Intra-Lingual and Cross-Lingual Prosody Modelling

Gopala Krishna Anumanchipalli

CMU-LTI-13-009

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Departamento de Engenharia
Electrotécnica e de Computadores
Instituto Superior Técnico
Portugal

Thesis Committee:
Alan W Black (Chair), LTI, CMU
Luís C. Oliveira (Chair), INESC-ID, IST
Justine Cassell, HCII, CMU
Mário Figueiredo, IT, IST
Bhiksha Raj, LTI, CMU
Isabel Trancoso, INESC-ID, IST
Paul Taylor, Google Inc.

*Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.*

Keywords: Text-to-speech synthesis, Statistical Parametric Speech Synthesis, Intonation Modelling, Prosody, Fujisaki model, Intonational Phonology, Audiobook Synthesis, Voice Conversion, Cross-lingual intent transfer, Speech-to-Speech Machine Translation.

Palavras-chave: síntese de fala a partir de texto (TTS), síntese de fala com parâmetros estatísticos, modelo de entonação, Prosódia, modelo Fujisaki, fonologia da entonação, síntese de livros falados, conversão de voz, conversao de intenção entre línguas, tradução automática de fala para fala.

Abstract

Statistical Parametric Speech Synthesis (SPSS) offers flexibility and computational advantage compared to other methods for Text-to-Speech Synthesis. While the speech output is intelligible, statistically trained voices are less natural due to the amount of signal processing and statistical averaging that goes into building the models. Much of the blame for the lack of naturalness falls on the inappropriate and monotonous prosody in synthesized speech. The voice source, which directly effects the prosody, is a complementary source of information than the vocal tract and has its own patterns that need to be dealt with appropriately. Under this hypothesis, this thesis investigates the representations and optimal strategies for prosody modelling within the SPSS paradigm.

We propose the Statistical Phrase/Accent Model (SPAM) of intonation as a framework that is both (i) a computational model with associated training and synthesis methods for prosody and (ii) has strong theoretical basis for prosodic description. The SPAM framework combines the strengths of existing complementary views of intonation like Autosegmental Metrical Phonology, Production paradigms like the Fujisaki Model and purely computational approaches like the TILT model. We demonstrate Accent Groups, a new data derived phonological unit, as the optimal representational level to model Pitch accents and integrate it within a multi-tier phonological model to synthesize natural and expressive intonation contours. In addition to improving text-to-speech synthesis, the framework is shown to improve voice conversion, both intra-lingually across speakers, and cross-lingually across languages.

We apply the proposed techniques on synthesis of Audiobooks by incorporating richer semantic and contextual features beyond the sentence. We also look at the closely related problem of voice conversion within the SPAM framework to more effectively capture the speaking style of a target speaker. The techniques are also applied for the case of cross-lingual voice conversion, in the context of speech-to-speech machine translation which aims to automatically dub a video into a target language, while preserving the speaker's intent in the original language after translation. Appropriate objective and subjective evaluations are conducted to show the performance of the proposed techniques.

Resumo

A síntese de fala com parâmetros estatísticos (SPSS - Statistical Parametric Speech Synthesis) oferece maior flexibilidade e vantagens computacionais quando comparada com outros métodos de síntese de fala a partir de texto (TTS Text-to-Speech Synthesis). Apesar da fala resultante ser inteligível, as vozes produzidas por treino estatístico são menos naturais, em resultado do processamento de sinal e do tratamento estatístico necessário para a construção dos modelos. Uma grande parte da responsabilidade pela menor naturalidade deve-se a uma prosódia inadequada e monótona da fala sintetizada. A fonte sonora, que afeta diretamente a prosódia, com informação complementar à do trato vocal, tem padrões próprios, que precisam ser tratados de forma adequada. Sob esta hipótese, a tese investiga as representações e estratégias ideais para a modelação da prosódia no paradigma SPSS.

Propõe-se o modelo de entonação estatística SPAM (Statistical Phrase/ Accent Model) como uma estrutura que é simultaneamente (i) um modelo computacional tendo associados métodos para o treino e a síntese da prosódia e (ii) uma forte base teórica para a descrição prosódica . O modelo SPAM combina os pontos fortes de pontos de atuais visões complementares da entonação como a fonologia métrica autossegmental, paradigmas de produção como o modelo Fujisaki e abordagens puramente computacionais como o modelo TILT. Demonstramos que os grupos entoacionais (Accent Groups), uma nova unidade fonológica extraída de dados, como o nível de representação ideal para modelar acentos tonais e integrá-los num modelo fonológico multinível para sintetizar contornos entoacionais de forma natural e expressiva. Além de melhorar a síntese de fala a partir de texto, o modelo mostrou melhorar a conversão de voz, tanto entre falantes na mesma língua, como entre línguas diferentes.

As técnicas propostas são aplicadas na síntese de livros falados, incorporando uma mais rica informação semântica e contextual para além da frase. Também se aborda o problema relacionado da conversão de voz no âmbito do modelo SPAM de forma a capturar de modo mais eficaz a forma de falar de um indivíduo. As técnicas são também aplicadas para o caso de conversão de voz entre línguas, no contexto de tradução automática de fala para fala, com o objetivo de efetuar a dobragem automática de um vídeo numa outra língua preservando, apiós a tradução, a intenção do falante na língua original. As técnicas propostas foram alvo de avaliações objectivas e subjetivas adequadas para mostrar o seu desempenho.

Acknowledgements

As a thesis conducted under two advisors, two schools in two countries, there are a lot of personal and professional thanks due.

My foremost thanks go to my advisors Alan Black and Luis Oliveira. Alan has had a tremendous influence on my thought and research process and has been a great advisor through the last five years. I am indebted to him for putting up with my procrastinating ways and often going out of his way to bail me out. Luis has been a great advisor at INESC-ID and his shrewd understanding and direction gave me a number of good ideas that refined this work. The most impressive aspect of this thesis yet may be its stellar committee! I thank all the members — Justine Cassell and Bhiksha Raj at CMU; Mario Figueiredo and Isabel Trancoso at IST and to the external juror Paul Taylor for their time and comments reviewing this thesis. I only wish I had more time working with each one of them. Outside of this committee, my discussion with Hiroya Fujisaki influenced the design of the core contributions of this thesis.

I spent the formative years and some key moments in conducting this thesis work at CMU and owe a lot to the training and culture I received there. Thanks to Brian Langner, Udhaykumar Nallasamy, Prasanna Muthukumar, Sunayana Sitaram, Vamshi Ambati, Pooja Reddivari, Aasish Pappu and Alok Parlikar for being good friends, bearing with me, cheering me up and helping me many different ways.

Most of the research was conducted in Lisbon, Portugal where I had a productive and enjoyable stay. I was fortunate to know and work with a number of people at INESC-ID starting with Isabel Trancoso who I cannot thank enough for all the help, advice and encouragement. Thanks also to a number of my friends in Portugal who either helped me with work or made sure I had a life outside of it — Vladimir Ivannikov, Yaim Cooper, Farid Bozorgnia, Jose Portelo, Tiago Luis, Wang Ling, Paula Vaz, Helena Moniz, Hugo Meneido and friends at ISKCON Lisboa, Rupa, Govinda and Kalindi.

I gratefully acknowledge the CarnegieMellon|Portugal programme and particularly the Fundação para a Ciência e a Tecnologia, Portugal for funding my doctoral studies and research.

My biggest gratitude goes to my parents and sister for the unending encouragement and prayers. This thesis wouldn't have been possible without my wife Satya's support and understanding to whom I can only promise to make up for all the time lost in the long distance !

Contents

Part I

Introduction

Chapter 1

Text-to-Speech Synthesis

Text-to-synthesis is the technology that aims to generate speech output from text input. There has been a long tradition in the practice of speech synthesis from the early attempts by Wolfgang von Kempelen in eighteenth century using physical models, to today's pervasive use of the technology in popular products like Apple's Siri, Google Now etc. This chapter provides a brief overview of the field over the decades with a description of the current state-of-the art. The narrative is attempted to lay the foundation for the rest of the thesis. While this chapter is aimed to provide a holistic view of the technology, some concepts beyond the scope of this thesis may be grossly omitted or treated only cursorily.

1.1 Human Speech production

Before reviewing speech synthesis by computers, it is illustrative to understand the speech production process in humans. The first stage in speech production in humans is at speech planning in the brain at the semantic level. The planned concepts are then realized as a sequence of words, and phonemes in the language in the speech motor cortex and the associated stimuli are sent to the speech production faculties, namely the lungs, vocal cords and the speech articulators [Bouchard et al., 2013]. Figure 1.1 shows the speech production apparatus in humans.

Air from the lungs when released under pressure passes through the vocal cords causing them to vibrate at regular intervals to produce *voicing* in vowel sounds like */ae/* (as in *cat*) or */iy/* (as in *heat*), and remain relaxed, not vibrating during the production of unvoiced consonants like */s/* (as in *hiss*) or */hh/* (as in *hat*). The

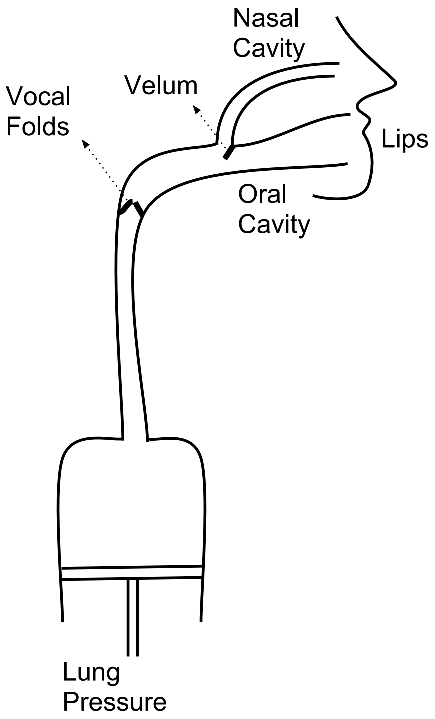


Figure 1.1: Illustration of the human speech production showing the voice source and vocal tract

frequency at which the vocal folds open and close is referred to as the fundamental frequency (denoted F_0) of speech, it is therefore non-existent during the production of unvoiced consonants. The excitation of the vocal folds is further modified by the other articulators like the velum, palate, teeth, lips and nose. These along with the shape of the mouth, tongue and the position of the tongue constriction (the place and manner) add higher frequency resonances to the speech that enable us to produce different *phonemes* of the language. Phonetics is the subfield of linguistics that deals with aspects of production and perception of different sounds [Stevens, 2000]. A related field is of Phonology that deals with the systematic organization of the sounds to form different words in the language [Chomsky and Halle, 1968].

Much like in the case of text, realizing notions of grammar, syntax and semantics is not straightforward in speech, which is a more complex and continuous phenomenon [Steedman, 2013]. Hence the variance when different speakers utter the same sentence or the same speaker utters a sentence in different ways in different contexts. The speech produced also depends on para-linguistic factors that effect the speaker like emotional state, context of the sentence, speaking environment, listener's level of interest and the social distance. Given the same string of words, there is more than one way of uttering it in order to convey different meanings or varying degrees of expression of the same meaning as relevant to its context.

Given a sentence, it is uttered in phrases that are separated by long or short pauses. Each phrase is delivered by speaking every word with the right level of intensity and emphasis to effectively convey the underlying intent of the sentence as appropriate for its context. There are hence two aspects in spoken language (i) *what* has been said in an utterance ? — the underlying sequence of words and (ii) *how* it has been said ? — all non-verbal detail that cannot be alluded to in text (also known as its prosody, more on this in Chapter 2).

1.2 Historical Overview

Early attempts at speech synthesis were done through physical models of the human speech production apparatus [von Kempelen, 1791]. Air was pumped through constrictions and tubes to simulate the vocal tract. However, these attempts were purely meant for educatory purposes to advance our understanding of phonetics and were only moderately successful in synthesizing various speech sounds. The next well known synthesizer was the *Voder* [Dudley, 1938] by Bell labs which was a combination of several electro-mechanical devices which were manually operated to produce certain English phonemes. The most significant development then was the invention of the spectrogram, a spectro-temporal representation of speech that spawned a sub-area called Acoustic Phonetics which dealt with the study of spectral properties of vowels and consonants [Jakobson et al., 1952]. These studies lead to the development of Formant synthesizers that aim to completely reconstruct speech from a carefully recorded set of parameters and a cascade of filters designed to mimic the resonances of the vocal tract [Klatt, 1980]. These values were manually recorded by experts by observing the spectrograms of phonemes recorded in different contexts.

The next paradigm of concatenative speech synthesis is made possible largely due to the advances in digital storage and processing techniques. Segmented speech instances are joined together optimizing certain acoustic distance measures over a voice database of the target speaker's recorded speech. These could be subword instances like diphones [Olive et al., 1993], or longer units, as available in a 'large' database from the desired speaker and speaking style. These techniques are known for the naturalness in the resulting voice, because segments of natural speech are joined together, pitch synchronously to generate the target sentence. It is worth mentioning that Unit Selection [Hunt and Black, 1996], a concatenative synthesis technique (or a close variant of it) remains the preferred technique for most commercially available speech synthesizers and is still widely researched in several languages.

A complementary paradigm of statistical speech synthesis [Tokuda et al., 1995] has emerged with advances in automatic speech analysis and synthesis techniques. The so called Statistical Parametric Speech Synthesis (SPSS, [Zen et al., 2009]) systems are currently among the most researched systems and are garnering immense interest due to their reliability, ease of training and relaxed requirements on the size and consistency of the speech database. The techniques presented in this thesis are developed for (and within) the SPSS paradigm and a brief introduction is provided in the following section.

1.3 Statistical Parametric Speech Synthesis

The premise of statistical parametric speech synthesis is that speech can be reconstructed from parametric representations that can be automatically extracted and predicted from text. This is the main contrast that differentiates it from the concatenative techniques in that it doesn't contain any real speech as the model, but just the statistics of speech representations [Zen et al., 2009]. In this section, we will briefly look at the components of an SPSS system.

As with all statistical systems, there is a training phase which involves building source and filter models that are used at test time to generate speech. Each word is expanded as a sequence of its constituent phonemes. Within the Clustergen SPSS system [Black, 2006], each phoneme is modelled as a 3-state Markov sequence (shown in Fig. 1.2). An appropriate spectral representation of speech is selected and the speech is analyzed into this feature representation. Usually these are LPC or MFCC features over a short window (called frame, analysed over a window of 25 millisecond with a shift of about 5 millisecond) of speech. All the speech frames belonging to a single state are pooled together and their statistics (Gaussian mean/variance) are estimated.

The first step in creation of an SPSS voice is the sub-phonetic segmentation of the entire training data. Data-driven techniques using Hidden Markov models are often employed for segmentation [Prahallad et al., 2006]. In the next stage,

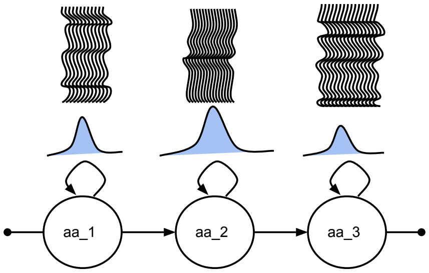


Figure 1.2: Modeling each phoneme as a 3 state Markov chain in SPSS

speech is analyzed into parametric representations. The models are then built to predict these representations only from text input. In order to effectively capture the variance within the entirety of training instances, the contexts are typically clustered using decision trees using linguistic, syntactic and positional features, shown in Fig. 1.3. The intermediate nodes of these trees include discrete or continuous valued questions about each training instance and the leaf nodes contain the statistics of the frames that fall within those set of questions. A fully trained voice consists of models for the spectra, F_0, and duration for each phoneme state.

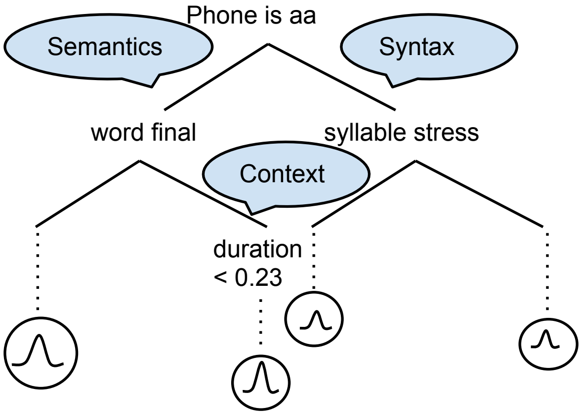


Figure 1.3: Contextual decision trees for storing statistics of spectra, F_0 and duration

At runtime, for synthesizing speech, text input is analyzed and normalized for expansion of abbreviations, number sequences etc [Sproat et al., 1999]. A phrasing model then 'parses' the words into groups of words that are uttered with pauses between them [Parlikar and Black, 2011]. This is followed by a dictionary lookup (or prediction using letter-to-sound rules) to convert the word string to a string of phonemes. At runtime, the trained spectral and prosodic decision trees are traversed to predict the optimal parameters [Tokuda et al., 2000] that are converted to speech using an inverse filter [Tokuda et al., 1995]. Figure 1.4 illustrates this sequential pipeline architecture showing the different stages in conversion of text to speech.

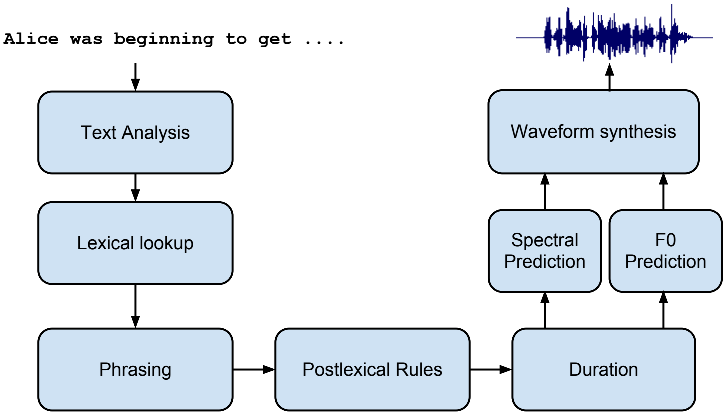


Figure 1.4: Runtime architecture of an SPSS system, synthesizing speech from text

The SPSS paradigm offers several advantages in that it can generate acceptable and intelligible voices without requiring large, consistent and high quality speech databases. Also the design of a parametric representation makes it amenable to a variety of transformation scenarios like speaker transformation [Toda et al., 2007], changing the voice characteristics for a noisy ambience [Anumanchipalli et al., 2010] or for synthesizing a different language [Anumanchipalli and Black, 2010]. However, SPSS techniques are far from sounding natural, with most qualitative assessments characterizing the synthetic voice as sounding robotic and unnatural. Current research in SPSS has been to largely improve the naturalness of statistical voices and in this thesis, we improve SPSS by focussing on the prosody modelling.

1.4 Thesis Statement

We've seen that SPSS techniques have advanced to a stage where they are completely intelligible. This is very much to the credit of advances in spectral modelling and optimal prediction strategies. While the voice quality of SPSS based systems is acceptable and intelligible, often better than Unit Selection systems, they score low on the naturalness and preferrability metrics than the latter [King and Karaiskos, 2009].

Qualitatively, SPSS voices are perceived to be lacking in expression compared to high quality Unit Selection voices, mainly due to the monotonous and inappropriate prosody of the synthesized speech due to the averaging out of speech frames from different contexts. Earlier techniques have dealt with F_0 as yet another parameter stream with the spectrum and the same modelling techniques have been used to model spectra and F_0 [Black, 2006, Zen et al., 2007]. This is sub-optimal since F_0 has its distinct perceptual function and its own acoustic patterns that distinguish it from the more quasi-static spectral features. To corroborate this hypothesis, this thesis aims to show that —

"It is possible to model intonation in a way that is both theoretically sound and computationally feasible for (i) synthesis of expressive prosody in Text-to-Speech; and (ii) for conversion of speaking style across speakers and languages."

We propose a model for computational prosodic description — including its analyses, modelling and conversion. We evaluate the proposed model by application to creation of high quality text-to-speech systems in prosodically diverse tasks. The framework is further exploited for the closely related problem of voice conversion. This includes applications to conversion between two speakers in a language and in the context of speech-to-speech machine translation between two languages.

1.5 Thesis Contributions

The following are the contributions of this thesis in speech synthesis and voice conversion —

- **Prosody modelling in SPSS**: We improve state-of-the art statistical parametric speech synthesis by proposing a multi-tier phonological intonation model (SPAM) for generation of F_0 at synthesis. We consolidate complementary views of intonation to empirically determine the optimal phonological unit to model intonation, a very practical engineering approach to validate findings in speech science and for application in speech synthesis technology.

- **Automatic discovery of Accent Groups**: We use a linear time algorithm that determines accent groups, a data-derived phonological unit which, besides improving SPSS, can be exploited to bootstrap speech synthesizers in languages without writing systems.

- **Audiobook Synthesis**: We apply the proposed models for synthesis of tasks with increasing prosodic variety. We incorporate richer semantic features and propose techniques for synthesis in the context beyond the sentence, evaluating this on multi-paragraph text input like audiobooks.

- **Style capturing voice conversion:** A 2-level approach is proposed for F_0 transformation in voice conversion using the SPAM framework. It is shown to be more sensitive to the speaking style of the target speaker than existing approaches.

- **Intent transfer in Speech Translation:** We describe an approach to transfer the original speaker's prosody appropriately into a target language within the context of speech-to-speech machine translation. The method uses a parallel speech corpus from a bi-lingual speaker to learn the right transform functions. These approaches are extended to automatic dubbing of videos into a target language while preserving speaker intent in the original language.

1.6 Organization of this Thesis

The necessary background into Prosody and related work in intonation is provided in Chapter 2. Chapter 3 investigates various phonological units as the modelling unit for intonation. Chapter 4 introduces the proposed Statistical Phrase/Accent Model as a framework for intonation modelling within SPSS and provides the necessary empirical justifications for it.

As applications of the proposed SPAM framework, Chapter 5 describes synthesis strategies for Audiobooks, with the integration of higher level linguistic information. Further, the SPAM framework is applied for voice conversion, in Chapter 6 presenting a case for a 2-level approach to voice conversion across speakers. Chapter 7 describes an approach for transferring prosodic prominence cues across languages in a speech-to-speech Machine Translation system. Conclusions and potential future work based on this thesis are presented in Chapter 8.

Chapter 2

Prosody in Text-to-Speech

This chapter gives an overview of prosody — its parts, functions and theories as a reference for the concepts introduced in the following chapters. Prosody is that part of spoken language dealing with *how* a sentence is delivered, hence encompassing all non-verbal aspects of the utterance. There are many competing views on prosody, often at odds with one another, both in analysis and synthesis. These are all valid theories with no one unified paradigm of description, alluding to the complexity and richness in prosodic form that humans employ in spoken communication. Once again, it is impossible to review all existing paradigms of all aspects of prosody. This chapter only highlights those parts that are relevant for the reminder of the thesis. More comprehensive accounts on Prosody may be found in [Taylor, 2009, Chapters 6 & 9].

2.1 Prosody in Spoken Language

As the broad phenomena that manifests the manner in which a sentence is uttered, there are several acoustic components to what constitutes prosody. These include the phrasing patterns, rhythm and intonation, which is the pitch or tune of the utterance. Of these, intonation forms the most important part of prosody, to the extent that both the terms are used interchangeably. Phrasing and duration are the other major elements that constitute prosody.

Duration (Segmental duration) refers to the time that the speaker spends within each phoneme in delivering the utterance. Several low level factors like phonetic context, prominence of the associated syllable, word and phrase effect the segmental duration.

Phrasing refers to the process where in speakers group words within an utterance, each group separated by a short or long pause. Given a sentence, there are several valid ways of grouping the words and laying different pause durations in speaking it. While there exists an agreement that phrasing occurs in systematic ways that are generic across speakers (to a certain extent), there is no universally accepted theory on how to describe phrasing. This is because there is no direct relation between the linguistic syntax and prosodic phrasing. It depends on many acoustic and phonological constraints factors like the speaking rate, number of words and more importantly, the general idiosyncratic traits of the speaker.

Intonation forms a major part of prosody (and this thesis). The term intonation is used to refer to the systematic way in which speakers employ *Pitch* to convey the underlying meaning in a sentence. *Pitch* is the perceptual correlate to what listeners perceive as the overall tune of the utterance. Pitch itself is directly related to the Fundamental Frequency (or F_0) of the speaker's vocal cords while speaking the sentence. Informally, the terms Pitch contour, Fundamental Frequency and Intonation all refer to the same notion of spoken communication. Fig 2.1 shows an example F_0 contour of an utterance by a female news reader from the Boston University Radio Speech Corpus [Ostendorf et al., 1996]. (All F_0 contours illustrated in this thesis are extracted using methods described in [Yegnanarayana and Sri Rama Murty, 2009], and interpolations are carried out over all non-silence, unvoiced regions to model F_0 as a continuous phenomenon over each phrase).

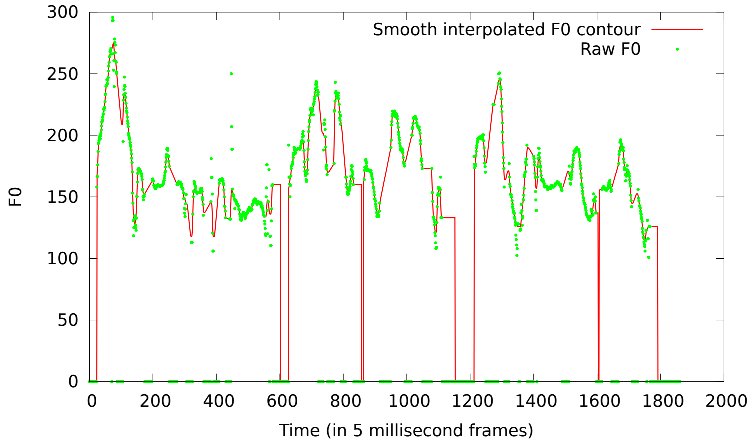


Figure 2.1: A smooth F_0 contour interpolated through unvoiced speech regions

It can be seen that the raw F_0 (green dots) is non-existent in unvoiced regions, in contrast to the smooth F_0. The smoothed F_0 interpolated through the unvoiced regions also removes some spurious candidates, there by generating contours that can be continuously analyzed with respect to the underlying text. Following the notation introduced by [Bolinger, 1986], Fig 2.2 shows the word sequence for the same utterance as Fig 2.1 where the words are placed roughly at their mean pitch during the delivery of that word.

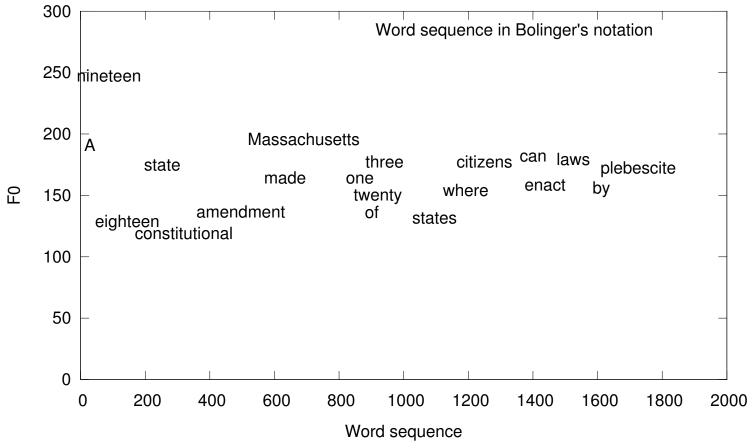


Figure 2.2: Pitch aligned word sequence in Bolinger's notation for the sentence *"A nineteen-eighteen state constitutional amendment made Massacheussetts one of twenty three states where citizens can enact laws by plebiscite."*

The relative pitch differences between the words signify different levels of importance, for the information conveyed in the sentence. This is given by the novelty of the concept introduced and the speaker's intention of where she wants her listeners to focus on. It can be seen that the speaker has quite effectively used her F_0 to lay emphasis on the words in the sentence in the context of the longer paragraph (not shown in Fig 2.1). Of course this representation shows only word level differences, but the same holds at the syllabic or phonetic levels as appropriate to the lexical stress (primary or secondary etc.,) or accentedness levels of the associated syllables.

The overall trend of the F_0 broadly reveals the type of the sentence. For example, in English, questions are likely to have a contour rising towards the end of the utterance and declarative/neutral statements may have a falling contour. The emotional state of the speaker (anger, happiness, sadness, fear, disgust and surprise) also effects the intonation showing a wider or narrow dynamic range, according to the speaker state.

The two important aspects that are widely discussed of the F_0 are the excursions (or tones) on the contour from the global trends, the highs and lows; and boundary tones, the trajectory of the contour immediately before the phrase boundaries. Bolinger himself was the first to coin the term *Pitch Accent*, to refer to the recurring excursions that add the intended meaning to the underlying text.

Beyond the phenomena identified, there are additional aspects like affective and augmentative prosody that add some redundant variance (and microposody, the local flutter) to the contour to ensure ample information is provided to the listener to decode the intended meaning.

2.2 Prosody Modeling in TTS

In TTS, Phrasing is the first step in the synthesis of prosody (refer Fig. 1.4) and has a direct bearing on all subsequent stages like F_0 prediction in the synthesis pipeline. Traditionally, a shallow decision is made at each word boundary, whether or not there should be a phrase break in that context. Parts of speech and positional information are used as features for classifiers that assigns breaks (or not) at each word boundary [Black and Taylor, 1997]. These models are trained on a segmentations of a standard speech database. Recent techniques show the speaker dependency of phrasing patterns and claim improvements by explicitly modeling speaker and style specific phrasing patterns [Parlikar and Black, 2011].

In the Clustergen SPSS system [Black, 2006], duration models are used to predict the segmental durations of each phoneme state. The spectra and F_0 are then predicted using trained CART models for each frame (5 millisecond duration). These decision tree models have syntactic and contextual questions that are both categorical and continuous) as the intermediate nodes. In all, the Clustergen system uses 61 base features in training the trees. Table 2.1 lists some of the features used.

While the modeling unit changes for the parameter streams (sub-phoneme state for duration and frame for spectral and F_0 modeling), the features used for clustering are about the same. There are some differences between different implementations of SPSS – HTS [Zen et al., 2007] predicts even F_0 and spectra at the level of the sub-phoneme state and performs optimal interpolation through the frames. OpenMary [Schröder and Trouvain, 2003] predicts a target value for each word before interpolation. For the frame-based Clustergen system, an example

Table 2.1: Some contextual/lexical features used in CART model training.

phoneme identity	durational features (if available)
phoneme state	relative and absolute positions in phoneme state
positional features	features of the phoneme/syllable/word
#content words in left/right	distance from phrase boundary
#syllable coda/rhyme characteristics	presence of minor/major phrase boundary
Parts of speech	All of the above for previous/next units
#syllables in word	⋮

predicted contour for an unseen test sentence is shown in Fig. 2.3. The same durations employed in the reference speech are used for prediction to allow the direct comparison of the F_0 contours.

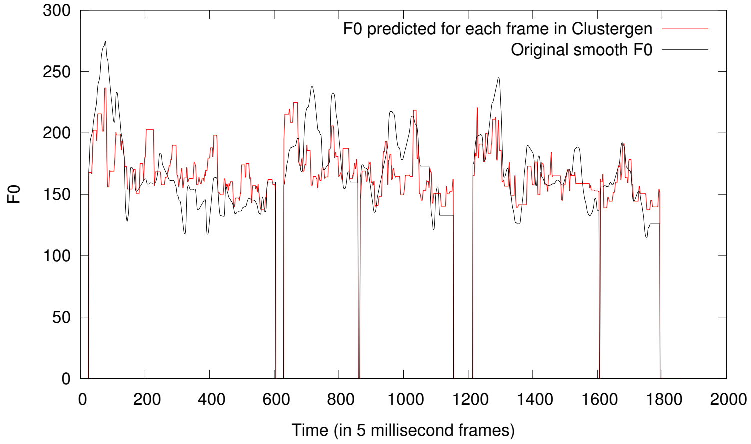


Figure 2.3: Predicted F_0 contour against an unseen sentence's smooth F_0 contour

It can be seen that the predicted contour (in solid red) is relatively not as smooth as the original contour. The same value for F_0 is predicted for many consecutive frames. This is because the features used for clustering do not have a low enough resolution to discriminate F_0 at the level of the frame. This causes the synthetic speech to sound unnatural. Smoothing techniques after prediction have shown little improvement to improve the synthesis quality.

Another problem is that of reduced dynamic range of the synthesized contours. Again, this is because the linguistic features used are very rudimentary and cluster semantically distinct regions together. This is what causes the synthetic speech to sound monotonous and void of any affect, which we saw was the main goal of intonation. To empirically confirm this, Table 2.2 shows the mean and standard deviations in the values of predicted and original F_0 contours for 45 test sentences for a Clustergen voice built on [Ostendorf et al., 1996, Speaker F2B]. Note the significant drop (about 13 Hz) in the variance of predicted contours.

Table 2.2: Comparison of F_0 mean/ranges for the original and synthesized F_0

F_0 source	**mean**	**range (std/dev)**
Reference F_0	167.85	30.28
Predicted F_0	168.67	18.55

These drawbacks exist across all SPSS systems, and work arounds like artificially increasing the variance have been proposed [Toda and Tokuda, 2007] to improve the perceived naturalness of synthesis outputs. The goal in this thesis is to improve the naturalness and acceptability of SPSS voices through improved modeling of F_0. Towards this goal, the following sections provide an overview of general paradigms in understanding intonation, and computational F_0 modeling techniques.

2.3 Paradigms in Intonation

Intonation has evolved into a field of its own that is thoroughly researched across diverse disciplines including, speech synthesis, psycholinguistics, language processing, speech pathology etc. While the interest in the latter fields of enquiry is in analyzing how pitch systematically complements or presents underlying intent, in speech synthesis, there is the additional goal of generating appropriate contours only from text input. With this in view, we also comment here on the relevance of each approach discussed for use in Text-to-Speech synthesis.

2.3.1 Intonational Phonology

The premise of Intonational Phonology, is quite literally that there is a phonological organization to pitch contours. While this is obvious for tonal languages like Mandarin, where relative pitch differences among the tones are phonetically perceived to refer to different words, the hypothesis here is that even in intonational languages like English, there is a phonological basis, that is to serve some linguistic purpose to convey the underlying semantics [Ladd, 1996].

Pierrehumbert's influential work on the phonetics and phonology of American English [Pierrehumbert, 1980], and earlier works on Autosegmental Metrical(AM) phonology [Liberman, 1975, Bruce, 1977] laid the groundwork for the current scope of Intonation Phonology. This was (and is still being) extended to several European and Asian languages. Ladd identifies the following four tenets of the AM theory that form the basis of much of the discussion in Intonation Phonology. These are —

1. *Sequential tonal structure*: Describing the pitch contour as a sequence of *events* associated with the underlying segmental string. Between such events, the contour is phonologically unspecified and merely serves as transitions between the events

2. *Distinction between Pitch Accent and stress*: Though pitch accents serve as perceptual cues to stress or prominence, they are part of a larger prosodic organization of the sentence and are distinct both functionally and acoustically from lexical stress on accented syllables.

3. *Analysis of pitch accents in terms of level tones*: Pitch accents and boundary tones can be qualitatively analyzed as consisting of primitive level tones or pitch targets, High (H) and Low (L).

4. *Local sources for global trends*: The phonetic realizations of a pitch accent (actual trajectory of the contour) is entirely complementary to the tone identity as an H or L, that are dependent on other factors like its position in the utterance etc.

A practical exposition of these tenets is proposed and an analysis scheme for studying intonation has developed as the ToBI standard (Tones and Break Indices [Silverman et al., 1992]), extending Pierrehumbert's original recommendations for F_0 analysis. The scheme recommends combinations of the H and L symbols marking all phonologically interesting events with this notation, denoting with a '*' for the accents and a '%' for the phrase boundary. Fig. 2.4 shows an expert annotation of a pitch contour under the ToBI scheme.

Once annotated, discussion in intonational phonology then is on the sequence of events, and what they functionally signify in the speech, how each speaker, dialect or language may use these uniquely shaped events to cater to distinct communicative needs. As a framework to 'explain' the pitch contour in tandem with the underlying

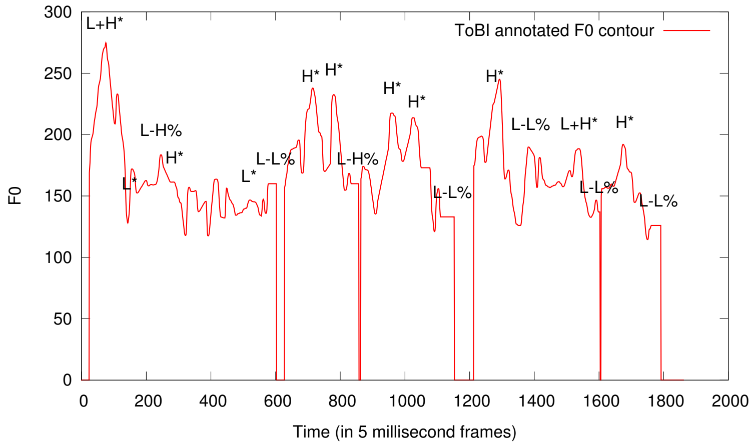


Figure 2.4: A smooth F_0 contour annotated with ToBI labels

segment sequence, ToBI is very successful and spawned several parallel lines of enquiry.

The biggest bottleneck in this approach is getting the annotations itself. ToBI annotation is very subjective and takes lot of effort to identify which events are phonologically relevant. Also, the inter annotator disagreement is quite high even among experts. Though there are current efforts to automate the annotation process [Rosenberg, 2010], the fundamental qualitative nature of the framework makes it unsuitable for TTS as such. There have been methods for synthesis [Anderson et al., 1984, Black and Hunt, 1996] that convert tone sequences to a pitch contour but these are either rule driven and/or expect a tone sequence, thereby not scaling to general TTS systems. This also holds for rest of the research in speech sciences which remain qualitative and descriptive but offer little predictive knowledge about intonation [Xu, 2012], an invariable requirement for TTS.

2.3.2 Production Models

A parallel view of intonation based on human speech production was independently proposed as the now famous Fujisaki model [Fujisaki, 1983]. The model is rooted in the fact that while speaking, the sub-glottal pressure decreases over the length of the phrase. This causes lesser velocity of air passing through the vocal cords causing the fundamental frequency to decrease towards the phrase end.

Consequently, the approach recommends 3-level additive modeling of the logarithmic F_0[1] contour comprising the baseline, phrase and accent commands. The baseline contributes to the minimum value of pitch for the speaker. The phrase and accent commands themselves are each modelled as critically damped second order filters, generating steep declining contours, controlled by parameters. While the original formulation recommended using a constant value for a speaker, later techniques have shown benefit with variable values depending on the task and sentence type. Figure 2.5 illustrates the Fujisaki model where the F_0 is formed additively by the three underlying baseline, phrase and accent components.

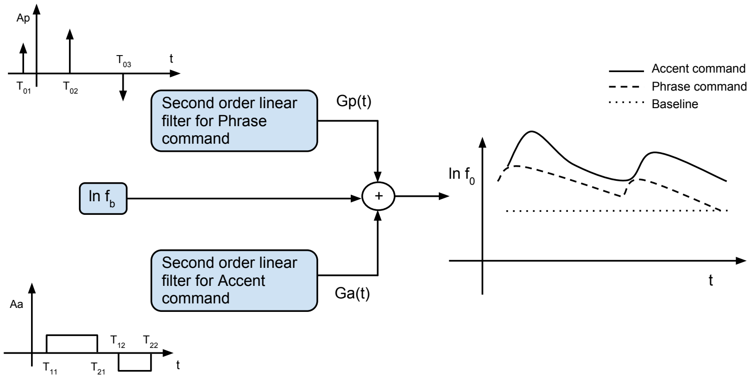


Figure 2.5: Fujisaki model — ln (F_0) as a superposition of the baseline, phrase and accents.

Though Fujisaki model was initially developed for Japanese declarative sentences, it is later developed in other languages. While the original method prescribed expert annotation to mark the beginning of the accent and phrase commands, automatic techniques are now used for extraction of Fujisaki parameters from directly from speech data [Mixdorff, 2000]. The Fujisaki model offers an elegant explanation for the general overall trend of 'declination' (or 'downdrift') of F_0 over the phrase, however, the model is incapable of synthesizing arbitrary F_0 contours. By design, the model assumes a falling contour, consequently not generating phrase final rises (like question utterances that have a definite rise towards the end). Also the recommendation of a constant value for the filter parameters cannot explain arbitrary pitch contours with a wide variety of events on them, not necessarily steeply falling.

While providing a sound mathematical formulation for F_0, the Fujisaki model doesn't relate the process (or the commands) to an underlying linguistic structure,

[1] Logarithmic F_0 is related to semitones, making the notion of superposition tangible.

making the model unsuitable for use directly in a TTS system. Where as in TTS, we have only text from which F_0 contours need to be predicted.

The idea of modeling F_0 as an additive phenomenon, however, gained traction with many techniques falling within the paradigm of super-positional contours [Sun, 2002, Bailly and Holm, 2005, van Santen et al., 2005, Sakai et al., 2009, Anumanchipalli et al., 2011, Wu and Soong, 2012], the recent techniques even under the SPSS paradigm.

2.3.3 The Tilt Model

With a view to develop a purely engineered description of the pitch contour, Taylor developed the Tilt model [Taylor, 2000a], with exact analysis and synthesis methods. Fundamentally, Tilt subscribes to the AM theory in describing the F_0 contour as a sequence of events connected by transitions. However, rather than categorizing these shapes, Taylor proposes using continuous feature tuples that can losslessly (almost) code the shape in terms of the event amplitude duration and a shape descriptor that can represent any arbitrary shape between a complete rise (+1) to a complete fall (-1).

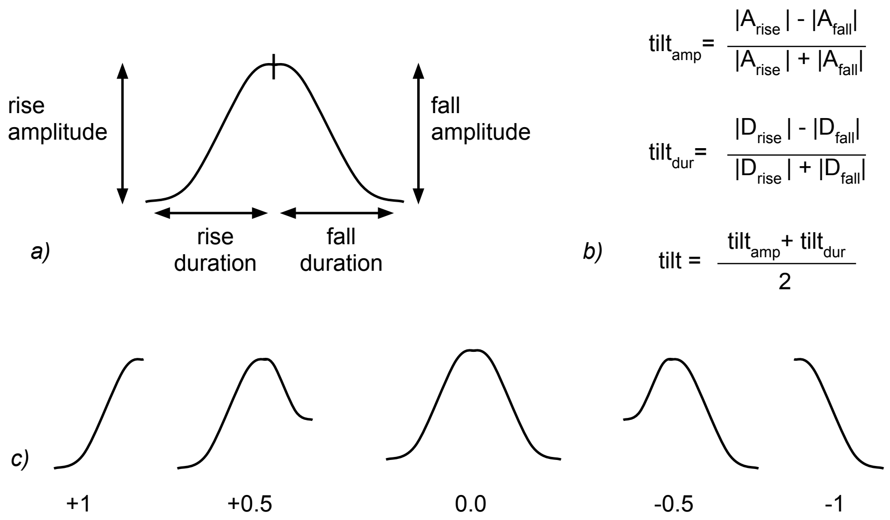

rise amplitude
fall amplitude
rise duration
fall duration
a)
b)
$\text{tilt}_{amp} = \frac{|A_{rise}| - |A_{fall}|}{|A_{rise}| + |A_{fall}|}$
$\text{tilt}_{dur} = \frac{|D_{rise}| - |D_{fall}|}{|D_{rise}| + |D_{fall}|}$
$\text{tilt} = \frac{\text{tilt}_{amp} + \text{tilt}_{dur}}{2}$
c)
+1
+0.5
0.0
-0.5
-1

Figure 2.6: a) Analyzing each intonational event using its rise/fall values. b) three parameters to code any arbitrary rise/fall event c) Examples of 5 pitch accents with the continuous *tilt* value ranging from -1 to +1

Similarly, Taylor proposed synthesis methods that use the tilt parameter tuple to synthesize an appropriate event bearing those values. The approach also 'links' each intonational event to an underlying segmental sequence of syllables, where every accented syllable is linked to an event on the pitch contour. The connections are prescribed to be linear from the end of one event to the beginning of another.

While the Tilt model is not fundamentally rooted in the physiology or phonology of speech, it offers several practical advantages in that it can automatically analyze and synthesize arbitrary contours (hence evading the problems of the other models). [Dusterhoff et al., 1999] further applies the Tilt model to predict F_0 contours from text in a TTS system, by first predicting the accented syllables and then the associated Tilt vector.

2.3.4 Other Methods

While we have seen three very different views to understanding intonation, there still are relevant others. There are stylization methods like [d 'Alessandro and Mertens, 1995] which model the contour with piecewise linear, yet perceptually lossless approximations; the MOMEL method of smoothing using quadratic splines [Hirst et al., 2000] that are later analyzed for phonological relevance. [Möhler and Conkie, 1998] employs vector quantization over parametric F_0 representations. There are also approaches borrowing the unit selection paradigm where there is no explicit notion of modeling prosody, but natural contours from similar contexts in a speech database are used after certain cost optimizations [Malfrere et al., 1998, Meron, 2001, Raux and Black, 2003].

Yet another complementary paradigm is of information structure where semantic aspects like theme and rheme are studied as relevant for intonation [Hirschberg and Pierrehumbert, 1986, Hirschberg, 1992, Prevost, 1996]. These are not alternative strategies but go alongside intonational phonology yet with a view to completely describe intonation in its entire pragmatic context [Prevost and Steedman, 1994]. A comprehensive review of these methods can be found in [Steedman, 2013].

2.4 TTS desirables for an Intonation Model

Despite the whole array of methods discussed to understand and model intonation, there still isn't a working framework for the synthesis of appropriate intonation within TTS for any arbitrary speaker and speaking style. This brings us to the question of what the criteria are which a model should satisfy to be considered optimal for Text-to-Speech. As practitioners, we identify the following criteria as being important —

1. Expressive and natural intonation : The primary goal of being able to generate appropriate intonation that is expressive over a variety of tasks and sounds natural doing so.

2. Automatic analysis and synthesis : The ease of automatic computational methods to analyze natural pitch contours and to synthesize them from the parametric representation.

3. Optimal use of training data : The model must make optimal use of the training data, elegantly dealing with cases where only minimal data is available, to cases where several hours of speech may be available from a speaker.

4. Amenable to transformation : Ideally the model is amenable to transformation between speakers, dialects and languages.

5. Predictable from text : The model should be able to use text-based based features with no other available sources of information (like specifying accented syllables).

6. Theoretical relevance : The model must be interpretable to either further our knowledge about speech communication or be able to validate existing theories.

Part II

Statistical Phrase/Accent Model of Intonation (SPAM)

Chapter 3

Optimal Unit for Intonation Modeling

Having reviewed the state-of-the art in general TTS and prosody modeling, this chapter brings us to the current investigation in this thesis to improve these components. In this chapter, we attempt to find the right phonological level to model intonation within SPSS.

3.1 Phonological resolution and synthetic F_0

In this section we describe our attempts at finding the right level to model F_0 for improved naturalness in synthesis. The unique challenge F_0 modelling presents is the complex relationship between the linguistic context and its phonetic realization. We have seen in Chapter 2 that there are some sentence level aspects like whether the sentence is declarative or interrogative which effect the general rise or fall of the contour. The speaker's expressiveness and speaking style effect the F_0 dynamic range (variance). The word level aspects like syntactic category and semantic role with the neighboring words effect the word level F_0 dynamics. The lexical stress patterns also manifest in relative intonational event lengths and duration. It is therefore important that the contributions from all these units are incorporated into training the contextual F_0 trees. However, in any given speech database, since the distribution of these units is skewed in favor of very short and local contexts, the richer linguistic features may not be chosen in model training which is optimized on entropy. Table 3.1 shows the distribution of phonological units in 1 hour of speech for [Ostendorf et al., 1996, Speaker F2B]. Only non-pause regions are counted for phonemes and frames.

Table 3.1: Distribution of phonological units in 1 hour of Radio news speech

Unit	Number of instances
Sentence	464
Phrase	1052
Word	9214
Syllable	14717
Phoneme	38523
Phoneme state	115569
Frame	592830

It is clear that there is a huge decrease in the number of available units to train from as we go to higher order phonological units. Since the data is analyzed at the frame level several training samples have the same values for the higher order features. As an example feature, word POS (part of speech, refer Table 2.1) will be the same for hundreds of frames, even as the value of F_0 may vastly vary over the duration of the word, making the feature irrelevant [Yu et al., 2010].

So it is important that the model bring features from all these levels to bear on its linguistic→prosodic mapping. The modeling unit is the most important aspect in the design of intonation models that effects the quality of this mapping. Here we set out to find the best unit, from among the ones accessible at training time.

3.1.1 Quantitative analysis

Using the Tilt model as a parameterization, CART trees are built with appropriate features for each modeling unit considered. The baseline is the standard Frame level CART tree built on the training data, with a full feature set containing questions from all phonological levels. The performance of such a model is discussed in Section 2.2, where we have seen predicted F_0, lacks the smoothness and dynamic range of natural F_0. Given that syllable is the smallest phonological unit with a valid notion of F_0, we test here, the hypothesis that higher levels of phonology may model F_0 better for these aspects. Table 3.2 compares the mean squared error between the predicted F0 and reference F0 for an unseen sentence for speaker F2B using several modeling units.

Table 3.2: Objective comparison of original and synthesized for F_0

Modeling unit	F0	
	Mean	Std/dev
Original	167.852	30.276
Frame Predicted	168.673	18.549
Syllable Predicted	175.254	16.484
Word Predicted	177.003	18.950

It can be seen that there is not much improvement in the the dynamic range using higher level units like syllable or word. In fact there may be a certain loss due to the overestimation of the mean F_0. This can be explained by the fact that though the general trend of the contour goes down towards the end of the utterance, there are only a few of the boundary units (word or syllable), that are overwhelmed by intermediate syllables, thus discounting their contribution. Moreover the Frame level model are exclusively optimized for the mean of the F0, without any longer scoped-representation, there by giving a better mean estimate.

Table 3.3 shows correlation and mean squared error measures for different modelling units compared against the reference smoothed F0 contours of 45 test sentences of speaker F2B.

Table 3.3: Objective comparison of predicted F_0 contours against references

Modeling unit	Predicted F0	
	RMSE	CORR
Frame	28.02	0.49
Syllable	30.33	0.40
Word	30.34	0.44

It comes as no surprise that frame level contours fare better on these metrics because they are optimized on mean squared error, where as the syllable and word level models have an inherent parameterization which also attempts to take longer aspects like peak position, event amplitude etc., into consideration. It should be noted that these measures are only tenuous in evaluating the quality of synthetic intonation [Clark and Dusterhoff, 1999]. The most reliable indicators remain human perceptual evaluations, where native listeners pick one stimulus over the another (as the current evaluation standard in annual Blizzard TTS challenges [Black and Tokuda, 2005]). The next section does a qualitative analysis of the nature of predicted intonation contours under different modeling units.

3.1.2 Qualitative Analysis

Predicted F_0 contours for the three phonological levels considered are shown in the Figure 3.1.

It can be visually seen that, the predicted F_0 contours are smooth in the syllable and word level predictions, mainly due to the design choice of the Tilt representation. It is interesting that higher level artefacts like peak alignment seem better in the word level modelling. The over prediction of the accents though is perhaps due to limited training data at the word level. Also, they are over smoothed over the length of the word. The syllable level predicted contour have relatively more detail, but lose out on peak alignment (consequently correlation, Table 3.3). The frame level point-to-point prediction has the highest resolution, but it is unnatural. These general trends also hold in listening tests, with word level contours sounding over-smoothed and frame predicted speech sounding buzzy. It should be noted that while we show the word level here for illustrative purposes, it is not a scalable unit for intonation modeling, at least under the Tilt representation which accommodates only one Rise-Fall event. Consider the 4-syllable word *Massacheussetts*, which is likely to have 2 pitch accents in any natural utterance of the word, which are sub-optimally handled at the word level. The problem is far more complex for agglutinative languages like German or Hungarian which allow increasingly more number of syllables per word.

What is clear from this analysis, though is that higher level modeling units can help improve the phonetic realizations and perception of synthetic contours for their smoothness and for improved modelling of longer range aspects of the text-tune alignment like event duration, peak position etc. Another inference is the fact that an accent may span multiple syllables, and the intonational event should be modelled as one unit over the constituent syllables rather than as individual syllables modelling sub-parts of one event. So the optimal unit is above the level of the syllable but not necessarily the word. This brings us to the notion of an Accent Group, detailed in the next section.

3.2 Accent Group as a unit for Intonation Modeling

Our goal now is to model each intonational event as itself, without modelling parts of it. We have seen that the pitch accent could be spread across multiple syllables. So the ideal unit must is beyond the level of the syllable. However, while the word

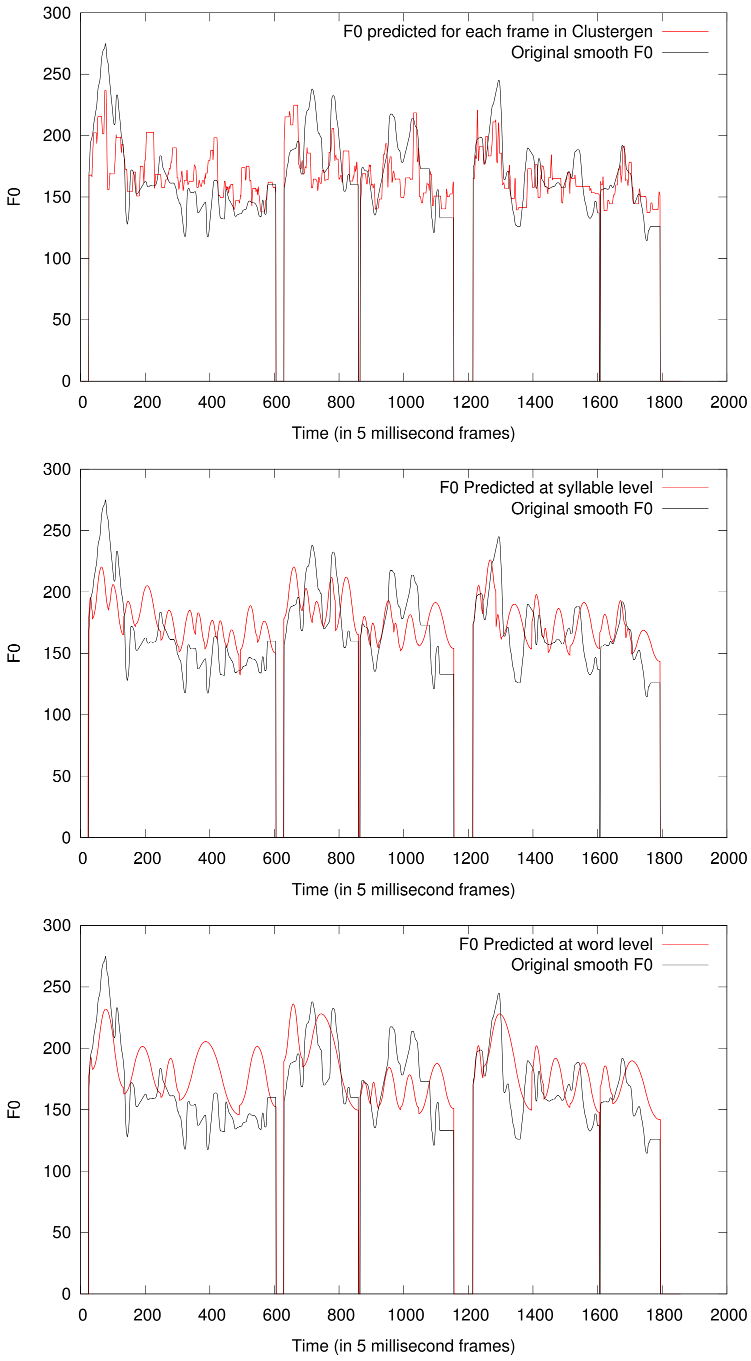


Figure 3.1: Prediction of parametric representation of F0 at the frame, syllable and word levels using Tilt as the representation for the syllable/word units.

is a discrete unit in text, it doesn't hold for intonation, because function words are (often) not phonetically remarkable and may be subsumed in the accents of content word syllables in their immediate context. On the other hand, there are words like *Massacheussetts* which can themselves have more than one pitch accent on them. So, the ideal unit is also not tied to word boundaries. We refer to this abstract unit as an *Accent Group*. Figure 3.2 illustrates this idea showing the underlying syllables that group together, to each intonational event.

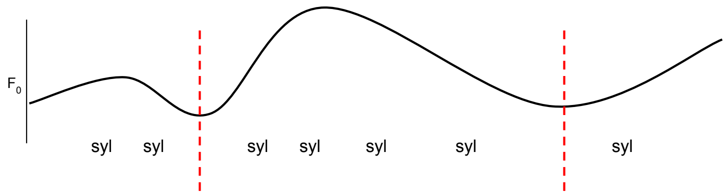


Figure 3.2: Illustrating the notion of an Accent group, an intonationally atomic unit. The vertical red lines mark the accent group boundaries

Needless to state, this illustration is on an artificial contour and real pitch contours are much harder to automatically 'parse' into a discrete set of events. One definite boundary is that of a pause i.e., an Accent Group cannot go across a pause, and has to be demarcated by it (into prosodic phrases). Within each prosodic phrase, however, there are no rules as to how syllables can be grouped together. While this is notionally very similar to *Metrical Foot* [Klabbers and van Santen, 2006], most prescriptions for what a foot should be do not hold when dealing with real speech. Hence, though we appeal to the idea of grouping syllables, we do not use any explicit definition of what an accent group should be — except that it should have only one pitch accent on it. We use a data-driven approach to automatically determine the accent grouping as appropriate to that particular speaker and speaking style used in the training speech data.

3.2.1 Accent Group within Intonational Phonology

The current discussion of a data-driven unit we referred to as the accent group in fact closely relates a number of previously proposed units in speech phonology like the stress group, metrical foot etc.,. It therefore merits a discussion to situate and contrast the current notion of Accent Group explicitly with existing phonological groupings.

The notion of identifying linguistically relevant (supra)segments to explain the rhythm of a language is most formally treated within Autosegmental Metrical Phonology [Liberman, 1975, Pierrehumbert, 1980, Goldsmith, 1990]. The broadest dichotomy in the context of duration between languages is being stress-timed (like English) or syllable-timed (like French). This is to identify what are the isochronous phonological units that explain rhythm in speech. A group of syllables beginning with a stressed syllable followed by any number of unstressed syllables is called a stress group. Similar extensions were proposed to formally define a group of syllables beginning with an accented syllable followed by any number of unaccented syllables [Mobius, 1997]. These groups are often each referred to as (metrical) foot, alluding to recurring patterns of syllable groupings in classical poetry. Beckman [1997] has argued that an intonational typology for spontaneous speech cannot be formally and exhaustively prescribed because of the variety of unknown factors contributing to prosodic variety. There is an interest, however, in empirically describing speech in terms of intonational coding and typology [Klabbers and van Santen, 2006, Gooden et al., 2009].

The purpose of the current definition of Accent Group is thus to serve as a unit that is both phonologically relatable as a group of syllables but to also be generalizable to explain the intonational phonetics of continuous speech. Also, the unit is inherently descriptive, with methods to automatically deduce metrical grouping from Pitch contours. This is in contrast to the prescriptive nature of earlier formal phonological groupings as discussed above. It is to be noted however, that the current Accent Group unit likely encompasses all of the above formal notions.

3.2.2 Data-driven Accent Group Discovery

In this section, we propose an automatic approach to extract Accent Groups, given the speech and an underlying segmental (syllable) sequence. Using Tilt as the parameterization, the technique tries to find the best Tilt-resynthesized approximation for a given pitch contour. Syllables are grouped together, if and only if, doing so helps improve the resynthesis error, or is within an acceptable threshold, accounting for microprosody (which is beyond the discussion at this stage). Incremental combinations of syllables are analyzed together to find the best-fit, where the contour can be approximated to a rise-fall contour with negligible resynthesis error. The exact procedure followed is given here as Algorithm 1.

ϵ is the acceptable error threshold within which a syllable is included within the accent group. ϵ can be empirically set to relax or tighten the condition, asymptoti-

Algorithm 1: Accent Group Discovery from Speech

```
for all phrases do
    accent_group initialized
    for all syllables do
        add syllable to accent_group
        syl_accent = tilt_analyze ($\ln(F_0)$) over syllable
        syl_err = $\ln(F_0)$ - tilt_resynth(syl_accent)
        agroup_accent = tilt_analyze ($F_0$) over accent_group
        agroup_err = $\ln(F_0)$ - tilt_resynth(agroup_accent)
        if ( agroup_err $\geq$ prev_agroup_err + syl_err + $\epsilon$) then
            accent_group= accent_group - { current syllable}
            /* accent group ended on previous syllable */
            output prev_agroup_accent
            accent_group = current syllable
            prev_agroup_err = syl_err
            prev_agroup_accent = syl_accent
        else
            prev_agroup_err = agroup_err
            prev_agroup_accent = agroup_accent
        end if
    end for
    if accent_group $\neq \phi$ then
        /* accent_group must end at phrase boundary */
        output prev_foot_accent
        accent_group = $\phi$
        prev_agroup_err = 0
    end if
end for
```

cally making each syllable its own accent group, to modelling the entire phrase as one rise-fall event. The method is linear in the number of syllables in the utterance and as output, gives a list of accent groups over the underlying syllable sequence.

3.2.3 Analysis of data-derived Accent Groups

In this section, we characterize the automatically detected accent groups both qualitatively and qualitatively. Figure 3.3 shows the performance of the above algorithm on a real utterance. The figure also shows the resynthesized F_0 contour over each Accent Group using the analyzed Tilt parameters. Each rise-fall event is one detected Accent Group.

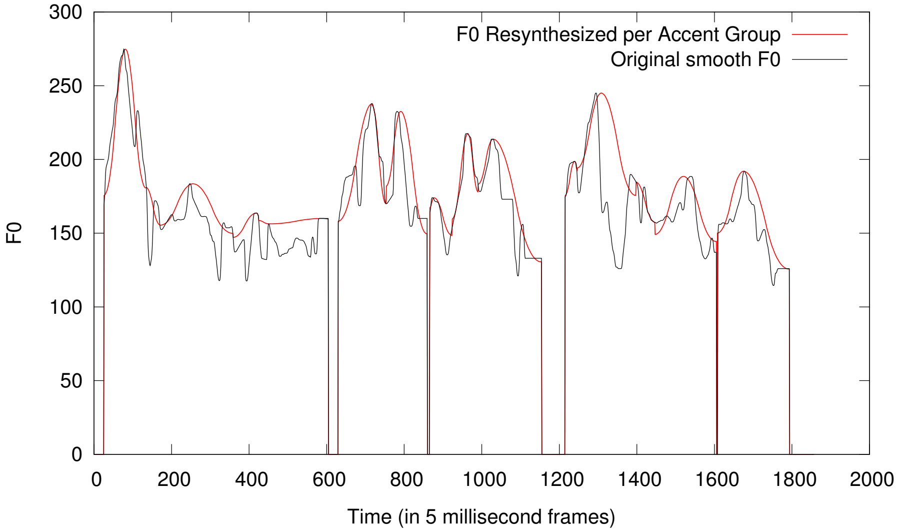


Figure 3.3: Resynthesized F_0 for automatically detected Accent Groups

It can be seen that the algorithm detects all major accents (all **H***s in the ToBI labelling, refer Figure 2.4) in the pitch contour. The local fluctuations are ignored by design. On average, over the entire training data of 1 hour, the correlation of the resynthesized contours against the originals was found to be 0.88, which is quite high, also suggesting that the procedure hasn't made gross errors in detection of the Accent Groups. The number of detected Accent Groups detected under this setting was between the number of syllables and words in the data as shown in Table 3.4. It is worth mentioning that our hypothesis about the ideal phonological unit was also that it was higher than the number of syllables but lesser than the number of words.

These numbers suggest that not all words have an accent on them but every other syllable is likely to have an accent, on average. This is not to be misinterpreted as a rise fall event on every second syllable however, e.g., in the Figure 3.3, the first

Table 3.4: Comparing data-derived Accent Groups against other phonological units

Unit	Number of instances
Sentence	464
Phrase	1052
Word	9214
Accent Group	**7751**
Syllable	14717
Phoneme	38523

accent in the contour is over 3 syllables, and the last accent is over 4 syllables. To exactly reveal the distribution, Figure 3.4 shows the histogram of the lengths (in number of syllables) of detected Accent Groups.

Indeed most Accent Groups are monosyllabic, and the number rapidly reduces over the increasing number of syllables per group. Since we also analyze the contour within each detected Accent Group using the Tilt parametrization, the shapes of the contours are available. A histogram of the shapes, in terms of the `tilt` parameter, is shown in Figure 3.5.

It can be seen that a majority of the Accent Groups are complete rises or falls, these can be interpreted as the connections between actual pitch contours. Discounting for these connections, the distribution is symmetric with around around 0, which is a bell-shaped symmetric rise-fall event (refer Figure 2.6), with the peak located half way through the accent group. These findings are consistent with earlier results on metrical feet in a hand-labelled speech database shown in [Klabbers and van Santen, 2006].

3.2.4 Accent Group Prediction from Text

In the previous sections, we have seen how an utterance can be prosodically ‘parsed’ into its constituent Accent Groups, and have characterized the automatically detected accent groups. The current problem at hand, is however, to incorporate an intonation model within a Text-to-Speech system where text is the only input. This section presents an approach to train a stochastic model that, given a sentence, predicts a valid Accent Grouping over its sequence of syllables.

This problem is analogous to parsing a sentence into its linguistic constituents or to the more closely related problem of phrase break prediction, where a decision

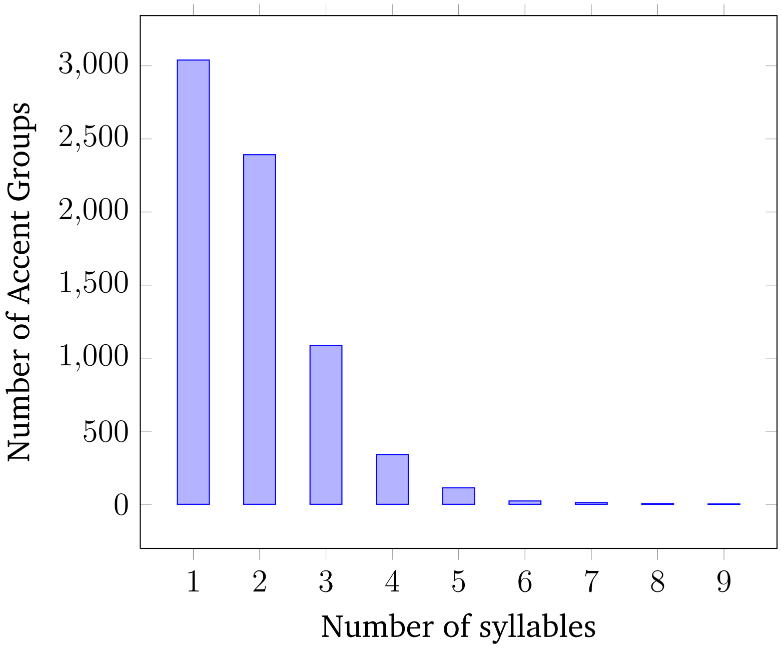


Figure 3.4: Histogram showing the number of Accent Groups against number of syllables in each

is made at each word boundary, whether or not there should be a phrase boundary following that word. In the current scenario, a decision has to be made after each syllable, if it is a valid candidate for an Accent Group boundary. Exploiting the recent results in style specific phrasing [Parlikar and Black, 2012], we employ a similar strategy by training a stochastic context free grammar (SCFG, [Pereira and Schabes, 1992]) of data-derived Accent Group parses of uniquely identified syllables. These form the unique set of terminals over which to train an SCFG, the syllables are tagged with six broad boolean descriptors — if the syllable is phrase final, initial, word final or initial, lexically stressed and has a predicted accent on it. Such a scheme uses about 30 combinations of tags in the 1 hour of Radio news speech presented. Higher number of tags lead to an increase in the number of terminal nodes to process, for which there may not be sufficient data to train an SCFG. To illustrate, a phrase having 4 syllables with 2 accent groups of 1 and 3 syllables each may be represented as —

```
(( syl_1_1_1_1_1_0 ) ( syl_1_1_1_1_0_0 syl_1_1_1_0_0_0 syl_1_0_0_1_0_0 ))
```

Such parses are created using the automatic accent group extraction method

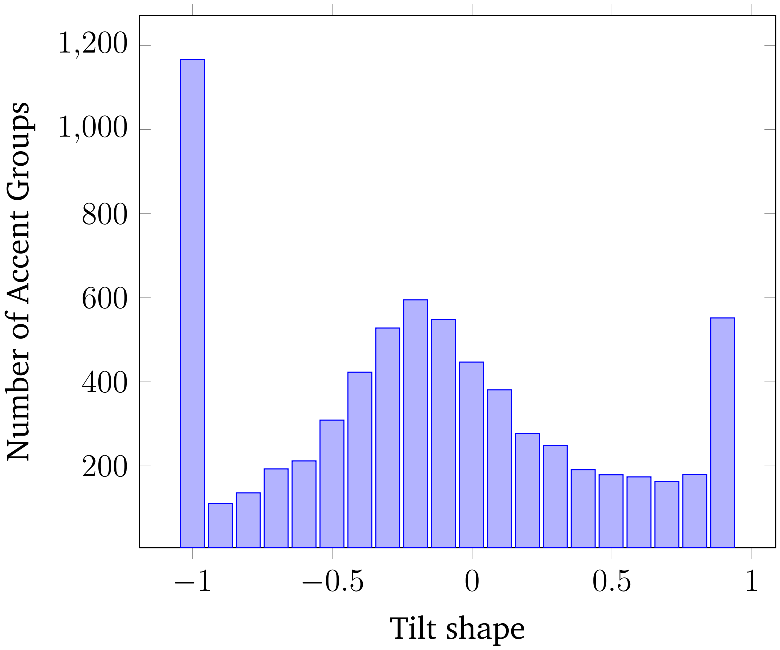


Figure 3.5: Histogram showing number of accent groups within each Tilt shape

and given as the input to the SCFG. Once trained, the grammar can produce parse representations for unseen sequences of syllables at runtime. While useful, these parses are not very accurate since they encode limited information via the boolean descriptors. To induce higher level linguistic features to determine the Accent Grouping decision, we use the automatically induced parses along with other contextual questions about the syllable in question to train a CART tree that can predict the accent boundary decision after each syllable. In all, 83 questions were designed from which decision trees are automatically trained for Accent boundary detection in syllable sequences of unseen text. In the experiments reported here, the models had over 70% accuracy in Break/Non-Break (B/NB) prediction at all syllable boundaries, compared to the reference sequences, obtained through application of the Data-driven Accent Group discovery (Section 3.2.2).

Additionally, to take the predictions on the neighboring syllables into context, an n-gram language model can be trained on B/NB sequences corresponding to the accent groupings in the syllables of the training data. The language model scores can then be used to condition the predictions as appropriate to their context. The language model score can also be scaled to alter the overall number of accent groups that compose a sequence of syllables.

3.2.5 Modeling F_0 phonetic detail over the Accent Group

The last section presented an approach for automatic discovery of phonological units where the F_0 is one contiguous event. This treatment makes it ideal to be modelled using the Tilt method which parametrizes each intonational event as a rise-fall event. The contour over each identified Accent Group is analyzed for the event amplitude, duration, peak position and a shape descriptor, also referred to as `tilt`. These four values form the representation of the F_0 contour bounded within the accent group. Though the connections between genuine pitch accents are also modelled similarly, their insignificance reflects in their amplitudes being small, and as the shape being a pure rise or fall (`tilt` $\simeq$ -1 or +1). The goal now is to model the shape of the contour, given an Accent Group. We do this by training a decision tree that predicts the 4-valued Tilt vector given contextual features about the accent group. A feature set specific for the Accent Group is designed, which includes linguistic and positional features related to the main syllable of the accent group, which we assume as being the first lexically stressed syllable of a content word, the features related to the first syllable, last syllable and word level features for these syllables etc.,. In all, 63 features were used for the clustering at this stage. Since in TTS, duration prediction happens before F_0 is predicted, the only parameters that need to be predicted are the Tilt amplitude, peak and tilt shape. Mean subtraction and variance normalization is done on these values over the entire training data so as not to bias the models towards one of these parameters.

At runtime, the decision trees are traversed for an unseen Accent Group to predict the most likely shape of the pitch contour, given the training data.

3.2.6 The Accent Group in SPSS

So far, we have seen approaches to all the necessary components to making an intonation model that can chunk an unseen sentence into groups of syllables and predict an appropriate pitch contour for it. The final step, is integration of the described intonation model into the SPSS system. Towards this, we introduce a new level within the Festival prosodic structure called the "Accent Group". Each Accent Group is linked to one or more syllables as its child nodes and has the *Phrase* as its parent node. The *Accent Group* level is explicitly not linked to the word level since accents could span syllables across words or a word itself can have multiple accents on it as seen in earlier sections. Fig. 3.6 illustrates the proposed Accent Group unit in the context of Festival TTS architecture [Black et al., 1998] that is used across all

SPSS implementations.

3.3 Objective Evaluation

To evaluate the performance of Accent Group as a modeling unit, the SPSS system was modified to integrate intonation models at different phonological resolutions. An unseen set of sentences in synthesized using the Accent Group intonation model described in the previous section. This includes the two stages of parsing the syllable sequence into Accent Groups and predicting the pitch contour as appropriate to each group. Figure 3.7 shows the contour for an unseen test sentence, the same sentence used in Figure 3.1 for other intonation models with the frame, syllable and word units.

It can be seen that the peak alignment is much better and the dynamic range seems improved relative to other modeling units. This is also shown in Table 3.5 where the Accent Group generated intonation is closer to the statistics of the original F_0 compared to the other levels.

Table 3.5: Comparing Accent Group against other Modeling units for synthetic F_0

Modeling unit	F0	
	Mean	Standard Deviation
Original	167.852	30.276
Frame Predicted	168.673	18.549
Syllable Predicted	175.254	16.484
Accent Group	**173.079**	**21.236**
Word Predicted	177.003	18.950

To further test each phonological level, Table 3.6 presents the objective results comparing the proposed units matched against reference unseen utterances for 3 speech databases. These are chosen for increasing prosodic complexity — i) Read speech (speaker SLT, [Kominek and Black, 2003]), ii) Radio News (speaker F2B, [Ostendorf et al., 1996]) and ii) Audiobook task, (The adventures of Tom Sawyer, Blizzard '12 Annual Speech Synthesis challenge). The root mean squared error (RMSE) and correlation with the reference contours are averaged across the test set.

The primary conclusions from this table are (i) read speech databases have predictable intonation values that statistical models seem to model well. (ii) As

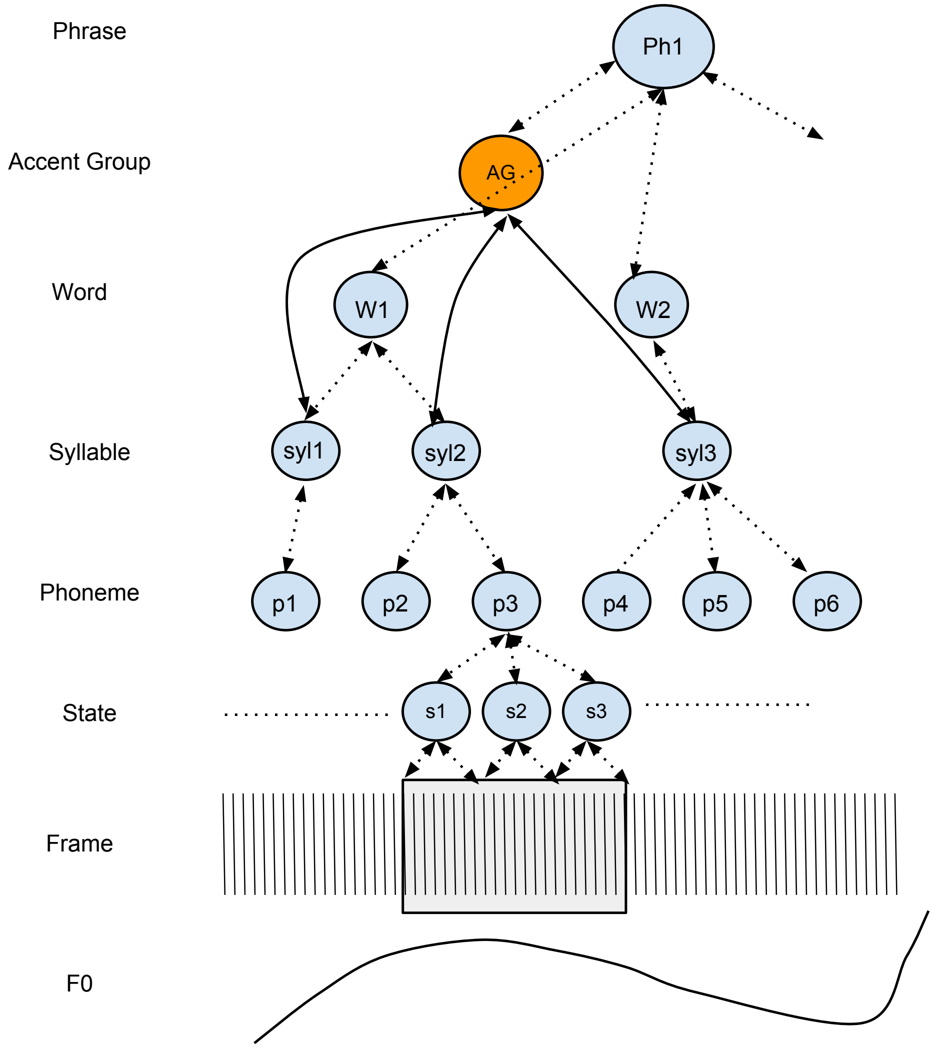


Figure 3.6: Illustration of the proposed prosodic structure in Festival. The new Accent Group relation is highlighted.

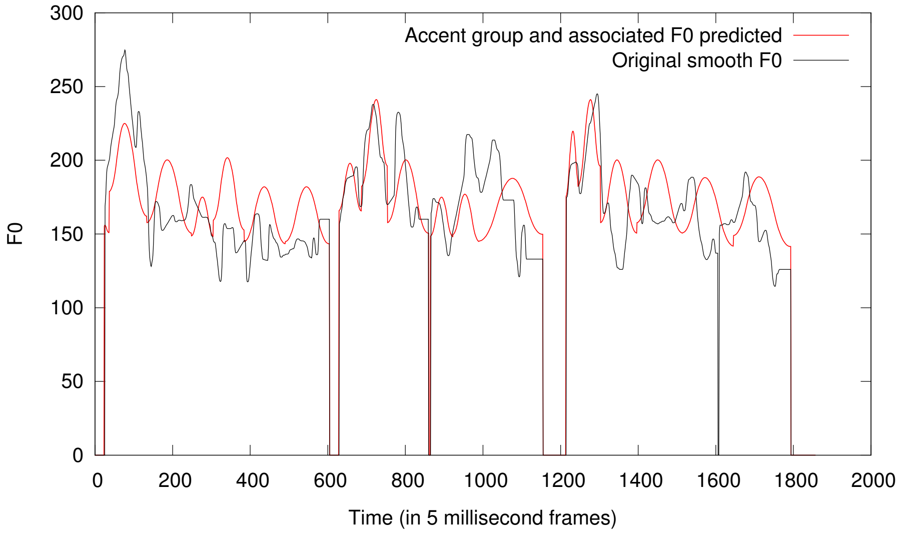


Figure 3.7: Predicted F_0 for predicted Accent Groups

the prosodic complexity increases, the default statistical models fail to capture the prosodic variance (iii) As increasingly more data is made available, models employing higher order phonological units tend to converge to similar predictions and (iv) Accent grouping is indeed a hidden part of intonation, when the true accent grouping is provided, F_0 estimates are more close to natural in all tasks— better than any other phonological unit.

3.4 Subjective Evaluation over the Mechanical Turk

As RMSE and correlation are not ideal metrics for evaluating perceptual goodness of synthetic intonation, subjective ABX listening tests on each pairs of the above models were carried out. The prosodically rich audio book task is chosen for this purpose.

One caveat with the subjective evaluation of intonation is the fact that listeners may not be sensitive to subtle differences in speech and may not be able to select one system over the other consistently. The nature of question asked during the listening test may also effect listener responses. These include questions like — *'which stimulus do you like ?' 'which voice is more understandable ?' 'which of these sounds more natural?'* etc., that are all valid aspects of intonation but could elicit

Table 3.6: Objective comparisons proposed vs. default models on three tasks

	SLT		F2B		TATS	
Unit	err	corr	err	corr	err	corr
Frame	10.97	0.62	37.22	0.38	29.95	0.079
Syllable	12.15	0.47	37.05	0.23	25.28	0.066
Word	12.65	0.46	36.30	0.33	25.80	0.0810
Accent Group	13.13	0.43	35.79	0.33	25.96	0.064
Accent Group Oracle	11.49	0.51	35.50	0.34	24.91	0.092

different listener responses. In all listening tests in the current work, we ask the question *'which of these stimuli do you prefer to hear ?'* assuming that it captures most of the desirables of good intonation in a synthetic voice. The challenge of reliability of listener preferences can be overcome by taking into account a large number of listeners, so that the preferences are both statistically reliable and generalizable accross listeners.

While we report subjective evaluation on a single dimension here (listener preference), there may still be value in a detailed inquiry into the task, stimuli and careful design setup of listening experiments [Hinterleitner et al., 2011]. Evaluation of higher level communicative functions like prosody is perhaps better done beyond the sentence level with implicit and explicit test design. Explicit tests may include questions on perceived salience of words etc where as implicit tests may include tasks like listening comprehensions with candidate synthetic voices and comparing task completion rates across listener populations using each voice. Another possibility is integration of competing TTS systems in real world dialog systems and comparing ease of use and end-user experience for each system. There may be psycholinguistic and sociolinguistic factors in the design of the latter experiments that are themselves worth investigating.

For the listening tests reported here, we synthesized a random 45 sentences from the test set. This set was synthesized by each of the candidate intonation models, all other TTS components remaining the same. The listening tests were carried out via crowd-sourcing on the Amazon Mechanical Turk [Parlikar, 2013], where listeners were asked to select the stimulus they prefer to hear. They can also choose a 'both sound similar' option. Each pair of stimuli was rated by 10 different listeners, making the following preferences reliable.

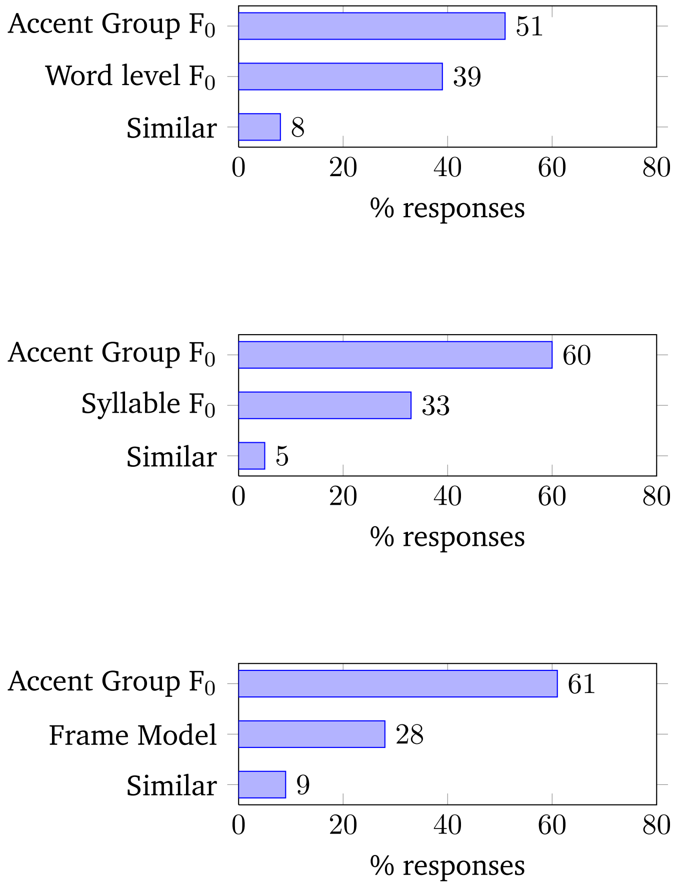
Even as these results are encouraging, the user studies also revealed more qualitative feedback about the nature of the generated speech. Since we are now able to model natural intonation, listeners are more aware of higher level aspects of laying emphasis on a wrong word or sounding like an accented speaker in the language. Interestingly, the American English we tried to model was characterized as mildly Scottish or Irish because of particular patterns generated by the model. Similarly for Tamil, it was perceived as sounding like a Sri Lankan speaker and Portuguese sounding like a particular dialect from the Alentejo region of Portugal.

It is necessary to understand here that by modeling Accent Groups, which are relatively few (only an order times more than number of words) in any given database, we are learning from insufficient data, causing bad clustering and averaging, which in turn manifests as an accidental modeling of another dialect. To address this problem of data scarcity, we propose an optimal additive strategy for F_0 in the next chapter.

3.5 Summary

This chapter presented an investigation into the optimal phonological level to model Pitch accents. A new data-derived phonological unit referred to as the Accent Group is proposed along with procedures for automatic extraction modelling as an intonation modelling unit. This is integrated into the Festival prosodic structure to enable TTS with the new intonation model. Thorough objective and subjective evaluation is conducted to show the superiority of the model compared to alternative modeling strategies.

Figure 3.8: Subjective Results: Listener Preferences for TTS with Accent Groups Vs other phonological units as the F_0 modeling unit



Chapter 4

A Multi-tier Phonological Model of Intonation

Chapter 3 proposed methods for improving F_0 modelling in SPSS by using a data-derived phonological unit, the Accent Group. In this chapter, we further improve intonation modelling by incorporating a multi-level additive strategy in F_0 generation. We motivate the idea of an additive architecture for optimal usage of the training data and propose computational techniques for automatically decomposing pitch contours into their respective components and appropriately modelling them. Objective and subjective evaluations are carried out for calibrating the performance of the approach.

4.1 Motivation for Additive Modeling

The perceptual improvements reported through the use of Accent Group are largely due to a more appropriate handling of pitch accents, that add the right prominence patterns to the underlying word sequence. Recalling Chapter 2, the two important aspects in F_0 are the pitch accents and boundary tones. In the framework detailed in the previous chapter, though the features about the phrase boundaries are incorporated into the decision tree training, they are not explicitly handled. It is also noted that higher level linguistic questions are not usually selected in favor of those within a shorter scope, causing the phrase level aspects to have minimal contribution into the model. So any aspects related to the boundary tones are only modelled by accident, and not design.

We have seen from the speech production paradigms (sec. 2.3.2) that there is an underlying long term trend to pitch, across the phrase and that the pitch accents are the local excursions that can be understood as being strung on top of the long term trend. This is seemingly advantageous from an intonation modelling perspective for multiple reasons, some of them listed here —

- *Explicit handling of boundary events* : Boundary tones are qualitatively and functionally very different from phrase-medial pitch accents. Through explicit modeling, it is expected that aspects like phrase final rises in interrogative utterances and declinations in a declarative or neutral statements can be formally induced into the model.

- *Improved explanation of variance*: None of the phonological units discussed in the previous chapter could preserve the variance of the natural intonation contours, as seen in the reduced dynamic ranges of synthetic contours (Table 3.5). This is perhaps because the statistical averaging is causing over-smoothing of the pitch accents, there by failing to effectively model the variance. Through a multi-level model, it is possible that the source of variance is more precisely modelled, by essentially distributing across different levels. This is also advantageous because appropriate richer linguistic and semantic features can be used for the higher order components, there by bringing them to effect the generated contour. The contribution of such features was noted to be only tenuous by pooling all the features together due to skewed nature of the distribution towards locally scoped features.

- *Optimal data usage* : A more practical modelling problem in the use of Accent Groups, is the reduced number of instances to train from (lesser than even the number of words available in training data, Table 3.4). Through the notion of additive modelling, it is likely that more data is available for the pitch accents to train from. To illustrate this strategy, Figure 4.1 identifies two excursions on the F_0 contour, that are qualitatively very similar, except for being at different levels of the contour. If the phrase component were subtracted, it is likely that the two accents are clustered together, than otherwise. This procedure increases the number of similar instances for the Accent Group clustering. This is beneficial both in i) providing more data for each cluster, and (ii) reducing the number of discrete shapes that phonetically characterize a speaker's intonation.

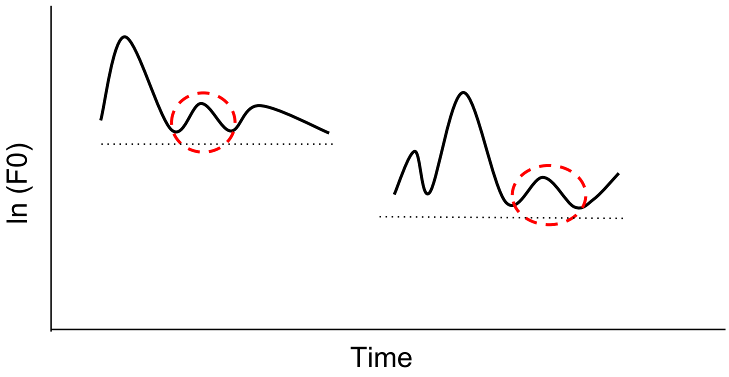


Figure 4.1: Equivalent pitch accents from two different phrases, similar in all respects but the peak amplitude, a baseline is marked for the two phrases using dotted lines.

4.1.1 What are the components ?

Under the hypothesis that the additive approach does offer a lot more to intonation modelling, the question then is what are the different components that must be considered as contributing to F_0. Starting with the Fujisaki model there have been several approaches within this paradigm. Earlier approaches have employed strategies to model pitch as a sum of multiple elementary contours that each serve a perceptual function (emphasis, attitude etc. [Bailly and Holm, 2005]); or additive contours from different phonological levels [Sun, 2002, Wu and Soong, 2012] or even completely agnostic to any underlying segmental structure [Sakai et al., 2009]. These strategies are only mildly successful because the representations of F_0 in these approaches isn't amenable to the notion of superposition.

We believe that Accent Groups as introduced in the last chapter are elegant in their representation of a pitch accent and are hence more suited to superposition. The second component is of course the long term trend, that is often modelled at the phrase level. These two conceptual levels apart, there is one other aspect to F_0, the microprosody, that is so far not properly dealt with in the current work. Microprosody refers to the jitter, the (random) segmental fluctuations in the contour that are shown to add a certain naturalness to speech. It may, hence also be considered another component.

Given these (complementary) sources, that aim to additively explain the F_0 contour, there are three issues to address for an intonation model. These are (i) Extracting these components automatically from speech data and (ii) Appropriately model these components, including prediction methods from text, and (iii) integration within the SPSS paradigm for real testing in TTS. These are elaborated in the following sections.

4.2 Data-driven Component extraction

This section presents a fully data-driven approach to component extraction from speech data. By ‘component’, we refer to each of the additive levels, phrase or the accent. A constrained, iterative procedure is employed to decompose pitch contours presented in the training data into their optimal estimates of individual long term trend (hereafter referred to as phrase) and short term excursion (referred to as the accent) components. The optimality criterion is chosen to be the objective RMSE and correlation errors of synthetic F_0 contours against reference test contours using intermediate estimates of the component models.

The iterative algorithm begins with an initialization of a phrase component. The residual after subtracting the phrase from the $ln(F_0)$ is modelled as the accent component. As an initial estimate of phrase command, the minimum value of $ln(F_0)$ over each Accent Group may be used. For each Accent Group, the residual (i.e., $ln(F_0) - phrase$) is parameterized. At this stage, to generalize over the entire training data, the following constraints are applied

- For the phrase components, at each iteration a CART tree is built to regress only from long range features, like phrase number, word number within phrase, syllable position in word etc., to the mean value of the phrase at each phoneme segment (done at the segment level for a sharper resolution and to capture microprosody, hence the multi-tier).
- For the accent components, the constraint is that the final codebook of pitch accents should be limited in number. A k-$means$ clustering is performed to identify the representative shapes of accents over all Accent Groups. Also these are forced to be predicted only from the local short range features at the level of the Accent Group.

Since the components are trained over the entire training data, they are also robust to utterance specific artifacts of the speaker or pitch detection routines. Also, these constraints are chosen to be minimally assuming and are generic across languages, speakers or speaking styles, giving the model more degrees of freedom. After the intermediate models are built (phrase CART tree and accent codebook), a new estimate of $\widehat{ln(F_0)}$ is reconstructed. The reconstruction error is added to the previous baseline and residuals are recomputed. This procedure is repeated till an objective criterion is met, here it is the minimum F_0 prediction error on an unseen set of sentences. The parameters that give the least error across the iterations are chosen as the optimal phrase and accent components. The exact procedure of this method is provided as Algorithm 2.

Algorithm 2: Constrained Component Extraction

```
for all utterances do
    for all AccentGroups do
        set $phrase$ to $min\{F_0\}$
        set $accent$ to $tilt(F_0 - phrase)$
    end for
end for
while $error \geq \epsilon$ do
    train an $accent$ codebook of size $k$ over all accents groups
    train a $codebook$ CART tree using local features
    train a $phrase$ CART tree using long range features
    for all utterances do
        Generate $\hat{F_0}$ using $phrase$ & $accent$ codebook
        for all AccentGroups do
            accumulate $error$ $(\hat{F_0} - F_0)$
            update $phrase$ to $(phrase + error)$
            update $accent$ to $tilt(F_0 - phrase)$
        end for
    end for
end while
```

For illustrating the process, one hour of speech in the Radio news genre is used and the described training algorithm is applied. Figure 4.2 presents the Root mean squared error measure of over the training utterances over the training (50 minutes) data and over an unseen development set (10 minutes).

It can be seen that the RMSE decreases over the iterations, reaching an optimum around iteration 8 for this database. The performance on an unseen development set is also consistent with the trend, suggesting the generalization of the phrase and accent models built. The same also holds for the correlation measures as shown in Figure 4.2.

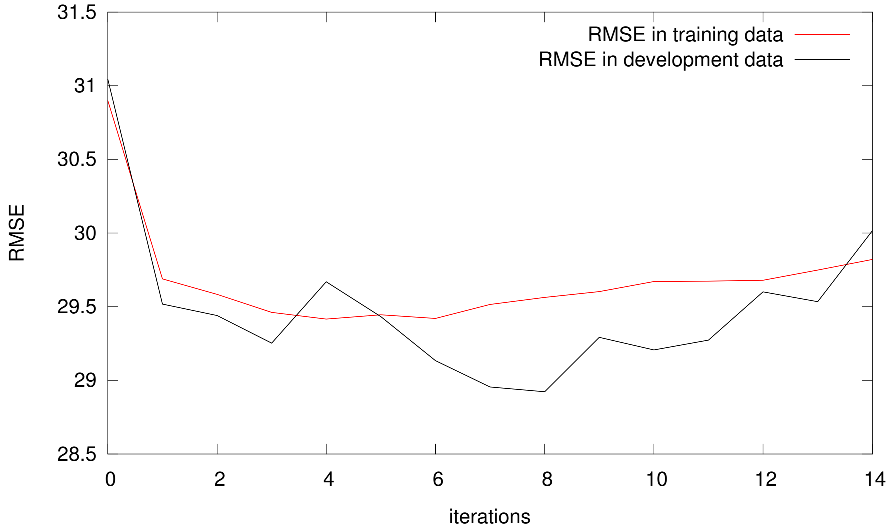


Figure 4.2: RMSE on Training and Development sets over iterations

The correlations improve over iterations during the training and similar patterns are observed over the unseen development set. These results show that over the iterations, the intermediate models improve the predictability of the respective phrase and accent components only from textual features. Figure 4.2 shows the best splits of the components on an example F_0 contour in the training data. The phrase component is set to minimum over each accent group, so that there are no discontinuities after superposition with the accent components.

Note that the model for phrase components generates a falling trend along the length of the phrase. This is consistent with established notions of declination of the contour, rooted in the physiology of speech production (Section 2.3.2). However, in the procedure described, such trends are not explicitly enforced but only emerged as a result of the described training procedure. This is quite valuable for discovering patterns within data in arbitrary tasks and languages, with minimal to no prior task knowledge.

4.3 Multi-tier F_0 modelling

As an output of the training process described in the earlier section, optimal phrase and accent models are available that can predict these components only from textual

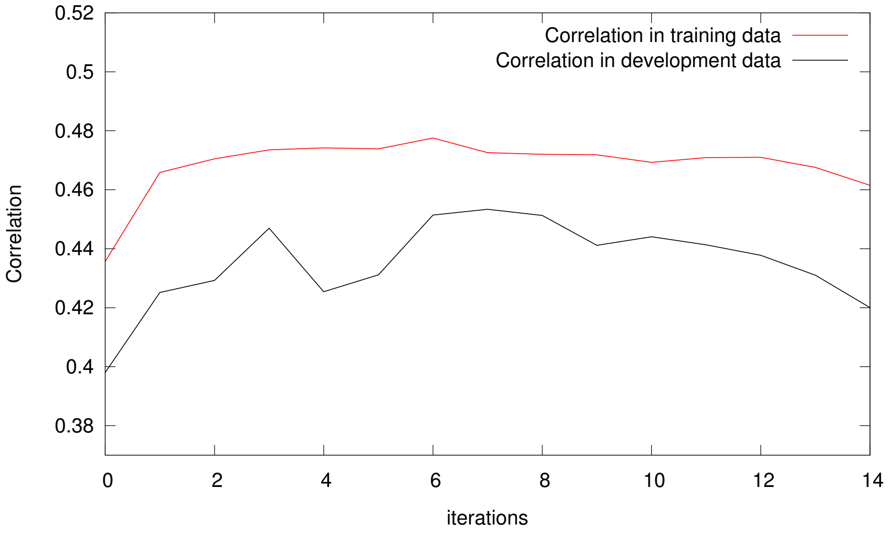


Figure 4.3: Correlation on Training and Development sets over iterations

(syntactic and contextual) features. These are integrated as the intonation modeling component in an SPSS system using a similar similar architecture described in Section 3.2.6.

The first stage in F_0 prediction is parsing an underlying syllable sequence into Accent Groups. The phrase component model is traversed for each phoneme segment to predict the long term trend over each phrase. The minimum over each accent group is assigned to avoid any audible discontinuities. The Accent Model is used over each accent group to predict a pitch accent appropriate for that context of the sentence. These two components are then added to synthesize the final F_0 for the sentence.

4.4 Evaluation

The evaluation of the proposed approach is carried out both objectively and subjectively. For an unseen set of test sentences from the same speaker in the radio news corpus, these are the same sentences used in the objective evaluations in Chapter 3 to allow direct comparison. The first test is of the mean and dynamic range of the synthesized contours, which is one of the motivations for the proposed multi-tier approach.

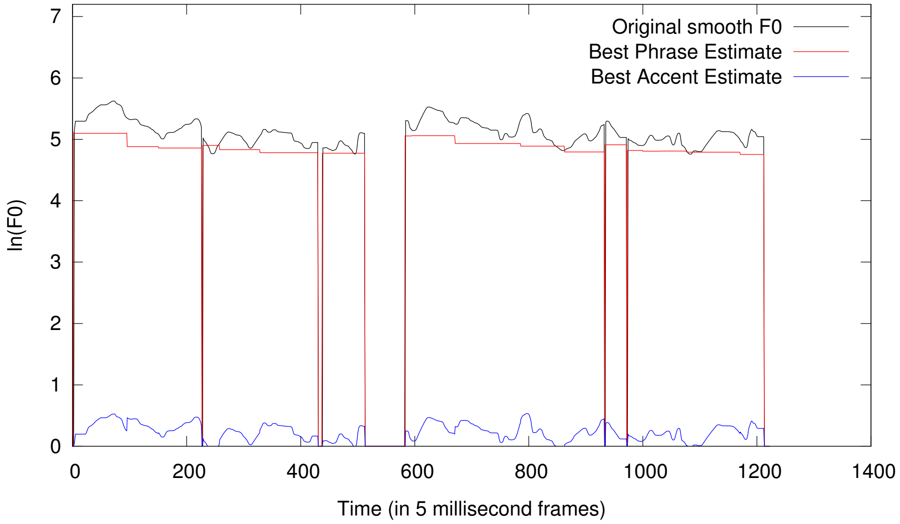


Figure 4.4: Best component splits after training

The comparable numbers for the modeling units discussed in the previous chapters are also provided. The table shows that the proposed Multi-tier approach generates contours that are most similar to the original smooth contours presented in the training data. This therefore confirms the hypothesis that the two-level modelling approach does preserve the rich variance in original intonation contours presented in the training data. Figure 4.5 shows an example synthesized contour using the described multi-tier modelling technique.

Note that the synthesized contour is indeed more varied over the speaker's pitch range compared to the strategies presented in the earlier chapter. This is without any external application of strategies like imposing global variance on synthetic contours [Toda and Tokuda, 2007]. The point-wise RMSE and correlation metrics for the synthesized contours over unseen sentences are presented for two settings in Table 4.2. The first, where intonation is completely predicted using the two-stage approach of accent group prediction followed by the application of multi-tier intonation model. The second is the case where true Accent Group boundaries from these sentences are provided.

Note that in both cases, the metrics are better than all modeling units for this database in Table 3.6. The oracle case again confirms that better accent grouping does improve the synthetic intonation significantly. Perceptual tests confirm the improvements by improved listener preferences (over 80%) to the synthesized

Table 4.1: Comparing Accent Group against other Modeling units for synthetic F_0

Modeling unit	F_0	
	Mean	Standard Deviation
Original	167.852	30.276
Frame Predicted	168.673	18.549
Syllable Predicted	175.254	16.484
Word Predicted	177.003	18.950
Accent Group	173.079	21.236
Multi-tier Accent Group	**168.77**	**26.41**

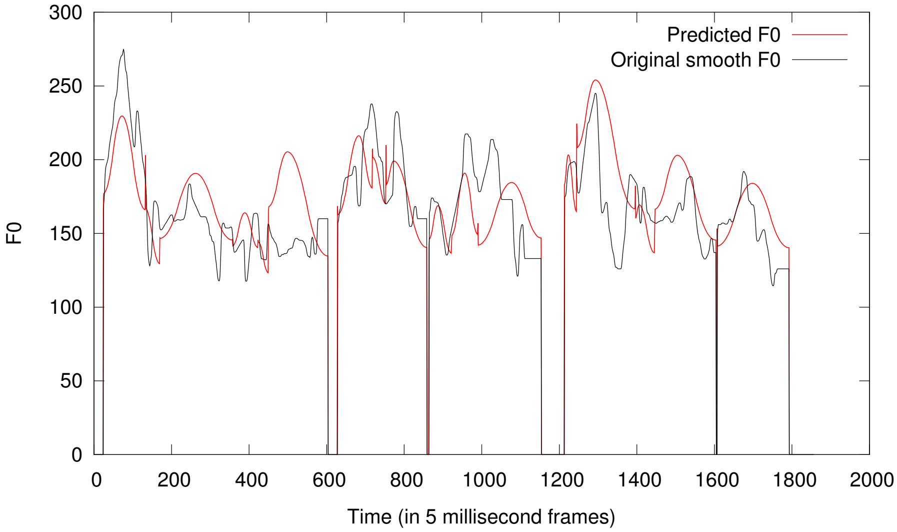


Figure 4.5: Prediction of F0 on an unseen sentence

Table 4.2: Objective comparison of synthetic F_0 on speaker F2B against references

Model	RMSE	CORR
Multi-tier Accent Group	32.0649	0.398776
Accent Group Oracle	28.9552	0.453386

contours using the Multi-tier intonation model.

4.5 Summary

This chapter presented a multi-tier strategy to improve the intonation models presented in the earlier chapter. By effectively modeling F_0 in two different layers, the variance of pitch contours is preserved in the modelling. We are also able to show improved objective measures, bringing the synthetic contours closer to the original. These improvements are also translated to perceptual preferences over a variety of tasks and speakers.

Part III

Applications of the SPAM framework

Chapter 5

Audiobook Synthesis

In this chapter, we will briefly look at aspects of prosody for Audiobook synthesis. We begin with a perceptual experiment to motivate the need for higher levels of processing for Audiobook synthesis. We later discuss the applicability of the SPAM intonation modelling framework for generation of high quality expressive voices from Audiobooks.

5.1 Empirical analysis of prosody in context

It is widely acknowledged that the factors that effect the prosody of an utterance exist in its entire context, even those beyond the sentence [Riester and Baumann, 2011]. We describe here a perceptual experiment, conducted in [Hovy et al., 2013] to study the effect of context in speech. Within the scope of this controlled experiment, we consider one previous sentence as the context of the target stimulus sentence.

We designed a sentence set that consists of about 100 sentence pairs: (i) a target sentence, and (ii) the previous sentence as its context. The source of this text is the Brown corpus [Francis and Kucera, 1971]. The sentences are balanced for genre and are selected randomly. The text is processed, tokenized, and sentences were chosen to be between 10 and 15 words.

Figure 5.1 shows an example of three sentences, a stimulus in isolation I, the context sentence C, and the stimulus in context of its previous sentence, denoted as C. The expected 'focus' word is underlined in the stimulus as the word likely to be perceived salient in its context.

I: “ What is your experience with **autistic** children ? ”

C: I try to give him as many normal experiences as possible.
S: “ What is **your** experience with autistic children ? ”

Figure 5.1: Example context (C) and stimulus sentence (S), both in isolation and in context. Expected prominent word underlined

Speech recordings of this set of 100 sentence pairs are collected from a voice talent. The sentences were recorded by a female graduate student who speaks standard American English. She was made aware of the purpose of the recordings as being the study of the phenomenon of prosodic focus, but was asked to deliver the sentences naturally. Recordings were performed in two settings. In the first session, we presented the sentence pair and recorded both the context sentence as well as the intended stimulus sentence together as a single utterance. The speaker was allowed to read the sentence pair ahead of the recording to make her aware of its context and allow her to plan accordingly. In a second session (after a few days), the speaker was presented only the stimulus sentence to record.

To study the difference between the recordings of the same sentence, spoken within context and in isolation, we conducted analysis on the pitch contours of the in recordings C, I and S.

To study the explicit effect of context on F_0, we measure the following global parameters: i) maximum value of F_0, ii) mean value of F_0, iii) mean F_0 of first content word, iv) mean F_0 of final content word, and v) dynamic range of F_0. All of these over the duration of the sentence.

Table 5.1 compares the Pearson’s correlations among various recording conditions. ‘isolated’ and ‘stimulus’ respectively correspond to the same sentence spoken without and with context; ‘context’ denotes the context sentence provided. The mean and range (standard deviation) of the F_0 for these conditions are shown.

The numbers show that the F_0 statistics are more correlated between the previous sentence and the stimulus recorded in context, as opposed to the previous sentence and the isolated recording. This implies the speaker employs systematic linear changes to the F_0 statistics when speaking in context. For the statistics corresponding to starting F_0, ending F_0 and maximum value of F_0, the averages of these values across all the utterances are shown in Figure 5.2.

Table 5.1: Correlations of F_0 mean/range for various conditions

Utterance pair compared	Correlation F0mean	Correlation F0range
isolated–stimulus	0.45	0.23
context–stimulus	0.23	0.22
context–isolated	0.13	0.13

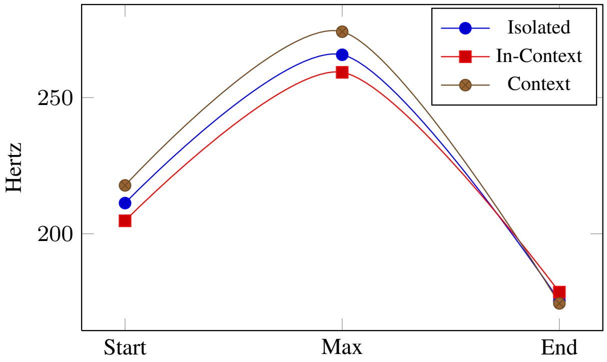


Figure 5.2: Analysis of mean starting/max/ending F_0 across for different utterances

It is clear that these F_0 statistics are consistently lower for sentences spoken in context than their counterparts spoken in isolation. These changes are likely in adjustment to the context of the previous sentence.

5.2 Audiobooks for creating TTS systems

The analyses presented in the previous section suggests that there is significant changes to the intonation based on the context of the sentence. These aspects are important for analysis and generation of multi-paragraph text, like Audiobooks.

Speech synthesis has traditionally used recordings of isolated sentences for creation of synthetic voices. With the availability of techniques for processing Multi-paragraph audio [Prahallad and Black, 2011], there is now access to prosodically richer Audiobooks. These are usually recorded by professionals or interested volunteers who read out the text of an entire story. Well recorded audiobooks are quite clean and provide an expressive narration of the story.

Since the recordings are done continuously, each sentence has substantial context provided for the speaker thereby adding more variety to intonation. This is both a benefit and a challenge, because while the richness in the data can help improve synthetic voices, none of existing algorithms are designed for optimally handle such high prosodic variance.

The other advantage Audiobooks offer is in the amount of speech provided from a single speaker. Commonly used TTS databases provide about one hour of speech from a single speaker, whereas a single Audiobook recording could provide over 10 hours of speech, depending on the length of the story.

5.3 Linguistic and Contextual Features

For the goal of generating natural and expressive speech output from text input, seemingly there should be a lot of sophisticated NLP that should go into building TTS systems. But, in practice, only rudimentary knowledge about the language goes into building TTS systems. Syntactic information (e.g., Parts of Speech) is about the most informative NLP feature that goes into the models. Table 5.2 lists some of the features (no sub-word features are listed) used in the CART tree clustering.

Table 5.2: Text & speech features used in CART training.

lexical	**acoustic**
word : content ?	word duration
word POS	position in phrase
#content words in left context	all above for neighboring words
#content words in right context	
#words in left context	
#words in right context	
#syllables in word	

5.3.1 Richer Semantic Features

As current TTS research efforts have moved from sentence level synthesis towards synthesis from audiobooks, there is a greater role for discourse level and pragmatic features, and generally richer NLP features into voice building process. Due to the availability of substantially more data through the use of Audiobooks, these features are more likely to be useful.

Current TTS systems use only syntactic information in context clustering. We've reiterated that one of the reasons for the lack of naturalness in current TTS voices is the lack of discriminative textual features causing semantically different speech regions to average out the model. There is a lot of current work in both supervised and unsupervised techniques for dependency parsing in several languages, that aim to provide semantic relations between words in a sentence. Fig 5.3 shows an example output of the Stanford dependency parser [De Marneffe et al., 2006], marking all semantic relations on the words.

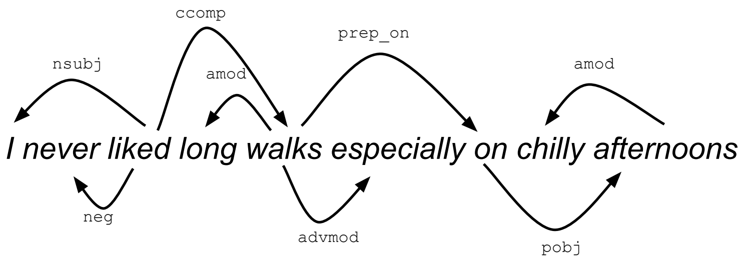


Figure 5.3: Output of a semantic parser, marking all the dependency relations on the words

It is non-trivial, however, to integrate these structures into the TTS architecture. We do this by designing features on such semantic graphs that can be used in building the SPSS models. Table 5.3 lists the features we employ to encode such information into the utterances.

Table 5.3: Dependency features on words for SPSS model training.

relation with head word	distance from root
relation with root word	number of siblings
relation with left word	number of daughters
relation with right word	boolean descriptors of the above features
distance from head	

These features are incorporated onto the words within the TTS's utterance structure, which can then be accessed at training and test time. Categorical and continuous valued questions on these features are included into the CART decision tree training for the spectral and SPAM intonation modeling.

5.4 Synthesizing from Audiobooks

Using the richer features described in the previous sections, we evaluate here the effect of providing more training data to the SPAM intonation model. The data used is an Audiobook for the story Jane Eyre, spoken by an American English female professional voice talent. The data is about 13 hours of speech. We train the intonation model as described in the last chapter. The evaluation is objectively carried out to using the Root Mean Squared Error and Correlation measures. Figures 5.4 and 5.5 show the measures on predicted intonation contours on unseen test sentences for each voice, similar in all respects except the amount of data provided for training.

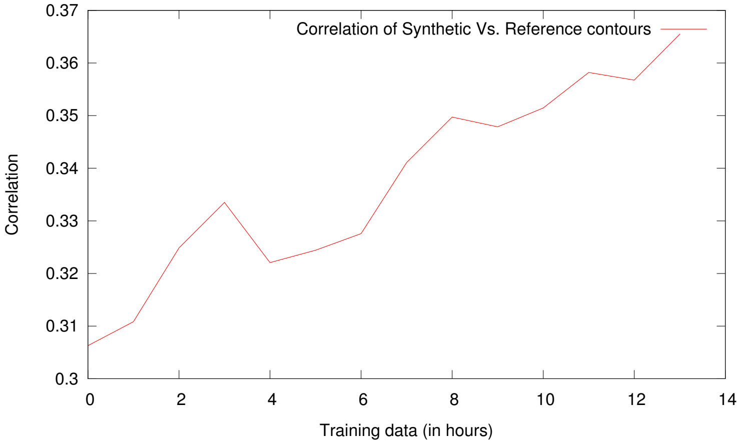


Figure 5.4: Correlation on test set using increasing amounts of training data

It can be seen that the spam intonation model improves as more training data is provided both on the error and correlation metrics, suggesting the optimal modelling of both the phrase and accent models respectively.

5.5 Summary

This chapter has presented motivation to use Audiobooks for building SPSS voices. A strategy is provided for incorporating semantic features into Text-to-Speech systems, which have only used syntactic features till date. The benefits of using increasingly more training data are presented by objective gains on predicted intonation within the SPAM framework.

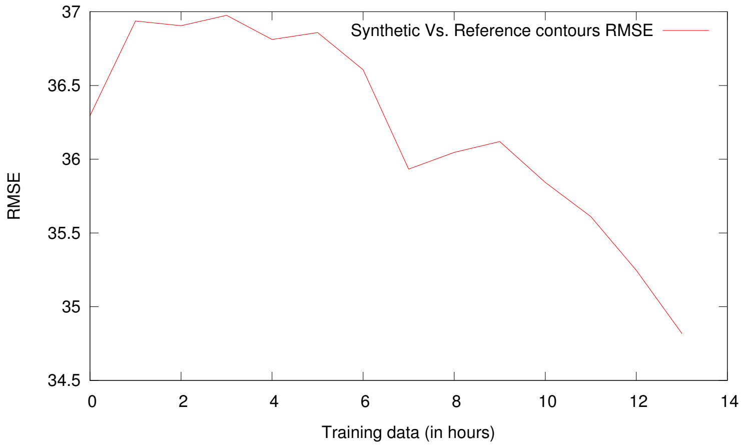


Figure 5.5: RMSE of test set with increasing amounts of training data

Chapter 6

A Style Capturing Approach to Voice Conversion

As an application of the proposed SPAM framework for intonation, this chapter presents an approach for improving voice conversion between speakers within a language by capturing the target speaker's speaking style. After a brief description of the state-of-the art in voice conversion, we motivate the proposed 2-level F_0 transformation technique [Anumanchipalli et al., 2013] for improved prosody conversion between the speakers.

6.1 F0 Transformation in Voice Conversion

Voice conversion aims to convert the speech from one speaker and make it sound like another speaker. The implications of the technology are numerous — for masking a speaker's identity for privacy [Jin et al., 2009]; for creation of synthetic voices where only a little data from a target speaker [Toda et al., 2007] or language is available [Anumanchipalli and Black, 2010].

Voice conversion has been an active research topic for over two decades with focus on source modification (energy, pitch etc.) and filter modifications (for the vocal tract). Between these, spectral transformation is more researched than source transformation [Abe et al., 1990, Stylianou and Cappé, 1998, Kain and Macon, 1998, Toda et al., 2007, Stylianou, 2009]. Many of these approaches worked on low level representations of the signal ignoring higher level aspects that are otherwise shown to be important cues to speaking style. [Zetterholm, 2006] finds that professional impersonators capture aspects of speaking style, particularly the rhythm, intonation and stress patterns across words and phrases. However, this aspect of speaking style capture has received little attention in voice conversion [Bänziger and Scherer, 2005]. [Stylianou, 2009] notes that the biggest challenge at this stage for voice conversion algorithms is the control (modeling, mapping and modification) of the speaking style of a speaker. In this chapter, we propose some directions to address this aspect of voice conversion within the SPAM framework.

Acoustically, the correlates of speaking style exist in the duration, phrasing, lexical stress patterns in words, prominence patterns, average pitch and overall pitch range. Each of these aspects is unique to a speaker, style or dialect, and intonation contributes primarily to these categorizations. Table 6.1 compares the F_0 statistics for different speakers of the Arctic read speech databases [Kominek and Black, 2003].

Table 6.1: F_0 statistics in speakers of CMU Arctic databases

Speaker	F_0 statistics	
ID	mean	std/dev
awb	132	25
bdl	128	36
ksp	133	23
rms	99	24
slt	172	27

Figure 6.1 illustrates the above by comparing the F0 contours of 5 speakers within the ARCTIC databases for the same sentence. Since speakers employ their own durations [Toth and Black, 2008] in saying the phrase, there are differences in the time axes among the speakers. It still is valid to talk about the stylistic aspects of intonation for these speakers. It can be seen that no two contours look identical (even if the times were to be normalized). The mean pitch is different for all the speakers and `slt`, the only female speaker notably has the highest F0 values. The number of peaks and the shapes of the tones are quite different for each speaker. Additionally, the contours of speakers `awb` and `ksp` are the most different from the rest of the speakers in their overall shape, and pitch accent patterns. These speakers also happen to be the only non-American English speakers within this set (`awb` is a Scottish English speaker and `ksp` is an Indian English speaker).

This illustration throws some light on the problem of intonation transformation

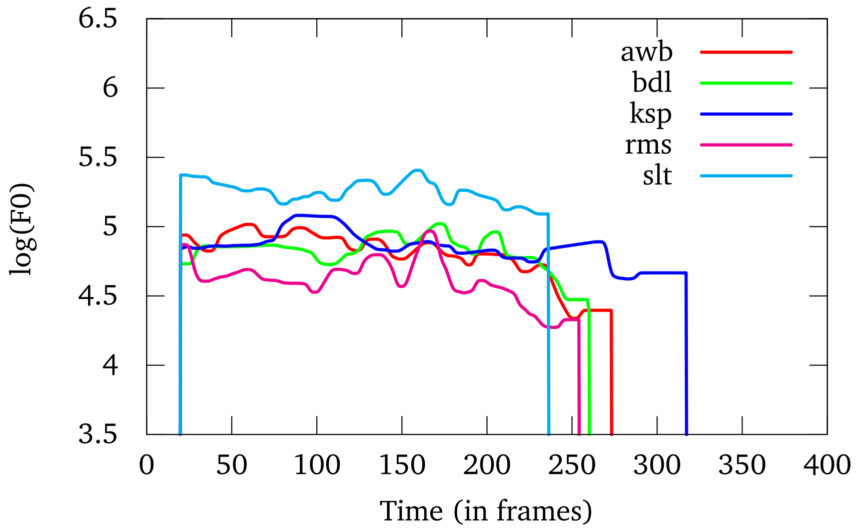


Figure 6.1: F0 contours of 5 arctic speakers for the phrase *"Will we ever forget it."*

and the challenges associated with such an attempt. Note that the example given is an excerpt from a read speech task. It can be expected that the style and range of expression through intonation is much higher for prosodically more complex tasks, like broadcast news, conversational speech or audiobooks.

Usually the problem setting of voice conversion is such that there is a large database of the source speaker's speech and a smaller set of speech recordings from a target speaker. This smaller set of target speaker's speech forms the adaptation data from which a conversion function from the source speaker's intonation patterns is derived to match the target speakers intonation patterns.

6.2 Related Work

Most voice conversion techniques, in practice, address F_0 transformation as optimizing these statistics towards the target speaker. The transformation itself is done on frame level representations of F_0 (in the order of 5-10 milliseconds) that are ill-equipped to capture prosodic phenomena that are spread over longer ranges, that of syllables, and beyond. Usually a variant of pitch range adaptation [Toda et al., 2007] is employed where the source F0 is transformed to the target speaker by employing the z-score transformation as follows –

$$F0_{(t)}^{tgt} = \frac{\sigma^{tgt}}{\sigma^{src}} \left(F0_{(t)}^{src} - \mu^{src} \right) + \mu^{tgt} \tag{6.1}$$

where $F0_{(t)}$ is the fundamental frequency at time instant t, and μ, σ denote the mean and standard deviation of F0's from the training and adaptation data for the source (src) and target (tgt) speakers respectively.

Therefore, while the mean and range averages are mapped to the target speaker, the transformation is blind to the aspects of identity and style that are spread over much larger contexts than the frame. For example, lexical stress is at the level of the syllables and prominence patterns can be explained at the level of Accent Groups. Recalling Figure 6.1, it can be easily seen that, while the above technique can only approximate the mean and range to the target speaker, it can not "move" the pitch accent. The problem is especially non-trivial when dealing with target speakers of a very different dialect as illustrated by the Indian male speaker `ksp`.

There are also approaches like multi-stage z-score transforms for the mean, baseline and topline over the F_0 contours [Gillett and King, 2003] for a higher resolution; Other interesting attempts for F0 conversion include [Helander and Nurminen, 2007], where syllable level codebooks are trained and CART trees are built to train a mapping from the source to the target speaker codebooks based on linguistic context. Inanoglu [2003] employed an utterance level codebook of intonation contours and used dynamic-time warping based "transplantation" of appropriate contours on a target utterance. Raux and Black [2003] impose intonation contours from a unit selection database for simulating emphasis. There are related techniques in the area of emotion conversion including rule-based techniques mentioned in [Schröder, 2001]; and unit-selection like data-driven techniques [Inanoglu and Young, 2009]. However emotion conversion is different from speaking style conversion in that, in the former, there is no requirement to match a target speaker, which poses more challenges. In the following sections, we present some approaches to transform a speaker's F0 characteristics to match another, which requires that the range, shape and peak positions of the target speaker are predicted only given the intonation of the original speaker.

Given the close relation between the definition of Accent Groups to pitch accents on F_0 contours (Chapter 3), the techniques in this chapter use these as the anchor units over which F_0 is analyzed for two speakers. This is rooted in the phonological hypothesis that the F0 is structured for conveying linguistic meaning of the underlying text. Our goal here is to convert an unseen F0 contour of the source speaker and predict the likely contour (with the appropriate pitch accents) that the target speaker may have employed in delivering the same sentence.

6.3 Analysis of Pitch Accents accross speakers

Using the accent group discovery procedure described in Section 3.2.2, pitch accents are analyzed for the speakers. The output of the algorithm for each utterance is a sequence of Tilt-parametrized Accent Groups that is detected on the contour.

This representation of F_0 is suitable for voice conversion in that it can parameterize the salient peaks and the overall cadence without significant loss in error and correlation, while removing unnecessary fluctuations within the contour that may not be as important in signifying speaking style.

Table 6.2 presents the mean error and correlations for natural and resynthesized contours of the speakers in the arctic database. It is rewarding that the same contour can be represented in about half or lesser number of parameters with minimal reconstruction loss, going from syllable to Accent Group level representations, in addition to giving a simple parametrization of the speaker style.

Table 6.2: Errors and correlations on resynthesized contours using different representational levels

Speaker	Syllable		Accent Group	
label	RMSE	CORR	RMSE	CORR
awb	0.13	0.77	0.14	0.73
bdl	0.09	0.82	0.12	0.76
ksp	0.08	0.79	0.11	0.73
rms	0.10	0.80	0.14	0.72
slt	0.07	0.73	0.09	0.69

For the current purpose of voice conversion, the main idea is to model any systematic way in which the nuclear accent (the peak) moves about from the source to the target and how the shape of the accent transforms over the Accent Group. The Accent Group discovery algorithm 3.2.2 is employed on the source speaker and the intonationally atomic units are detected. It should be noted that different speakers may have a different set of Accent Groups they choose to employ. For the arctic speakers, on an average, there is only about 38% of the Accent Groups matching per utterance for a random speaker pair. So, it is not easy to get parallel data with this setting. We deal with this problem by 'force aligning' the source speaker's Accent Groups on the target speaker, so an analysis is carried out on target speaker's speech within the same linguistic context, so that there is an alignment of the number of Accent Groups analyzed in each utterance pair. The contours over each Accent Group are then parameterized using the Tilt representation, which stores the peak, the total length over the contour, duration and the tilt shape parameter of the accent for both the source and target speakers. The corresponding Tilt parameterized vectors of each foot form the parallel data, from which to model a transformation. Table 6.3 shows the correlation matrix for the shape parameter(tilt) of corresponding Accent Groups in each speaker pair. Note that this matrix is not symmetric because the Accent Group boundaries vary as a different speaker is chosen as the source. It still is satisfying that there is a small but positive correlation between almost all speaker pairs on the shape parameter.

Table 6.3: Correlation matrix for the tilt shape parameter among speakers for corresponding Accent Groups

	awb	bdl	ksp	rms	slt
awb	1	0.139	0.333	0.293	0.254
bdl	0.155	1	0.290	0.244	0.250
ksp	0.336	0.218	1	0.301	0.260
rms	0.256	0.202	-0.01	1	0.213
slt	0.230	0.162	0.137	0.158	1

6.4 Style Capturing Voice Conversion

Chapter 3 has argued for Accent Grouping as being a speaker specific trait which is confirmed by the better objective performance in the oracle tests. Given this, a style-sensitive voice conversion technique must comprise 2 parts — i) Accent Group conversion, where the goal is to learn the systematic way in which the target speaker chunks an underlying syllable sequence to lay pitch accents on and ii) F_0 conversion, the conversion of the pitch accents of the source speaker to the target speaker.

The SPAM intonation model represents $ln(F_0)$ as a sum of the phrase, that models the long term trend of the contour and accents that model the local detours. Being the long term trend, the phrase component, by design, is comparable across speakers within a language. The interesting detail is in the Pitch accents that needs

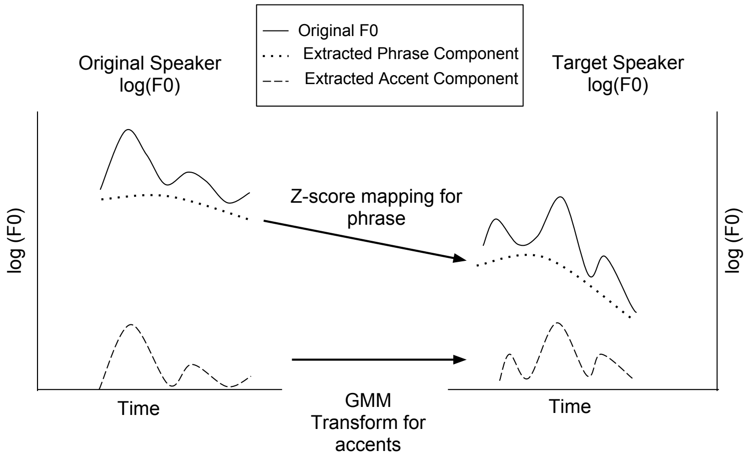

Original F0
Extracted Phrase Component
Extracted Accent Component
Original Speaker log(F0)
Target Speaker log(F0)
log (F0)
log (F0)
Z-score mapping for phrase
GMM Transform for accents
Time
Time

Figure 6.2: Proposed 2-stage F_0 transformation in Voice Conversion

to be handled more appropriately. Figure 6.4 illustrates the proposed F_0 conversion framework within the SPAM model.

We use the z-score transformation technique as shown in Equation 6.1 on the phrase contour because the mean pitch is roughly determined in the phrases and an affine transformation is sufficient to approximate the target speaker's phrase components. Accents are however complex since they are described by many aspects like the peak position, shape etc., that manifest as the speaker style. To transform accents, we train a mapping function between the two speakers' accent vectors using parallel data as described below.

The goal of the mapping that we learn from the parallel data of accent vectors is to apply it to accent shapes of an unseen utterance of the source speaker, and predict corresponding shapes of the target speaker. To accomplish this, we use the Gaussian mixture model(GMM) Joint density modelling technique, often used for spectral conversion [Stylianou et al., 1995]. The conversion can be realized by a continuous mapping based on soft clustering of the parallel accent features [Kain, 2001] for the source and target speakers.

Let x_t and y_t be the TILT accent vectors for corresponding Accent Groups in the source and target speakers. The joint probability density of the source and target vectors is modelled as the following GMM —

$$P(z_t|\lambda^{(z)}) = \sum_{m=1}^{M} w_m \mathcal{N}(z_t; \mu_m^{(z)}, \Sigma_m^{(z)}) \tag{6.2}$$

where z_t is the joint vector $\begin{bmatrix} x_t' \\ y_t' \end{bmatrix}$, with the GMM having M mixtures with a mean, covariance and mixture weight of the m'th Gaussian component denoted by w_m, $\mu_m^{(z)}$ and $\Sigma_m^{(z)}$ respectively. The Covariance matrix $\Sigma_m^{(z)}$ is constrained to be of the form $\Sigma_m^{(z)} = \begin{bmatrix} \Sigma_m^{(xx)} & \Sigma_m^{(xy)} \\ \Sigma_m^{(yx)} & \Sigma_m^{(yy)} \end{bmatrix}$, where each partial covariance matrix is set to be a full matrix, because some Tilt parameters (duration and tilt amplitude) have positive correlation between themselves Taylor [2000b].

At test time, given an accent shape x_t of the source speaker, the goal is to predict the corresponding corresponding y_t of the target speaker as follows –

$$\begin{aligned}
\hat{y}_t &= \sum_{i=1}^{M} p(m_i|x(t), \lambda^{(z)}) E(y_t|x_t, m_i, \lambda^{(z)}), \qquad (6.3) \\
E(y_t|x_t, m_i, \lambda^{(z)}) &= \mu_i^{(y)} + \Sigma_i^{(yx)} \Sigma_i^{(xx)^{-1}} (x_t - \mu_i^{(x)}), \qquad (6.4) \\
p(m_i|x(t), \lambda^{(z)}) &= \frac{w_i \mathcal{N}(x_t; \mu^i, \Sigma_i^{(xx)})}{\sum_{j=1}^{M} w_j \mathcal{N}(x_t; \mu_j^{(x)}, \Sigma_j(xx))} \qquad (6.5)
\end{aligned}$$

6.5 Evaluation

To evaluate the proposed transformation, we select speakers `awb`, `ksp`, `slt` and `rms` of the Arctic databases. SPAM intonation models were trained on the training data (90%) for each speaker. For each selected speaker pair, a transformation data of 200 sentences (about 12 minutes of speech) is randomly selected. The source speaker's intonation is analyzed as described in Sec 6.4. Since the transformation data is relatively small, a phrase CART tree cannot be trained for the target speaker, so the phrase model of the source speaker is used on the target speaker's utterance to predict a possible phrase contour. The phrase contour is shifted along the $log(f0)$ axis such that the residuals are all non-negative with a minimum at 0. For each Accent Group, the accent residual of the target speaker is also analyzed within the same linguistic context, to obtain a parallel set of accents for the speaker pair. GMM Joint densities are trained and the mapping function described in the previous section is computed. Also the means and standard deviations of the phrase components are computed to learn a z-score transformation for the phrase components of the two speakers.

For a test set of 100 sentences, the source speech is analyzed, the Accent Groups extracted and parameterized. The durations are modified in the parameterization to match those of the reference speech of the target speaker. This is done so as to be able to objectively compute the root mean squared error (`rmse`) and correlation (`corr`) metrics for each utterance. The phrase model of the source speaker's SPAM intonation model is used to predict a phrase curve over the source speaker, and the phrase level z-score transform is applied to estimate an approximate phrase contour for the target speaker. GMM transform is applied on the Tilt parameterizations of the accents over each Accent Group, to predict the possible accent shapes of the target speaker. Resynthesis of the transformed parameters is done and added with the transformed phrase contour to predict the F0 contour for the target speaker for the durations he employed. As a baseline to compare against, we use the traditional z-score mapping directly on the $ln(f0)$ contour – the resynthesized parameters of the source speaker for the durations of the target and the result mapped to the mean and range of the target speakers $ln(f0)$.

The predicted contours of the baseline and the proposed approaches are evaluated against the reference target speaker $ln(f0)$s, using RMSE and correlation measures. Table 6.4 compares the averages of these measures over the test set for several speaker pairs. All statistically significant differences in correlation are shown in bold font. It can be seen that the correlation of the transformed contours of the proposed approach are consistently improved compared to the baseline z-score mapping on the F0 contour.

6.6 Summary

In this chapter, we have successfully demonstrated the usefulness of the SPAM framework for Voice Conversion. A phonologically sensitive approach for F0 transformation is proposed. Corresponding Pitch accents of two speakers speaking the same underlying text are extracted and parameterized using the SPAM intonation model. A Gaussian mixture model based mapping is trained between the parametrized accent vectors. This mapping is used to convert unseen contours of utterances of the source speaker to predict the likely contour of the target speaker.

Table 6.4: Objective comparison of frame level z-score transformation and GMM transformation of Accent Group vectors

Speaker	Z-score transform		Foot based	
pair	RMSE	CORR	RMSE	CORR
bdl-slt	0.494	0.377	0.466	**0.521**
bdl-ksp	0.264	0.450	0.289	**0.526**
bdl-awb	0.305	0.528	0.310	**0.647**
bdl-rms	0.593	**0.461**	0.421	0.405
ksp-bdl	0.324	0.557	0.312	0.556
ksp-slt	0.470	0.423	0.438	**0.505**
ksp-rms	0.493	0.339	0.697	**0.513**
ksp-awb	0.334	0.561	0.304	**0.631**
rms-bdl	0.216	0.565	0.238	**0.590**
rms-slt	0.628	0.247	0.443	**0.487**
slt-bdl	0.638	0.465	0.350	**0.491**
slt-rms	0.915	**0.531**	0.475	0.307

Objective evaluations show that the method is better than the baseline frame-level z-score mapping technique for F0 conversion.

Chapter 7

Intent Transfer in Speech Translation

We have so far seen methods to model and transform intonation. We have noted that intonation cannot be entirely predicted from text or completely transformed to a target speaker is due to the lack of sufficient input information and the inherent variance in human speech. Speakers have a large variability and freedom to emphasize any concept they choose to, for which there are no cues in text. These form the ‘augmentative’ and ‘affective’ parts of prosody, the extra information conveyed through intonation to ensure that the intended message is decoded by listeners [Taylor, 2009, Chapter 6]. In the rest of this chapter, we use the term ‘intent’ to broadly refer to such aspects of speech.

While intent in its true sense is represented in the brain, likely in a language and speaker generic form, humans effectively use speech as the mode of communicating their intent to a listener. Intent manifests in all aspects of speech production right from the word choice, word ordering and also includes the the non-linguistic content in the speech signal through which the speaker employs prosodic devices like focus. Such aspects of speech intent are not merely stylistic choices, but also disambiguate between different semantic interpretations of a sentence. Prominence, for example, helps prioritize the concepts presented in the sentence by laying different levels of emphasis on each content word. For a TTS system to generate speech as good as a human, it needs to perform sophisticated analysis of the text to ‘understand’ the intent to then predict and impose the desired attributes in the generated speech signal. There is not enough richness in text, nor are NLP methods advanced enough to completely recover intent from it.

There are, however, certain domains where information about intent may be accessible to the speech synthesizer like speech translation. In this chapter, we propose to apply the SPAM framework and transformation techniques for the task of speech-to-speech machine translation (S2SMT). The goal of S2SMT systems is to take speech in one language as input and generate as output, the same sentence spoken in another language. Since the speech in the original language is available, intent may be analyzed and appropriately laid onto the target language synthesis.

7.1 Speech-to-Speech Machine Translation

Traditional approaches to S2SMT use the pipeline architecture where speech from a source language is passed through an automatic speech recognizer (ASR). The ASR output is translated to a target language using a statistical machine translation (SMT) system. The translation output is passed on to the TTS system to synthesize the translated text in a target language. Since all the component technologies are very much under development and are fragile in the real world, S2SMT systems have not yet become commonplace. This is partly due to the errors that each system contributes but also due to the cumulative loss of information along the pipeline. In this chapter, we propose tighter integration of the TTS system with the ASR and SMT components to improve the information sharing and overall performance of S2SMT.

While there has been considerable work in the ASR, SMT components and tightening the interface between the two to improve speech translation [Al-Onaizan and Mangu, 2007, Bertoldi et al., 2008, Wolfel et al., 2008], issues for speech synthesis within this framework remain to be studied ([Aguero et al., 2006, Parlikar et al., 2010]). Previously prosody in the source side has been used to improve the performance of the ASR system for verifying different linguistic hypotheses [Noth et al., 2000]. There is also work in cross-lingual conversion of spectral information that can be exploited to match the original speaker's voice after translation [Wu et al., 2009, Anumanchipalli and Black, 2010, Kurimo et al., 2010].

In this chapter, the goal is to further exploit the source prosodic information by imposing it appropriately on the target side after translation, in essence transferring the intent across in S2SMT. Transferring the word prominence from the source language utterances to the synthesized utterances in the target language is a formidable challenge requiring integration of several techniques within speech analysis, speech recognition, machine translation and speech synthesis frameworks. Figure 7.1 situates the current problem within the framework of speech translation. We address this problem by learning how prosodic correlates of prominence patterns

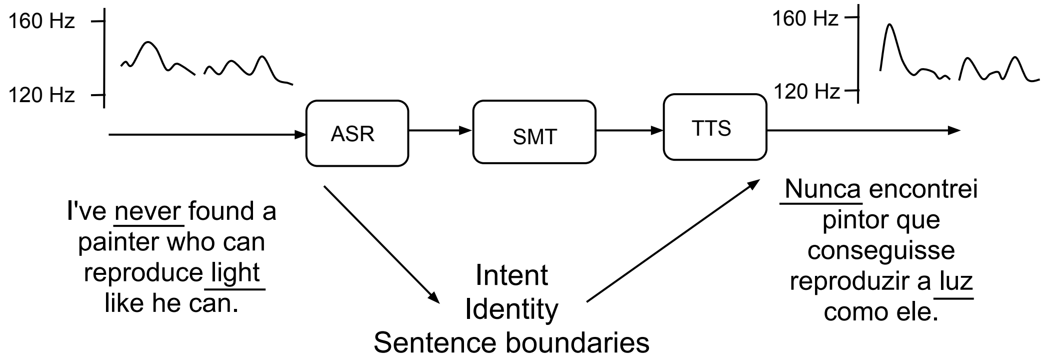

160 Hz
120 Hz
ASR
SMT
TTS
160 Hz
120 Hz
I've never found a painter who can reproduce light like he can.
Intent
Identity
Sentence boundaries
Nunca encontrei pintor que conseguisse reproduzir a luz como ele.

Figure 7.1: Schematic illustration of the proposed prominence transfer within the spoken language translation framework.

change across languages, considering the case of English↔Portuguese translation in this work. For this, We have created a database of parallel speech in these languages. Automatic word alignment information and accent prediction techniques are used to map the prominence patterns across this language pair.

While beyond the scope of this work, Fig. 7.1 also shows other outstanding problems at the source-target interface, i.e., speaker identity and sentence boundaries (relevant for better audio-visual synchronization in automatically dubbed videos). In this chapter, we only deal with transfer of speaker intent as conveyed through F_0. We address this problem by learning how the intonational correlates of focus change across languages, considering the case of two language pairs for translation in this work. Our results show that the approach effectively transfers the word prominence patterns cross-lingually.

We reiterate that correlates of focus also exist in the energy, duration and phrasing patterns around the associated concepts. In this work however, we deal only with the intonational aspects, manifested as appropriate pitch accents to convey word focus.

7.2 Parallel Speech Corpora

In order to learn the right mapping from the source language intonation to the target language, the requirement is the availability of parallel speech corpora. This is similar, in spirit to parallel text that is used in statistical machine translation [Koehn et al., 2007]. In the case of speech however, the term “parallel” needs to be elaborated. Within the scope of this thesis, we consider speech recordings of semantically identical sentences spoken with the same intent and level of expression. Such resources are not readily available in speech. It is also challenging to setup control for the requirement of similar prosodic expression in both languages. In this section, we describe the data and the process to build such a parallel speech corpus for the English-Portuguese language pair.

As the text corpus from which to record, we use *UP*, the in-flight magazine of Portugal’s national carrier, *TAP* airlines. The magazine has parallel articles, parallel at the paragraph and sentence levels, on a variety of topics including travel, cuisine, art and culture. This ensures a good coverage of proper nouns and syntactic constructions in the recording prompts, suited for training high-quality natural-sounding TTS systems. From a vast collection of articles, an optimal set of paragraphs (optimized for phonetic coverage) is chosen to be recorded by a native Portuguese speaker fluent in both English and in Portuguese The choice of recording at the paragraph level was deliberately made to give the speaker enough linguistic context for employing natural prosody, which is otherwise difficult to elicit in sentence level read speech recordings. The recordings were done alternatively for each paragraph, first in Portuguese and then in English, so that the speaker is likely to employ the same intent in the two languages. However the speaker is not given explicit instructions to maintain the same focus/prominence patterns in the two instances of recordings.

These paragraph level utterances are automatically chunked at the sentence level and are phonetically segmented using the `islice` module [Prahallad and Black, 2011] within Festvox voice building suite. The duration of the speech is approximately 1 hour in each language. The corpus statistics are as presented in the Table 7.1.

Additionally, we also present results of automatic focus analysis on an English-German (`en-de`) parallel speech database generously provided by the EMIME project [Kurimo et al., 2010]. The statistics of this corpus is presented in Table 7.2.

The EMIME databases are read speech recordings at the sentence level. Note

Table 7.1: Statistics of the English-Portuguese parallel speech corpora

Language	English	Portuguese
#Paragraphs	84	84
#Sentences	420	420
#Tokens	8184	8211
#Words	2934	3283
#Tokens/Sentence	19.49	19.55
Duration(in mins)	60.36	59.47

Table 7.2: The EMIME English-German parallel speech corpus

Language	English	German
Speaker ID	GM1	GM1
#Paragraphs	—	—
#Sentences	145	145
#Tokens	1301	1198
#Words	763	697
#Tokens/Sentence	8.97	8.26
Duration(in mins)	11.68	11.87

that the #Tokens are higher in number and more comparable in the `en-pt` language pair since it is more free style magazine content, and due to the fact that German is agglutinative. Also note that the `en-de` corpus is much smaller in data size per speaker.

7.2.1 Word Alignment through Statistical Machine Translation

In comparing intonation of two speakers within a language, prosody is studied across the same linguistic entities (words/ phrases etc). On similar lines, it is necessary to determine comparable linguistic anchors for comparing prosody across languages. To study the correspondence in intonation, we obtain the mapping between the words in the source and target language sentences. We use GIZA++ [Och and Ney, 2000] tool to align the sentences within each language pair. A word alignment model trained on the parallel text in these databases, seeded with the respective Europarl data [Koehn, 2005] in these languages is used to obtain the word mappings between the languages. This word alignment information is necessary both in the analysis and synthesis phases. Note that we use word as the anchor unit for which the parallel speech analysis is carried out because word is the generic unit comparable across languages, accent group for instance is language dependent.

7.3 Cross-lingual Analysis of Intent

To empirically investigate the relevance of the current problem of cross-lingual intent transfer, in this section we report analysis on a subset of data from the `en-pt` parallel speech corpus.

7.3.1 Manual Analysis of Cross-lingual Focus

A random set of 75 sentences (about 10 minutes of speech) is chosen from the *TAP-UP* corpus and annotated for focus by a trained linguist, fluent in both the languages. The annotator was asked to mark all the focussed word(s) in each sentence. The annotations for each language were carried out in different sessions so as to limit the influence on perception of language stimuli on one another. These annotations are of explicit focus, hence could also include perceived emphasis through energy and duration cues. Table 7.3 summarizes the focus annotations in both the languages.

Table 7.3: Results of manual annotation of focus in parallel speech

Language	Total #words	focussed words	non-focussed words	#focussed/ sentence
English	1569	298	1271	3.97
Portuguese	1585	285	1300	3.8

It is no surprise that the expert annotator marked comparable number of words as focussed in either language. To further analyze how much agreement there is, in focussed words across languages, we use SMT alignments between the parallel sentences. Of the 75 sentence subset, alignments were generated only for 1110 `en-pt` word pairs. These also include many-to-one and one-to-many mappings between the words. Among the 1110 word pairs, 336 were marked with focus in the English word and 303 were marked as focussed on the Portuguese word. The intersection between the marked focussed words (focus on both words in pair) is found to be in 145 word pairs (about 48% match). This result is worse than, yet comparable to inter-annotator agreement of prominence on the same set of speech stimuli within a language [Mo et al., 2008].

It is therefore clear that there is a substantial overlap in the relative prominence across the two languages in the `en-pt` task. This number is still an underestimate, given the mis-alignments from SMT output and the tough nature of the task (paragraph level recordings, without explicit instruction to maintain similar prosody). This analysis reinforces our conjecture that word focus is similar across languages for the semantically same sentence. Hence this information, if available, must be exploited to improve S2SMT.

7.4 Learning Intonation Transformation from Bi-lingual Speech Corpora

In this section we present an approach for conversion of intonation from one language to another. Given a natural utterance in the source language, the goal here is to predict an appropriate intonation contour for the TTS system to synthesis the translated sentence in the target language.

We use the same framework described in Chapter 6, but rather than parametrizing Pitch accents over accent groups, here we use word level analysis which is more appropriate in the cross-lingual setting. Word level TILT vectors are used as the features, and word alignment information from SMT is used to create the parallel data. However, since word focus is primarily on content words, the parallel data is constructed only from the accents over content words. Also, a threshold is determined on accent probability to remove non-accented words from the parallel data. This data is used for training the conversion function between the accent vectors of the source and target language. Joint density of the source and target accents of the corresponding words are modelled as a Gaussian mixture model (described in Sec 6.4).

The trained function can be used on an novel source utterance word accents along with the translation and the word alignment information to predict an intonation contour with appropriate prominence patterns as used in the original speech. At synthesis time in the target language, the default word-level intonation models predict a Tilt vector for each word of the translated sentence. For the content words translated, the associated Tilt vector of the original utterance $x(t)$ are converted to $\hat{y}_t$, using the trained conversion function, overriding the default predicted word TILT vectors.

7.5 Evaluation

Clustergen synthetic voices are built for all the databases within the `en-pt` and `en-de` parallel speech data. Each voice has CART tree models for spectral and duration information. The word-level TILT intonation models are used as the intonation models. These voices are used as the baselines to compare the proposed method against. Essentially, the baselines are standard state-of-the-art TTS systems that only use the text input of the translated sentences.

As the test data, we set aside 10% of the sentences in the target language. We try to objectively measure the distance between the predicted intonation contours for the translated sentences from the reference intonation contours of the test set. We use the conventionally used root mean squared error (`rmse`) and correlation (`corr`). To enable this, the same durations as employed in the reference sentence are employed in synthesis of the test set.

As the proposed intonation model, we use a fusion of the predicted word level intonation model and the transformation model using the joint density GMM on the source utterance accent vectors. For all the function words in the translated sentence, the default predicted word level contour is retained. For the content words, the default is contour linearly interpolated with the transformed intonation contour with a simple mixing weight as given by,

$$F0_{\text{fused}} = (\phi)F0_{\text{w}ordtilt} + (1-\phi)F0_{\text{G}MMvc}$$

where $F0_{\text{w}ordtilt}$ is the default word level predicted intonation contour, $F0_{\text{G}MMvc}$ is the contour after applying the conversion function on the source utterance's accent vectors and $0 \leq \phi \leq$ is the interpolation weight. This was empirically determined to be 0.6 on a development set across language pairs. The fusion is done to improve the coherence of intonational accents, that could otherwise get effected if the conversion method is directly used, since the technique context insensitive. Table 7.4 compares the proposed and the baseline intonation contours using the `rmse` and `corr` measures.

It can be seen that the proposed method generates intonation contours much closer (lesser `rmse` and higher `corr`) to the reference than the baseline prediction

Table 7.4: Objective comparison of synthesized F0 contours

Lang Pair	Frame-based		SPAM-based	
	rmse	corr	rmse	corr
en-pt	17.60	0.51	16.59	0.54
pt-en	15.90	0.47	15.30	0.49
en-de	11.93	0.54	10.98	0.51
de-en	10.27	0.46	10.17	0.46

that doesn't exploit the source language prosody. It is also consistently effective in all language pairs, although the degree of improvement is understandably different. To further illustrate the performance of method proposed, Figure 7.2 shows the predicted intonation contours for three differently emphasized input Portuguese utterances of the sentence *'A lanterna é uma boa invenção'*. The three utterances are varied in which word, the emphasis is laid from among the three content words. In this illustration, the same durations of the baseline system are used across the three utterances for better visualization.

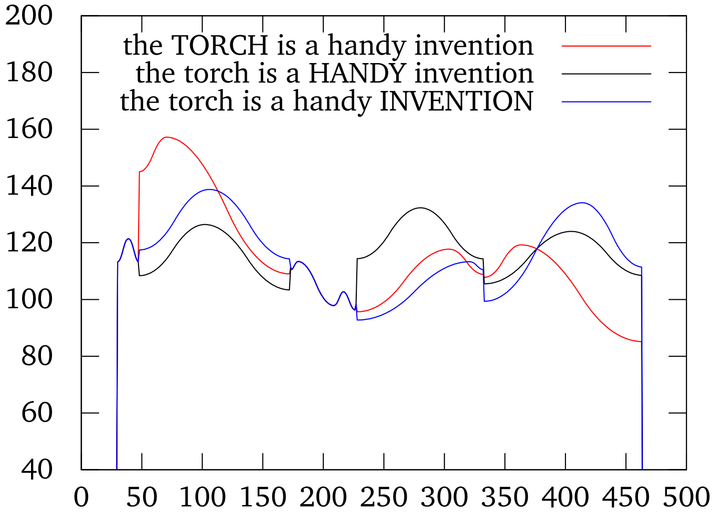


Figure 7.2: Synthesized F0's for differently focussed inputs of the Portuguese sentence *'A lanterna é uma boa invenção'*

It can be seen that the synthesized intonation contours in English are also varied to reflect the same prominence patterns as the input. This is quite elegant compared to default TTS systems that invariably produce the same intonation contours for all intents of the underlying text.

7.6 Summary

We have seen in this chapter an approach to transfer prosodic aspects accross languages within the SPAM framework. Starting with an analysis to motivate the need for Intent transfer, we have shown that there is enough correlation of prosodic aspects accross languages when semantically the same sentence is being uttered in two languages. The transfer itself uses a transformation function trained on a parallel speech corpus recorded by a bilingual speaker, exploiting automatically obtained word alignment information. The approach is evaluated objectively on unseen target language sentences, given their natural speech productions in the source language. It is clear that synthesis for an expected intent of the sentence is better when the same utterance in a source language is made available at synthesis time.

Part IV

Outlook

Chapter 8

Conclusion and Future Work

This thesis has explored a computational framework for modelling and transforming intonation. A statistical phrase/accent model is proposed for describing F_0 contours, consolidating complementary views of intonation into a unified model that can be automatically trained from speech data, and is also rooted in theoretical paradigms like intonational phonology.

The applicability of the framework is demonstrated for improved performance of Text-to-Speech synthesis within the framework of Statistical Parametric Speech Synthesis and in voice conversion applications for F_0 transformation between speakers and languages.

8.1 Contributions of this thesis

The techniques described in this thesis advance the state-of-the-art in Text-to-speech synthesis and Voice Conversion in the following ways —

- *Data-driven Accent Groups*: This thesis has argued for an optimal phonological unit appropriate to statistically model intonation, which we propose as the Accent Group. Through subjective and objective evaluations, we establish the superiority of the Accent Group compared to other levels like the frame, phoneme, syllable or word. Data-driven methods are proposed to automatically discover the accent groups, only using the speech data and an underlying segmental sequence.

Methods are also proposed to automatically parse an unseen sentence into its constituent accent groups. Accent Group is integrated into the Festival speech synthesis system for synthesis of intonation for improved speech outputs of Statistical Parametric Speech Synthesis voices.

- *Statistical Phrase/Accent Model*: A language-independent, multi-tier architecture is proposed with associated training and runtime algorithms for improving the naturalness of synthetic speech through appropriate modeling of variance in natural intonation contours. The method is shown to bring higher levels of contexts to bear on synthesized speech due to the appropriate decomposition of F_0 into constituent components.

  The method is shown to be optimal to for a range of prosodically diverse tasks including Radio news and Audiobooks; and for languages, reported here for English and Portuguese.

- *F_0 transformation in Voice Conversion*: As an application of the SPAM framework, we propose a more detailed method for F_0 transformation in voice conversion. The approach is shown to be objectively better at generating contours that are more correlated to those of the target speaker, than conventional techniques

- *Speech Translation*: We motivate the need for transferring speaker prosody from the source language to the target language for synthesis to ”completely” translate the speaker’s intent. We describe methods for tighter integration of the TTS system to the source language analysis and Machine Translation components, that have so far been neglected in speech translation.

  A conversion method trained on parallel speech data in the language pair is described for generation of F_0 contours in a target language, that takes into account the intonation in the source language.

8.2 Future Directions

While the framework and methods presented in this thesis demonstrate success in their respective tasks, these results are only an under-estimate of the real potential of what’s presented. This thesis only barely scratches the surface of slowly emerging paradigm of “Computational Prosody”. The following are viable lines of research that can improve or further exploit the contributions of this thesis —

- *Alternative design choices* : All design choices in this thesis, Classification and Regression Trees (CART) as the Machine Learning Model, Tilt for quantitative encoding of Pitch Accents, and GMM-Joint density estimation for transformation are only chosen for the practical ease and availability at the time of conducting this research. There is however potential in improving each of these choices, exclusively with a view to modelling and transforming intonation.

- *Synthesizing Audiobooks* : This thesis has only begun to address the issue of Synthesis for the Audiobooks task. We demonstrate improved performance of TTS systems by using more data as available from audiobooks. But the synthesis itself is, in spirit, at the level of isolated sentences. While we show the importance of beyond the sentence context for prosody, we haven't really addressed the issue of synthesizing multi-paragraph text input, like a story. This gaping void in the Audiobook synthesis can be filled by integrating the proposed techniques in this thesis with latest findings in discourse information structure and pragmatics that deal with contexts beyond the sentence.

  These are certainly logical extensions to the proposals in this thesis, where richer features can be used at the appropriate levels in the model to build synthesizers that are more context-aware.

- *Speaker and Task Characterization* : There is a need for a more comprehensive study of the models trained in the proposed SPAM framework. This can be potentially useful for characterizing the speakers and languages. For example, a boring or monotonous speaker may be described only by fewer shapes in the Pitch Accent codebook and an enthusiastic speaker may employ many more shapes. It is valuable to automatically categorize speakers depending on the model components. Additional analysis may include the nature of the shapes themselves and identifying what are the intonational aspects and what are merely idiosyncratic to a given speaker or speaking style.

  There are immediate applications for this in speaker verification, emotion detection and studying intonation of contact languages/creoles.

- *Transformations of all aspects of Prosody* : This thesis addressed only aspects of F_0 in voice conversion which constitutes only a part of speaking style. There is only limited work in addressing the other prosodic components like duration and rhythm etc. Conversion of Accent Grouping strategies is yet another dimension to this, immediately extending the work in the current thesis.

- *Improving Automatic Dubbing* : The prosody transfer methods described in this thesis are only a proof-of-concept for the idea of tighter integration of TTS with other components of the Speech Translation pipeline. These methods can be integrated to automatically dub videos from one language to another. There is however the practical requirement for time alignment of the video and the translated audio at some phonological level (sentence ?). There is more benefit in more closely integrating the TTS with the SMT systems so that an optimal translation hypothesis can be output (from among the n-best paraphrases) depending on the duration of the video and expected duration of each translation alternative.

  Another dimension to speech translation is to develop techniques for graceful degradation of the strategies in the face of ASR and SMT errors.

- *Transfer of Para-linguistic Aspects in Speech Translation* : The assumption in this thesis has been that spoken language is the same as written language, hence the assumption of sentence level synthesis in speech translation, and expectation of fluent language from the SMT output. This is, however, not true because spoken language has many para-linguistic aspects like revisions, repetitions, speech fillers and pauses, some of them functional to communication and some that are redundant. These aspects need to be further explored and appropriate translation techniques must be researched for all these aspects for making speech translation seamless on a variety of natural spoken language inputs.

- *Integration in Other Speech Applications* : While we have demonstrated the techniques on a few applications, there are still a host of others that can exploit the findings of this thesis. These include context-aware TTS in spoken dialog systems where the synthesizer adapts to the dialog state and responds accordingly. Another application is for virtual agents where an agent can entrain itself with respect to the speaking style and attitude of a user.

- *Psycholinguistic Evaluation* : Given the level of naturalness and expressivity reached by SPSS systems in this thesis, there is an immediate need for a full scale psycholinguistic evaluation comparing human responses to natural and synthetic speech stimuli. These include evaluation of Audiobooks that are naturally spoken and automatically synthesized to that of voice conversion and dubbing techniques presented here.

Appendix A

Related Peer-reviewed publications

1. G. K. Anumanchipalli, L. C. Oliveira, and A. W. Black. Data-driven Intonational Phonology. *166'th meeting of the Acoustical Society of America*, 2013d.

2. D. Hovy, G. K. Anumanchipalli, A. Parlikar, C. Vaughn, A. Lammert, E. Hovy, and A. W. Black. Analysis and Modeling of "Focus" in Context. In *Interspeech 2013*, Lyon, France, 2013.

3. S. Sitaram, G. K. Anumanchipalli, J. Chiu, A. Parlikar, and A. W. Black. Text to speech in new languages without a standardized orthography. In *8th ISCA Workshop on Speech Synthesis*, Barcelona, Spain, 2013.

4. G. K. Anumanchipalli, L. C. Oliveira, and A. W. Black. A Style Capturing approach to F0 transformation. In *ICASSP 2013*, Vancouver, Canada, 2013b.

5. G. K. Anumanchipalli, A. W. Black, and L. C. Oliveira. Accent group modeling for improved prosody in statistical parametric speech synthesis. *IEEE ICASSP 2013*, 2013a.

6. G. K. Anumanchipalli, L. C. Oliveira, and A. W. Black. A Statistical Phrase/Accent Model for Intonation Modeling. In *Interspeech 2011*, Florence, Italy, 2011.

7. G. K. Anumanchipalli, P. K. Muthukumar, U. Nallasamy, A. Parlikar, A. W. Black, and B. Langner. Improving speech synthesis for noisy environments. In *7th ISCA Workshop on Speech Synthesis*, Keihanna, Japan, September 2010.

8. G. K. Anumanchipalli and A. W. Black. Adaptation techniques for speech synthesis in under-resourced languages. In *Spoken Language Technologies for Under-resourced languages*, Penang, Malaysia, 2010.

Bibliography

M. Abe, K. Shikano, and H. Kuwabara. Voice conversion through vector quantization. *J. Acoust. Soc. Jpn. (E)*, 11:71–76, 1990. 65

P. Aguero, J. Adell, and A. Bonafonte. Prosody generation for speech-to-speech translation. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, page I, may 2006. doi: 10.1109/ICASSP.2006.1660081. 76

Y. Al-Onaizan and L. Mangu. Arabic ASR and MT integration for GALE. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–1285 –IV–1288, april 2007. doi: 10.1109/ICASSP.2007.367312. 76

M. Anderson, J. Pierrehumbert, and M. Liberman. Synthesis by rule of English intonation patterns. In *Proceedings of ICASSP 84*, pages 2.8.1–2.8.4, 1984. 18

G. K. Anumanchipalli and A. W. Black. Adaptation techniques for speech synthesis in under-resourced languages. In *Spoken Language Technologies for Under-resourced languages*, Penang, Malaysia, 2010. 8, 65, 76

G. K. Anumanchipalli, P. K. Muthukumar, U. Nallasamy, A. Parlikar, A. W. Black, and B. Langner. Improving speech synthesis for noisy environments. In *7th ISCA Workshop on Speech Synthesis*, Keihanna, Japan, September 2010. 8

G. K. Anumanchipalli, L. C. Oliveira, and A. W. Black. A Statistical Phrase/Accent Model for Intonation Modeling. In *Interspeech 2011*, Florence, Italy, 2011. 20

G. K. Anumanchipalli, L. C. Oliveira, and A. W. Black. A Style Capturing approach to F0 transformation. In *ICASSP 2013*, Vancouver, Canada, 2013. 65

G. Bailly and B. Holm. SFC: A trainable prosodic model. *Speech Communication*, 46 (34):348 – 364, 2005. ISSN 0167-6393. doi: 10.1016/j.specom.2005.04.008. 20, 47

T. Bänziger and K. R. Scherer. The role of intonation in emotional expressions. *Speech Communication*, 46(34):252 – 267, 2005. ISSN 0167-6393. doi: 10.1016/j.specom.2005.02.016. 66

M. Beckman. A typology of spontaneous speech. In Y. Sagisaka, N. Campbell, and N. Higuchi, editors, *Computing Prosody*, pages 7–26. Springer US, 1997. ISBN 978-1-4612-7476-6. doi: 10.1007/978-1-4612-2258-3_2. URL http://dx.doi.org/10.1007/978-1-4612-2258-3_2. 31

N. Bertoldi, R. Zens, M. Federico, and W. Shen. Efficient speech translation through confusion network decoding. *IEEE Transactions on Audio, Speech & Language Processing*, 16(8):1696–1705, 2008. 76

A. Black and A. Hunt. Generating F_0 contours from ToBI labels using linear regression. In *ICSLP96*, volume 3, pages 1385–1388, Philadelphia, PA., 1996. 18

A. Black and P. Taylor. Assigning phrase breaks from part-of-speech sequences. In *Eurospeech97*, volume 2, pages 995–998, Rhodes, Greece, 1997. 14

A. Black and K. Tokuda. Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets. In *Interspeech 2005*, Lisbon, Portugal, 2005. 27

A. Black, P. Taylor, and R. Caley. The Festival speech synthesis system. http://festvox.org/festival, 1998. 37

A. W. Black. Clustergen: A statistical parametric synthesizer using trajectory modeling. In *Interspeech 2006*, Pittsburgh, PA, 2006. 6, 9, 14

D. Bolinger. *Intonation and its Parts*. Stanford University Press, 1986. 13

K. E. Bouchard, N. Mesgarani, K. Johnson, and E. F. Chang. Functional organization of human sensorimotor cortex for speech articulation. *Nature*, 495(7441):327–332, 2013. 3

G. Bruce. *Swedish word accents in sentence perspective*, volume 12. LiberLäromedel/Gleerup Malmo, 1977. 17

N. Chomsky and M. Halle. *The Sound Pattern of English*. MIT Press, 1968. 4

R. Clark and K. Dusterhoff. Objective methods for evaluating synthetic intonation. In *Proc. Eurospeech 1999*, 1999. 27

C. d 'Alessandro and P. Mertens. Automatic pitch contour stylization using a model of tonal perception. *Computer Speech & Language*, 9(3):257 – 288, 1995. ISSN 0885-2308. doi: http://dx.doi.org/10.1006/csla.1995.0013. URL http://www.sciencedirect.com/science/article/pii/S0885230885700137. 21

M.-C. De Marneffe, B. MacCartney, and C. D. Manning. Generating typed dependency parses from phrase structure parses. *Proceedings of LREC*, 6:449–454, 2006. 61

H. W. Dudley. System for the artificial production of vocal or other sounds, 06 1938. URL http://www.google.com/patents/US2121142. Patent. 5

K. Dusterhoff, A. Black, and P. Taylor. Using decision trees within the tilt intonation model to predict F0 contours. In *in Proc. Eurospeech 1999*, pages 1627–1630, 1999. 21

W. N. Francis and H. Kucera. Brown corpus. *Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers. Revised*, 1971. 57

H. Fujisaki. Dynamic characteristics of voice fundamental frequency in speech and singing. In P. MacNeilage, editor, *The Production of Speech*, pages 39–55. Springer-verlag, 1983. 18

B. Gillett and S. King. Transforming F0 contours. In *Interspeech 03*, Geneva, Switzerland, 2003. 68

J. A. Goldsmith. *Autosegmental and metrical phonology*, volume 11. Blackwell Oxford, 1990. 31

S. Gooden, K.-A. Drayton, and M. Beckman. Tone inventories and tune-text alignments prosodic variation in'hybrid'prosodic systems. *Studies in Language*, 33(2): 396–436, 2009. 31

E. Helander and J. Nurminen. A Novel method for prosody prediction in voice conversion. In *ICASSP 2007*, Hawaii, 2007. 68

F. Hinterleitner, G. Neitzel, S. Möller, and C. Norrenbrock. An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks. In *Blizzard 2011*, 2011. 41

J. Hirschberg. Using discourse content to guide pitch accent decisions in synthetic speech. In G. Bailly and C. Benoit, editors, *Talking Machines*, pages 367–376. North-Holland, 1992. 21

J. Hirschberg and J. Pierrehumbert. The intonational structure of discourse. In *Proceedings of 24th Conference of the Association for Computational Linguistics*, pages 136–144, 1986. 21

D. Hirst, A. D. Cristo, and R. Espesser. Levels of representation and levels of analysis for the description of intonation systems., 2000. 21

D. Hovy, G. K. Anumanchipalli, A. Parlikar, C. Vaughn, A. Lammert, E. Hovy, and A. W. Black. Analysis and Modeling of "Focus" in Context. In *Interspeech 2013*, Lyon, France, 2013. 57

A. Hunt and A. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *ICASSP96*, volume 1, pages 373–376, Atlanta, GA, 1996. 6

Z. Inanoglu. Transforming pitch in a voice conversion framework. Master's thesis, Cambridge University, 2003. 68

Z. Inanoglu and S. Young. Data-driven emotion conversion in spoken english. *Speech Communication*, 51(3):268 – 283, 2009. ISSN 0167-6393. doi: 10.1016/j.specom.2008.09.006. 68

R. Jakobson, G. Fant, and M. Halle. *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates*. MIT Press, 1952. 5

Q. Jin, A. R. Toth, T. Schultz, and A. W. Black. Speaker de-identification via voice transformation. In *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, pages 529–533. IEEE, 2009. 65

A. Kain. *High Resolution Voice Transformation*. PhD thesis, OGI School of Science and Engineering, Oregon Health and Science University, 2001. 71

A. Kain and M. Macon. Spectral voice conversion for text-to-speech synthesis. In *ICASSP-98*, volume 1, pages 285–288, Seattle, Washington, 1998. 65

S. King and V. Karaiskos. The Blizzard Challenge 2009. In *Blizzard Challenge 2009*, Edinburgh, UK, 2009. 9

E. Klabbers and J. van Santen. Clustering of foot-based pitch contours in expressive speech synthesis. In *ISCA Speech Synthesis Workshop V*, Pittsburgh, PA, 2006. 30, 31, 34

D. H. Klatt. Software for a cascade/parallel formant synthesizer. *The Journal of the Acoustical Society of America*, 67(3):971–995, 1980. doi: 10.1121/1.383940. 5

P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit*, pages 79–86, Phuket, Thailand, September 2005. 79

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: open source toolkit for statistical machine translation. In *ACL 2007: Interactive Poster and Demonstration Sessions*, pages 177–180, Prague, Czech Republic, 2007. 78

J. Kominek and A. Black. The CMU ARCTIC speech databases for speech synthesis research. Technical Report CMU-LTI-03-177 http://festvox.org/cmu_arctic/, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, 2003. 38, 66

M. Kurimo, W. Byrne, J. Dines, P. N. Garner, M. Gibson, Y. Guan, T. Hirsimki, R. Karhila, S. King, H. Liang, L. Saheer, M. Shannon, S. Shiota, J. Tian, K. Tokuda, M. Wester, Y. jian Wu, and J. Yamagishi. Personalising speech-to-speech translation in the EMIME project. In *ACL 2010*, 2010. 76, 78

R. D. Ladd. *Intonational Phonology*. Cambridge University Press, 1996. 16

M. Y. Liberman. *The intonational system of English.* PhD thesis, Massachusetts Institute of Technology, 1975. 17, 31

F. Malfrere, T. Dutoit, and P. Mertens. Automatic prosody generation using suprasegmental unit selection. In *3rd ESCA Workshop on Speech Synthesis*, pages 323–327, Jenolan Caves, Australia, 1998. 21

J. Meron. Prosodic unit selection unit an intonation speech database. In *4th ISCA Workshop on Speech Synthesis*, Pitlochry, Scotland, 2001. 21

H. Mixdorff. A novel approach to the fully automatic extraction of fujisaki model parameters. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 3, pages 1281–1284. IEEE, 2000. 19

Y. Mo, J. Cole, and E.-K. Lee. Naïve listeners' prominence and boundary perception. In *Speech Prosody*, 2008. 81

B. Mobius. Synthesizing german intonation contours. In J. V. P. Santen, J. P. Olive, R. W. Sproat, and J. Hirschberg, editors, *Progress in Speech Synthesis*, pages 401–415. Springer New York, 1997. ISBN 978-1-4612-7328-8. doi: 10.1007/978-1-4612-1894-4_32. URL http://dx.doi.org/10.1007/978-1-4612-1894-4_32. 31

G. Möhler and A. Conkie. Parametric modeling of intonation using vector quantization. In *In Proc. 3rd ESCA Workshop on Speech Synthesis (Jenolan Caves, Australia)*, pages 311–316, 1998. 21

E. Noth, A. Batliner, A. Kießling, R. Kompe, and H. Niemann. Verbmobil: The use of prosody in the linguistic components of a speech understanding system. *Speech and Audio Processing, IEEE Transactions on*, 8(5):519–532, 2000. 76

F. J. Och and H. Ney. Improved statistical alignment models. In *ACL 2000*, pages 440–447, Hongkong, China, October 2000. 79

J. Olive, A. Greenwood, J. Coleman, and A. Greenwood. *Acoustics of American English Speech: A Dynamic Approach*. Springer Verlag, 1993. 5

M. Ostendorf, P. Price, and S. Shattuck-Hufnagel. Boston university radio speech corpus. 1996. 12, 16, 26, 38

A. Parlikar. *Style-Specific Phrasing in Speech Synthesis*. PhD thesis, Carnegie Mellon University, 2013. 41

A. Parlikar and A. W. Black. A grammar based approach to style specific phrase prediction. In *Interspeech*, Florence, Italy, September 2011. 7, 14

A. Parlikar and A. W. Black. Data-driven phrasing for speech synthesis in low-resource languages. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, March 2012. 35

A. Parlikar, A. W. Black, and S. Vogel. Improving speech synthesis of machine translation output. In *Interspeech*, pages 194–197, Makuhari, Japan, September 2010. 76

F. Pereira and Y. Schabes. Inside-outside reestimation from partially bracket corpora. In *Proceedings of the 30th conference of the Association for Computational Linguistics*, pages 128–135, Newark, Delaware, 1992. 35

J. B. Pierrehumbert. *The Phonology and Phonetics of English Intonation*. PhD thesis, MIT, 1980. Published by Indiana University Linguistics Club. 17, 31

K. Prahallad and A. W. Black. Segmentation of Monologues in Audio Books for Building Synthetic Voices. *IEEE Transactions on Audio, Speech & Language Processing*, 19(5):1444–1449, 2011. 59, 78

K. Prahallad, A. W. Black, and R. Mosur. Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, pages I–I. IEEE, 2006. 6

S. Prevost. An information structural approach to spoken language generation. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 294–301, Santa Cruz, CA, 1996. 21

S. Prevost and M. Steedman. Specifying intonation from context for speech synthesis. *Speech Communication*, 15:139–153, 1994. 21

A. Raux and A. Black. A unit selection approach to F0 modeling and its application to emphasis. In *ASRU2003*, St Thomas, USVI, 2003. 21, 68

A. Riester and S. Baumann. Information structure annotation and secondary accents. *Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena*, pages 111–127, 2011. 57

A. Rosenberg. AuToBI - A Tool for Automatic ToBI Annotation. In *Interspeech 2010*, Chiba, Japan, 2010. 18

S. Sakai, T. Kawahara, T. Shimizu, and S. Nakamura. Optimal learning of p-layer additive f0 models with cross-validation. In *ICASSP*, pages 4245–4248, 2009. 20, 47

M. Schröder. Emotional speech synthesis: A review. In *Seventh European Conference on Speech Communication and Technology*, 2001. 68

M. Schröder and J. Trouvain. The german text-to-speech synthesis system mary: A tool for research, development and teaching. *International Journal of Speech Technology*, 6(4):365–377, 2003. 14

K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. ToBI: a standard for labelling English prosody. In *Proceedings of ICSLP92*, volume 2, pages 867–870, 1992. 17

R. Sproat, A. Black, S. Chen, K. Shankar, M. Ostendorf, and C. Richards. Normalization of non-standard words. Final Report of Johns Hopkin University Summer Workshop, 1999. 7

M. Steedman. The surface-compositional semantics of english intonation. *Language*, 2013. 4, 21

K. N. Stevens. *Acoustic Phonetics*. MIT Press, 2000. 4

Y. Stylianou. Voice transformation: A survey. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 3585 –3588, April 2009. 65, 66

Y. Stylianou and O. Cappé. A system voice conversion based on probabilistic classification and a harmonic plus noise model. In *ICASSP 1998*, pages 281–288, Seattle, WA, 1998. 65

Y. Stylianou, O. Cappé, and E. Moulines. Statistical methods for voice quality transformation. In *Eurospeech95*, pages 447–450, Madrid, Spain, 1995. 71

X. Sun. F0 generation for speech synthesis using a multi-tier approach. In *INTERSPEECH*, 2002. 20, 47

P. Taylor. Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America*, 107 3:1697–1714, 2000a. 20

P. Taylor. Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America*, 107 3:1697–1714, 2000b. 72

P. Taylor. *Text-to-Speech Synthesis*. Cambridge University Press, 2009. 11, 75

T. Toda and K. Tokuda. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE - Trans. Inf. Syst.*, E90-D: 816–824, May 2007. ISSN 0916-8532. doi: 10.1093/ietisy/e90-d.5.816. 16, 52

T. Toda, A. Black, and K. Tokuda. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8):2222 –2235, nov. 2007. 8, 65, 67

K. Tokuda, T. Kobayashi, and S. Imai. Speech parameter generation from hmm using dynamic features. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 660–663. IEEE, 1995. 6, 8

K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for hmm-based speech synthesis. In *ICASSP*, volume 3, pages 1315–1318, 2000. 8

A. R. Toth and A. W. Black. Incorporating durational modification in voice transformation. In *INTERSPEECH*, pages 1088–1091, 2008. 66

J. van Santen, A. Kain, E. Klabbers, and T. Mishra. Synthesis of prosody using multi-level unit sequences. *Speech Communication*, 46(3-4):365 – 375, 2005. ISSN 0167-6393. 20

W. von Kempelen. *Mechanismus der menschlichen Sprache nebst der Beschreibung seiner sprechenden Maschine*. Stuttgart-Bad Cannstatt, 1791. 5

M. Wolfel, M. Kolss, F. Kraft, J. Niehues, M. Paulik, and A. Waibel. Simultaneous machine translation of German lectures into English: Investigating research challenges for the future. In *Spoken Language Technology Workshop, 2008. SLT 2008. IEEE*, pages 233–236. IEEE, 2008. 76

Y.-J. Wu and F. K. Soong. Modeling pitch trajectory by hierarchical hmm with minimum generation error training. In *ICASSP*, pages 4017–4020, 2012. 20, 47

Y.-J. Wu, Y. Nankaku, and K. Tokuda. State mapping based method for cross-lingual speaker adaptation in hmm-based speech synthesis. In *Interspeech 2009*, Brighton, U.K.,, 2009. 76

Y. Xu. Speech prosody: A methodological review. *Journal of Speech Sciences*, 1(1), 2012. 18

B. Yegnanarayana and K. Sri Rama Murty. Event-based instantaneous fundamental frequency estimation from speech signals. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(4):614–624, 2009. ISSN 1558-7916. doi: 10.1109/TASL.2008.2012194. 12

K. Yu, F. Mairesse, and S. Young. Word-level emphasis modelling in hmm-based speech synthesis. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 4238 –4241, march 2010. 26

H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda. The hmm-based speech synthesis system (hts) version 2.0. *Proc. of Sixth ISCA Workshop on Speech Synthesis*, pages 294–299, 2007. 9, 14

H. Zen, K. Tokuda, and A. W. Black. Review: Statistical parametric speech synthesis. *Speech Communication*, 51:1039–1064, November 2009. 6

E. Zetterholm. Same speaker different voices: A study of one impersonator and some of his different imitations. In *Intl Conf on Speech Science and Technology*, pages 70–75, 2006. 66