusuke Mihara, Hiroya Hashimoto, Daisuke Saito, and Nobuaki Minematsu. Adaptation of prosody in speech synthesis by changing command values of the generation process model of fundamental frequency. In *Proceedings of Interspeech*, pages 2793–2796, Florence, August 2011.

Keikichi Hirose, Hiroya Hashimoto, Jun Ikeshima, and Nobuaki Minematsu. Fundamental frequency contour reshaping in HMM-based speech synthesis and realization of prosodic focus using generation process model. In *Speech Prosody*, May 2012.

Daniel Hirst and Albert Di Cristo. *Intonation systems: a survey of twenty languages*. Cambridge University Press, 1998.

Daniel Hirst and Robert Espesser. Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut Phonétique d'Aix*, pages 75–85, 1993.

Daniel Hirst, Albert Di Cristo, and Robert Espesser. Levels of representation and levels of analysis for the description of intonation systems. In *Prosody: Theory and experiment*, pages 51–87. Springer, 2000.

Pierre-Edouard Honnet and Philip N. Garner. Importance of prosody in Swiss French accent for speech synthesis. In *Nouveaux cahiers de linguistique française*, September 2014.

Pierre-Edouard Honnet and Philip N. Garner. Emphasis recreation for TTS using intonation atoms. In *Proceedings of the 9th ISCA Speech Synthesis Workshop*, pages 14–20, Sunnyvale, CA, USA, September 2016a.

Pierre-Edouard Honnet and Philip N. Garner. Intonation atom-based emphasis transfer. Idiap-RR Idiap-RR-14-2016, Idiap, 5 2016b.

Pierre-Edouard Honnet, Alexandros Lazaridis, Jean-Philippe Goldman, and Philip N. Garner. Prosody in Swiss French accents: Investigation using analysis by synthesis. In *Speech Prosody*, Dublin, Ireland, May 2014.

Pierre-Edouard Honnet, Branislav Gerazov, and Philip N. Garner. Atom decomposition-based intonation modelling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4744–4748, Brisbane, Australia, April 2015. IEEE.

Qiong Hu, Korin Richmond, Junichi Yamagishi, and Javier Latorre. An experimental comparison of multiple vocoder types. In *Proceedings of the 8th ISCA Speech Synthesis Workshop*, pages 135–140, Barcelona, Spain, 2013.

C. Huang, T. Chen, S. Li, E. Chang, and J. Zhou. Analysis of speaker variability. In *Proceedings of Eurospeech*, pages 1377–1380, Aalborg, Denmark, 2001.

Rongqing Huang, John HL Hansen, and Pongtep Angkititrakul. Dialect/accent classification using unrestricted audio. *IEEE Transactions on Audio, Speech and Language Processing*, 15 (2):453–464, 2007.

Andrew J Hunt and Alan W Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 373–376. IEEE, 1996.

Hirokazu Kameoka, Jonathan Le Roux, and Yasunori Ohishi. A statistical model of speech F0 contours. In *Proceedings ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition (SAPA)*, pages 43–48, September 2010.

Hirokazu Kameoka, Kota Yoshizato, Tatsuma Ishihara, Kento Kadowaki, Yasunori Ohishi, and Kunio Kashino. Generative modeling of voice fundamental frequency contours. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(6):1043–1052, June 2015.

Vasilis Karaiskos, Simon King, Robert AJ Clark, and Catherine Mayo, editors. *The Blizzard Challenge 2008*, 2008.

Liu Wai Kat and Pascale Fung. Fast accent identification and accented speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 221–224. IEEE, 1999.

Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain de Cheveigné. Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, 27(3):187–207, 1999.

Lyndon S Kennedy and Daniel PW Ellis. Pitch-based emphasis detection for characterization of meeting recordings. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 243–248. IEEE, 2003.

Simon King and Vasilis Karaiskos, editors. *The Blizzard Challenge 2011*, 2011.

Werner M Kistler, Wulfram Gerstner, and J Leo van Hemmen. Reduction of the Hodgkin-Huxley equations to a single-variable threshold model. *Neural Computation*, 9(5):1015–1045, 1997.

Pierre Knecht. Le français en Suisse romande: aspects linguistiques et sociolinguistiques. In *Le français hors de France*, pages 249–258. Valdman, A., Paris, 1979.

Greg Kochanski and Chilin Shih. Stem-ML: language-independent prosody description. In *Proceedings of Interspeech*, pages 239–242, 2000.

Greg Kochanski and Chilin Shih. Prosody modeling with soft templates. *Speech Communication*, 39(3):311–352, 2003.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86, 2005.

John Kominek and Alan W Black. The CMU Arctic speech databases. In *Fifth ISCA Workshop on Speech Synthesis*, 2004.

Sebastian Kraft and Udo Zölzer. BeaqleJS: HTML5 and JavaScript based framework for the subjective evaluation of audio quality. In *Linux Audio Conference*, Karlsruhe, Germany, 2014.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

Sacha Krstulovic̀ and Rémi Gribonval. MPTK: Matching pursuit made tractable. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 3, pages 496–499. IEEE, 2006.

Mikko Kurimo, William Byrne, John Dines, Philip N. Garner, Matthew Gibson, Yong Guan, Teemu Hirsimäki, Reima Karhila, Simon King, Hui Liang, Keiichiro Oura, Lakshmi Saheer, Matt Shannon, Sayaka Shiota, Jilei Tian, Keiichi Tokuda, Mirjam Wester, Yi-Jian Wu, and Junichi Yamagishi. Personalising speech-to-speech translation in the EMIME project. In *Proceedings of the ACL 2010 System Demonstrations*, pages 48–53, Uppsala, July 2010.

Lori F Lamel, Jean-Luc Gauvain, and Maxine Eskenazi. BREF, a large vocabulary spoken corpus for French. In *Proceedings of EUROSPEECH*, pages 505–508, 1991.

Javier Latorre, Mark J.F. Gales, Sabine Buchholz, Kate Knill, Masatsune Tamura, Yamato Ohtani, and Masami Akamine. Continuous F0 in the source-excitation generation for HMM-based TTS: do we need voiced/unvoiced classification? In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4724–4727, 2011.

Alexandros Lazaridis, Elie Khoury, Jean-Philippe Goldman, Mathieu Avanzi, Sébastien Marcel, and Philip N. Garner. Swiss French regional accent identification. In *Odyssey: The Speaker and Language Recognition Workshop*, 2014a.

Alexandros Lazaridis, Jean-Philippe Goldman, Mathieu Avanzi, and Philip N. Garner. Syllable-based regional Swiss French accent identification using prosodic features. In *Nouveaux cahiers de linguistique française*, 2014b.

Hui Liang. *Data-driven Enhancement of State Mapping-based Cross-lingual Speaker Adaptation*. PhD thesis, École Polytechnique Fédérale de Lausanne, 2012.

Hui Liang. Detecting emphasised spoken words by considering them prosodic outliers and taking advantage of HMM-based TTS framework. In *Speech Prosody Conference*, pages 69–73, Boston, USA, 2016.

Hui Liang and John Dines. Phonological knowledge guided HMM state mapping for cross-lingual speaker adaptation. In *Proceedings of Interspeech*, pages 1825–1828, 2011.

Wai Kat Liu and Pascale N Fung. MLLR-based accent model adaptation without accented data. In *Proceedings of the International Conference on Spoken Language Processing*, volume 3, pages 738–741, 2000.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*, 2015.

Stéphane G. Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on*, 41(12):3397–3415, 1993.

André Martinet. *La prononciation du français contemporain*. Droz, Geneva, Switzerland, 1971.

Jean-Pierre Métral. Le vocalisme du français en Suisse romande. considérations phonologiques. *Cahiers Ferdinand de Saussure*, (31):145–176, 1977.

Jessica Sertling Miller. *Swiss French prosody: intonation, rate, and speaking style in the Vaud canton*. PhD thesis, Graduate College of the University of Illinois, Urbana-Champaign, 2007.

Hansjörg Mixdorff. A novel approach to the fully automatic extraction of Fujisaki model parameters. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 3, pages 1281–1284, Istanbul, Turkey, 2000.

Yves Charles Morin. Le français de référence et les normes de prononciation. *Cahiers de l'Institut de linguistique de Louvain*, 26(1):91–135, 2000.

Hideharu Nakajima, Hideyuki Mizuno, and Sumitaka Sakauchi. Emphasized accent phrase prediction from text for advertisement text-to-speech synthesis. In *The 28th Pacific Asia Conference on Language, Information and Computation*, Phuket, Thailand, December 2014.

Shuichi Narusawa, Nobuaki Minematsu, Keikichi Hirose, and Hiroya Fujisaki. A method for automatic extraction of model parameters from fundamental frequency contours of speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 509–512, 2002.

Keiko Ochi, Keikichi Hirose, and Nobuaki Minematsu. Control of prosodic focus in corpus-based generation of fundamental frequency contours of Japanese based on the generation process model. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4257–4260. IEEE, 2009.

Sven Öhman. *Word and sentence intonation: A quantitative model*. Speech Transmission Laboratory, Department of Speech Communication, Royal Institute of Technology, 1967.

Mohamed Kamal Omar and Jason Pelecanos. A novel approach to detecting non-native speakers and their native language. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4398–4401. IEEE, 2010.

Keiichiro Oura, Keiichi Tokuda, Junichi Yamagishi, Simon King, and Mirjam Wester. Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4594–4597, March 2010.

Douglas B. Paul and Janet M. Baker. The design for the Wall Street Journal-based CSR corpus. In *Proceedings of the Workshop on Speech and Natural Language*, pages 357–362, Stroudsburg, PA, USA, 1992.

Xianglin Peng, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda. Cross-lingual speaker adaptation for HMM-based speech synthesis considering differences between language-dependent average voices. In *Proc. of ICSP*, pages 605–608, October 2010.

Janet Pierrehumbert. Synthesizing intonation. *Journal of the Acoustical Society of America*, 70: 985–995, 1981.

Réjean Plamondon. A kinematic theory of rapid human movements: Part I: Movement representation and generation. *Biological Cybernetics*, 72(4):295–307, March 1995.

Santitham Prom-on, Yi Xu, and Bundit Thipakorn. Modeling tone and intonation in Mandarin and English as a process of target approximation. *Journal of the Acoustical Society of America*, 125:405–424, January 2009.

Lawrence R Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

I. Racine, S. Schwab, and S. Detey. Accent(s) suisse(s) ou standard(s) suisse(s) ? Approche perceptive dans quatre régions de Suisse romande. In A. Falkert, editor, *La perception des accents du français hors de France.*, pages 41–59. 2013.

Manuel Sam Ribeiro and Robert A. J. Clark. A multi-level representation of F_0 using the continuous wavelet transform and discrete cosine transform. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4909–4913, Brisbane, Australia, April 2015. IEEE.

Manuel Sam Ribeiro, Oliver Watts, Junichi Yamagishi, and Robert AJ Clark. Wavelet-based decomposition of F_0 as a secondary task for DNN-based speech synthesis with multi-task learning. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5525–5529. IEEE, 2016a.

Manuel Sam Ribeiro, Oliver Watts, and Junichi Yamagishi. Parallel and cascaded deep neural networks for text-to-speech synthesis. In *Proceedings of the 9th ISCA Speech Synthesis Workshop*, pages 119–124, Sunnyvale, CA, USA, September 2016b.

Albert Rilliard, Alexandre Allauzen, and Philippe Boula de Mareüil. Using Dynamic Time Warping to compute prosodic similarity measures. In *Proceedings of Interspeech*, pages 2021–2024, Florence, Italy, August 2011.

Arup Saha, Tulika Basu, Anal Haque Warsi, Keikichi Hirose, and Hiroya Fujisaki. Subjective evaluation of joint modeling of pause insertion and F_0 contour generation in text-to-speech synthesis of Bangla. In *Oriental COCOSDA*, pages 6–9, 2011.

Sandra Schwab and Isabelle Racine. Le débit lent des suisses romands: mythe ou réalité? *Journal of French Language Studies*, pages 281–295, 2013.

Sandra Schwab, Mathieu Avanzi, Jean-Philippe Goldman, Pascal Montchaud, Isabelle Racine, et al. An acoustic study of penultimate accentuation in three varieties of French. In *Proceedings of Speech Prosody*, 2012.

Kim EA Silverman, Mary E Beckman, John F Pitrelli, Mari Ostendorf, Colin W Wightman, Patti Price, Janet B Pierrehumbert, and Julia Hirschberg. TOBI: a standard for labeling English prosody. In *ICSLP*, volume 2, pages 867–870, Banff, October 1992.

Frank Soong and B Juang. Line spectrum pair (LSP) and speech data compression. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 9, pages 37–40. IEEE, 1984.

Adriana Stan, Yoshitaka Mamiya, Junichi Yamagishi, Peter Bell, Oliver Watts, Robert Clark, and Simon King. ALISA: An automatic lightly supervised speech segmentation and alignment tool. *Computer Speech & Language*, 35:116–133, 2016.

Helmer Strik. *Physiological control and behaviour of the voice source in the production of prosody*. PhD thesis, Dept. of Language and Speech, Univ. of Nijmegen, Nijmegen, Netherlands, October 1994.

Volker Strom, Robert AJ, and Simon King. Expressive prosody for unit-selection speech synthesis. In *Proceedings of Interspeech*, Pittsburgh,PA,USA, 2006.

Volker Strom, Ani Nenkova, Robert AJ Clark, Yolanda Vazquez-Alvarez, Jason Brenier, Simon King, and Daniel Jurafsky. Modelling prominence and emphasis improves unit-selection synthesis. In *Proceedings of Interspeech*, 2007.

Antti Suni, Daniel Aalto, Tuomo Raitio, Paavo Alku, and Martti Vainio. Wavelets for intonation modeling in HMM speech synthesis. In *8th ISCA Workshop on Speech Synthesis*, pages 305–310, Barcelona, Spain, August 2013.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112, 2014.

György Szaszák, Máté Ákos Tündik, Branislav Gerazov, and Aleksandar Gjoreski. Combining atom decomposition of the F_0 track and HMM-based phonological phrase modelling for robust stress detection in speech. In *International Conference on Speech and Computer*, pages 165–173, Budapest, Hungary, 2016. Springer.

Shinji Takaki, Junichi Yamagishi, and Zhenzhou Wu. A function-wise pre-training technique for constructing a deep neural network based spectral model in statistical parametric speech synthesis. In *First International Workshop on Machine Learning in Spoken Language Processing*, 2015.

Paul Taylor. Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America*, 107:1697–1714, March 2000.

Carlos Teixeira, Isabel Trancoso, and António Serralheiro. Accent identification. In *Proceedings of the International Conference on Spoken Language Processing*, volume 3, pages 1784–1787. IEEE, 1996.

Ingo R. Titze and Daniel W. Martin. Principles of voice production. *Journal of the Acoustical Society of America*, 104(3), 1998.

Tomoki Toda and Keiichi Tokuda. Speech parameter generation algorithm considering global variance for HMM-based speech synthesis. In *Proceedings of Interspeech*, pages 2800–2804, 2005.

Keiichi Tokuda, Takao Kobayashi, Takashi Masuko, and Satoshi Imai. Mel-generalized cepstral analysis-a unified approach to speech spectral estimation. In *Proceedings of the International Conference on Spoken Language Processing*, 1994.

Keiichi Tokuda, Takashi Masuko, Noboru Miyazaki, and Takao Kobayashi. Multi-space probability distribution HMM. *IEICE TRANSACTIONS on Information and Systems*, 85(3):455–464, 2002a.

Keiichi Tokuda, Heiga Zen, and Alan W Black. An HMM-based speech synthesis system applied to English. In *Proc. of 2002 IEEE SSW*, pages 227–230. IEEE, 2002b.

Humberto Torres and Jorge Gurlekian. Novel estimation method for the superpositional intonation model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24 (1):151–160, 2016.

Andreas Tsiartas, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. A study on the effect of prosodic emphasis transfer on overall speech translation quality. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, Canada, 2013. IEEE.

Aaron 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. *arXiv preprint arXiv:1609.03499*, 2016.

Karel Veselỳ, Arnab Ghoshal, Lukás Burget, and Daniel Povey. Sequence-discriminative training of deep neural networks. In *Proceedings of Interspeech*, pages 2345–2349, 2013.

Mirjam Wester. The EMIME bilingual database. Technical report, The University of Edinburgh, 2010.

Mirjam Wester, Cassia Valentini-Botinhao, and Gustav Eje Henter. Are we using enough listeners? No! An empirically-supported critique of Interspeech 2014 TTS evaluations. In *Proceedings of Interspeech*, pages 3476–3480, Dresden, Germany, 2015.

Reiner Wilhelms-Tricarico, Brian Mottershead, Rattima Nitisaroj, Michael Baumgartner, John Reichenbach, and Gary Marple. The Lessac Technologies system for Blizzard challenge 2011. In *Blizzard Challenge 2011*, 2011.

Yi-Jian Wu, Simon King, and Keiichi Tokuda. Cross-lingual speaker adaptation for HMM-based speech synthesis. In *Chinese Spoken Language Processing, 2008. ISCSLP'08. 6th International Symposium on*, pages 1–4, 2008.

Yi-Jian Wu, Yoshihiko Nankaku, and Keiichi Tokuda. State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis. In *Proceedings of Interspeech*, pages 528–531, 2009.

Junichi Yamagishi. *Average-voice-based speech synthesis*. PhD thesis, Tokyo Institute of Technology, March 2006a.

Junichi Yamagishi. An introduction to HMM-based speech synthesis. Technical report, Tokyo Institute of Technology, 2006b.

Junichi Yamagishi and Takao Kobayashi. Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training. *IEICE Transactions on Information and Systems*, 90(2):533–543, 2007.

Junichi Yamagishi, Takao Kobayashi, Yuji Nakano, Katsumi Ogata, and Juri Isogai. Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(1): 66–83, 2009.

Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proceedings of EUROSPEECH*, pages 2347–2350, 1999.

Takenori Yoshimura, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda. Cross-lingual speaker adaptation based on factor analysis using bilingual speech data for HMM-based speech synthesis. In *Proceedings of the 8th ISCA Speech Synthesis Workshop*, Barcelona, September 2013.

Kota Yoshizato, Hirokazu Kameoka, Daisuke Saito, and Shigeki Sagayama. Statistical approach to Fujisaki-model parameter estimation from speech signals and its quantitative evaluation. In *Speech Prosody*, pages 175–178, 2012.

Steve J Young, Julian J Odell, and Philip C Woodland. Tree-based state tying for high accuracy acoustic modelling. In *Proceedings of the workshop on Human Language Technology*, pages 307–312. Association for Computational Linguistics, 1994.

Kai Yu and Steve Young. Continuous F0 modeling for HMM based statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech and Language Processing*, 19(5):1071–1079, July 2011.

Kai Yu, François Mairesse, and Steve Young. Word-level emphasis modelling in HMM-based speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4238–4241, 2010.

Heiga Zen and Haşim Sak. Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4470–4474, Brisbane, Australia, April 2015. IEEE.

Heiga Zen and Andrew Senior. Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3872–3876. IEEE, 2014.

Heiga Zen, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan Black, and Keiichi Tokuda. The HMM-based speech synthesis system (HTS) version 2.0. In *Proceedings of the 6th ISCA Speech Synthesis Workshop*, pages 294–299, 2007.

Heiga Zen, Keiichi Tokuda, and Alan W Black. Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064, 2009.

Heiga Zen, Andrew Senior, and Mike Schuster. Statistical parametric speech synthesis using deep neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7962–7966. IEEE, 2013.

Bufan Zhang, Zhenhua Ling, Long Qin, and Renhua Wang. Applying SFC model for Chinese expressive speech synthesis. In *Proceedings of International Symposium on Chinese Spoken Language Processing*, 2006.

Pierre-Edouard Honnet

PhD *Student at* EPFL,
Research Assistant at Idiap

+41 (0) 79 367 64 30
pierre-edouard.honnet@grenoble-inp.org
http://idiap.ch/~pehonnet

Summary

I am an electrical engineer specialised in signal processing, currently pursuing PhD studies in speech processing, involved in a personalised speech-to-speech translation project. I am interested in audio and speech processing, and more generally in new technologies.

Education

Nov. 2012 – Present **Doctoral Student in Electrical Engineering**, *Ecole Polytechnique Federale de Lausanne*, Switzerland.

Sep. 2009 – Jun. 2012 **MSc. in Electrical Engineering**, *Grenoble Institute of Technology (INP)*, France.

Sep. 2007 – Jun. 2009 **Classes Préparatoires aux Grandes Écoles**, *Lycée Henri Poincaré, Nancy*, France.

Work Experience

Nov. 2012 – Present **Research Assistant in Speech & Audio Processing Group**, *Idiap Research Institute*, Martigny, Switzerland.
Supervisors: Prof. Hervé Bourlard, Dr. Philip N. Garner

- Investigation of Swiss French accents using prosodic variations
- Developing a new intonation model
- Intonation prediction for speech synthesis
- Database design for a high quality French speaker dependent corpus
- Intonation-based emphasis transfer

Mar. – Jun. 2015 **Research Intern in SMG group**, *National Institute of Informatics*, Tokyo, Japan.
Supervisor: Dr. Junichi Yamagishi

- Evaluating perception of a previously developed intonation model
- Statistical modelling for intonation synthesis using state of the art methods

Feb. – Jul. 2012 **Research Student in Speech Communication Laboratory**, *Trinity College of Dublin*, Dublin, Ireland.
Supervisors: Prof. Nick Campbell, Dr. Céline De Looze, Dr. Thomas Hueber

- Investigating prosodic variations at topic change boundaries
- Building HMM-based speech synthesis systems
- Using speaker adaptation to build a speech synthesis system with topic changes

Jun. – Aug. 2011 **Software Development Engineer**, *Overkiz*, Archamps, France.
Supervisor: Boris Brezillon

- Investigating a home automation protocol and the existing software architecture
- Designing a solution to integrate the new protocol in the system
- Implementing the protocol in C++

Languages

French	Native
English	Fluent
German	Basic knowledge

Computer skills

Programming	C/C++, Java, Python, Bash, Z-shell
Tools	LaTeX, Git, Matlab, HTS (HTK), SPTK, MPTK
OS	Linux, Windows

Research Interests

Speech & Audio Processing

Speech Technologies

Human-Machine Interactions

Machine Learning

Activity and Interests

- Sports: all racket sports, skiing, snowboarding, hiking, unihockey.
- Watching movies, reading

Publications

- Pierre-Edouard Honnet and Philip N. Garner. *Emphasis Recreation for TTS Using Intonation Atoms.* In Proceedings of the 9th ISCA Speech Synthesis Workshop, pages 14–20, Sunnyvale, CA, USA, September 2016.
- Alexandros Lazaridis, Milos Cernak, Pierre-Edouard Honnet, and Philip N. Garner. *Investigating spectral amplitude modulation phase hierarchy features in speech synthesis.* In Proceedings of the 9th ISCA Speech Synthesis Workshop, pages 32–37, Sunnyvale, CA, USA, September 2016.
- Jean-Philippe Goldman, Pierre-Edouard Honnet, Rob Clark, Philip N. Garner, Maria Ivanova, Alexandros Lazaridis, Hui Liang, Tiago Macedo, Beat Pfister, Manuel Sam Ribeiro, Eric Wehrli, and Junichi Yamagishi. *The SIWIS Database: a Multilingual Speech Database With Acted Emphasis.* In Proceedings of Interspeech, pages 1532–1535, San Francisco, CA, USA, September 2016.
- Milos Cernak, Afsaneh Asaei, Pierre-Edouard Honnet, Philip N. Garner, and Hervé Bourlard. *Sound pattern matching for automatic prosodic event detection.* In Proceedings of Interspeech, pages 170–174, September 2016.
- Branislav Gerazov, Aleksandar Gjoreski, Aleksandar Melov, Pierre-Edouard Honnet, Zoran Ivanovski and Philip N. Garner. *Unified prosody model based on atom decomposition for emphasis detection.* In Proceedings of ETAI, Ohrid, Macedonia, 2016.
- Branislav Gerazov, Pierre-Edouard Honnet, Aleksandar Gjoreski, and Philip N. Garner. *Weighted Correlation-based Atom Decomposition Intonation Modelling.* In Proceedings of Interspeech, pages 1601–1605, Dresden, Germany, September 2015.
- Milos Cernak and Pierre-Edouard Honnet. *An Empirical Model of Emphatic Word Detection.* In Proceedings of Interspeech, pages 573–577, Dresden, Germany, September 2015.
- Pierre-Edouard Honnet, Branislav Gerazov, and Philip N. Garner. *Atom Decomposition-*

based Intonation Modelling. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pages 4744–4748, Brisbane, Australia, April 2015

- Pierre-Edouard Honnet and Philip N. Garner. *Importance of Prosody in Swiss French Accent for Speech Synthesis.* In Nouveaux cahiers de linguistique française, September 2014.
- Pierre-Edouard Honnet, Alexandros Lazaridis, Jean-Philippe Goldman, and Philip N. Garner. *Prosody in Swiss French Accents: Investigation Using Analysis by Synthesis.* In Speech Prosody, Dublin, Ireland, May 2014.
- Alexandros Lazaridis, Pierre-Edouard Honnet, and Philip N. Garner. *SVR vs MLP for phone duration modelling in HMM-based speech synthesis.* In Speech Prosody, Dublin, Ireland, May 2014.