Intra-Lingual and Cross-Lingual Prosody Transformation

Gopala Krishna Anumanchipalli

Ph.D. Thesis Proposal

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

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

Copyright © 2012 Gopala Krishna Anumanchipalli

This work is supported by the Fundação de Ciência e Tecnologia through the CMU|Portugal Program, a joint program between the Portuguese Government and Carnegie Mellon University.
Abstract

Speech technologies have a major role to play in future generation technologies like spoken language translation, intelligent virtual assistants, gaming, language learning etc., Text-to-speech synthesis (TTS) is one of the main components in these technologies, often acting as the interface between the rest of the software and the end-user. It is incumbent on TTS researchers now more than ever to make synthetic speech output more natural and appropriate in the wide variety of spoken language application domains.

The main criticism of the state-of-the-art in speech synthesis is the lack of expression and personality in synthetic speech outputs. This proposal takes a step closer in the direction of expressive and flexible speech synthesis by improving the current prosodic models used within TTS. In preliminary work, we propose a Statistical Phrase/Accent Model (SPAM) for intonation that can generate natural and expressive intonation contours within TTS. Through subjective evaluation, the model is shown to be significantly better than currently used intonation models. Future work within intonation modeling will include empirical investigation into optimal prosodic representations like segments, syllables and higher level phonological units like metrical feet. The efficiency of the proposed model will be tested on prosodically rich and diverse tasks like and newscasts and audiobooks.

We then propose techniques for conversion of a speaker’s intonation to that of a target speaker or style. This is done within a voice conversion framework to match a target speaker’s spectral and prosodic characteristics. Further, we extend the proposed model and techniques within a cross-lingual setting, where we learn from a parallel speech corpus and attempt to transfer affect (including prominence patterns) in a speech-to-speech machine translation task. We apply the cross-lingual transformation on lecture style TED videos for the case of English→Portuguese translation. Evaluation will be conducted through large-scale perceptual tests and appropriate objective comparisons.
Contents

1 Introduction 1
   1.1 Proposed Work 3
   1.2 Expected Contributions of This Thesis 4

2 Intonation Modeling 5
   2.1 Intonation: Its Parts and Uses 6
   2.2 A Review of Existing Intonation Models 7
   2.3 Preliminary Work: The SPAM Model 9
      2.3.1 The Iterative Phrase/Accent Decomposition Method 10
      2.3.2 Speech Databases 12
      2.3.3 Phrase/Accent Component Training 13
      2.3.4 Preliminary Evaluation 15
   2.4 Proposed Work 16
   2.5 Summary 18

3 Intonation Transformation 19
   3.1 Related Work 21
   3.2 Preliminary Work: GMM Mapping of Pitch Accents 22
      3.2.1 Automatic Extraction of Metrical Feet 22
      3.2.2 2-Level F0 Conversion Technique 25
      3.2.3 Preliminary Experiments 26
<table>
<thead>
<tr>
<th>3.3 Proposed Work: Identity and Style Transformation</th>
<th>28</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.4 Summary</td>
<td>29</td>
</tr>
<tr>
<td>4 Affect Transfer in Speech-to-Speech Machine Translation</td>
<td>33</td>
</tr>
<tr>
<td>4.1 Preliminary Work: Parallel Speech Corpora</td>
<td>35</td>
</tr>
<tr>
<td>4.1.1 Analysis of Prominence</td>
<td>37</td>
</tr>
<tr>
<td>4.1.2 Cross-lingual F0 conversion</td>
<td>38</td>
</tr>
<tr>
<td>4.1.3 Initial Experiments Transferring Prominence Across Languages</td>
<td>39</td>
</tr>
<tr>
<td>4.2 Proposed Work</td>
<td>39</td>
</tr>
<tr>
<td>4.3 Summary</td>
<td>40</td>
</tr>
<tr>
<td>5 Evaluating Intonation</td>
<td>43</td>
</tr>
<tr>
<td>5.1 Proposed Objective Evaluations</td>
<td>44</td>
</tr>
<tr>
<td>5.2 Proposed Subjective Evaluations</td>
<td>45</td>
</tr>
<tr>
<td>6 Thesis Timeline</td>
<td>47</td>
</tr>
<tr>
<td>6.1 Estimated Timeline:</td>
<td>47</td>
</tr>
<tr>
<td>6.2 Proposed Tasks</td>
<td>48</td>
</tr>
<tr>
<td>Bibliography</td>
<td>49</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Speech is the most natural form of human communication. While the underlying string of words is indeed what is said in an utterance, a lot of information is encoded in how it gets said. Information that can only be alluded to in writing, can be richly expressed in sound [Bolinger, 1986]. Humans express a range of emotions, convey mental states and display attitudes through speech [Juslin and Scherer, 2005]. Using devices like word emphasis (how much to stress each word), melody (the recurrent highs and lows in pitch) and pace (the rightly timed pauses), we draw the listener's attention to different parts of a sentence or convey different shades of its meaning. The same sentence can be spoken in many ways to make it sound incredulous, sarcastic or pleading by differently employing these devices. In some cases, speech becomes indispensable in understanding the speaker's true intent, disambiguating from among the several senses the words may mean.

‘Prosody’ is the overarching term that refers to all those elements of spoken language that may not be encoded in grammar or the choice of words. Intonation is perhaps the most ‘affective’ part of prosody that, through variation in pitch, conveys the speaker’s true intent in a sentence. Apart from communicating the meaning, intonation also gives away our personal/ethnic identity and social class to the listener. Unlike the choice of vocabulary, which can be controlled to an extent, it is hard to mask these cues that are abundantly contained in speech and become an essential part of speaker identity. We use intonation to identify who a speaker is [Abberton and Fourcin, 1978]; or recognize a speaking style [Hirschberg, 2000]; to place him in an accent group [Grabe and Post, 2002] or to identify a language [Ohala and Gilbert, 1979]. This is the general space of this thesis — computationally modeling the systematic ways in which meaning is conveyed.
through intonation, and how aspects of speaking style and identity are captured in prosody.

Text-to-Speech (TTS) is the technology that most closely deals with and has a major stake in prosody. The goal of TTS is to synthesize speech using computers to convey the underlying meaning in a way that is both natural and appropriate for a piece of text, in a given context. Furthermore, it is desirable for TTS systems to sound as close as possible to the speaker used to train the synthetic voice.

Rule-based approaches to intonation generation were the mainstay in the early TTS systems [Mattingly, 1966, Young and Fallside, 1980, Anderson et al., 1984, Klatt, 1987]. Learnings from studies in speech science and analysis were carefully programmed on computers to generate speech with the desired properties. But with the advent of data-driven techniques like Unit Selection [Hunt and Black, 1996], there was a paradigm shift in how speech synthesis was done. These techniques, broadly termed concatenative synthesis approaches, stitch together bits and pieces of speech from different utterances of a speaker and smooth the transitions to synthesize a new sentence. The resulting speech is natural and can be of high quality, given a clean, consistent and large database of speech recordings. However, these techniques put the onus of prosody entirely on the data. To synthesize a compassionate or an authoritative voice meant that a different speech database with a different style of speech recordings had to be used [Yamagishi et al., 2003]. This is obviously expensive, inflexible and unscalable to a variety of speakers, speaking styles and languages.

To address the concerns with concatenative speech synthesis strategies, the TTS community has moved towards what is referred to as statistical parametric speech synthesis (SPSS) [Tokuda et al., 2000, Black, 2006, Zen et al., 2009]. As the name rightly suggests, these techniques analyze and store spectral and prosodic parameters, usually of short intervals of speech (frames) within a statistical model. For synthesizing a new sentence, the model is used to predict the appropriate parameters and the inverse operation is applied on the parameters to generate the speech waveform output. The benefits afforded by SPSS approaches include minimal requirements on the quantity and consistency of training speech data, and the flexibility to transform the model parameters to achieve desired output of speech. Though the resulting synthetic speech in SPSS is natural and understandable, it may lack personality and sound monotonous if the underlying parameterization or modeling framework are sub-optimal. In this thesis, we present a statistical framework for intonation modeling that is more suited for expressive speech synthesis within SPSS.
1.1 Proposed Work

As speech technology becomes more pervasive, so does the cost of a distasteful listener experience caused by a bad speech output in the form of dull, or inappropriate intonation. This is more significant for complex systems involving other technologies like Speech-to-Speech Machine Translation (S2SMT), Computer assisted Language Learning (CALL) etc., where TTS has to synthesize as appropriate to the context (e.g., emphasize the right word as in the source language, speak according to a learner’s level of comprehension and fluency etc.). In this thesis we apply the proposed model and cross-lingual transformation techniques for the case of speech translation of lecture videos from English ↔ Portuguese.

1.1 Proposed Work

With increased use of statistical and parametric approaches, a new avenue is opened for prosody modeling within TTS, which we intend to explore in this thesis. Specifically, we propose computational approaches for modeling and transformation of prosody, primarily intonation, between speakers, speaking styles, and languages.

As a first step towards this goal, we propose a new framework for intonation modeling for affective speech synthesis, that can synthesize natural, expressive and appropriate intonation within SPSS. The framework is designed with a view to capture aspects of speaking style and identity within the intonation model parameters. The proposed model is evaluated on a diverse set of speech tasks over a variety of intonation styles like — read speech, radio newscast and audiobooks. The language neutrality of the framework will also be tested by applying the model for different languages.

We then propose techniques for style-sensitive transformation of intonation between speakers and speaking styles. We extend these techniques further and propose techniques to transform prosody across languages, where there are more challenges with respect to word reordering etc. The framework and transformation techniques will be used in a real world speech system that aims to automatically translate lecture videos from one language to another (automatic dubbing). Traditionally, the approach for speech translation is thought of as being a pipeline of Automatic Speech Recognition (ASR), Statistical Machine Translation (SMT) and TTS. However, the translation is incomplete without transferring the speaker’s affect, which includes aspects like word focus, and overall speaking style in the original language. We will address this aspect to complete the process of speech translation by imposing the source speaker affect on the target language side after translation.
While human preferences are the true test of how good intonation is, it is useful to have reliable and consistent objective metrics that correlate well with human judgments. This is important as objective metrics can be used to judge subtle improvements in intonation and can also be used as criteria to optimize in computational methods. However, the lack of good objective measures is one of the unsolved problems in prosody research. In this thesis, we will explore a range of metrics that can be used to automatically compare several models of intonation.

1.2 Expected Contributions of This Thesis

1. **A computational model of intonation for affective TTS**: An affective intonational model for statistical parametric speech synthesis along with associated training and synthesis algorithms. Investigation into the optimal level of prosodic representation including frame, phoneme, syllable and metrical feet.

2. **Techniques for transformation of Intonation**: Identity and style preserving transformations of intonation between speakers and speaking styles.

3. **Techniques for cross-lingual intonation transformation**: Training F0 conversion functions from corpora of parallel speech and word alignment information.


5. **Evaluation metrics for intonation**: Objective and subjective measures to evaluate the modeling and transformation techniques for intonation.
Chapter 2

Intonation Modeling

As motivated in Chapter 1, intonation modeling is one the most important aspects in text-to-speech synthesis (TTS). Perhaps due to the many different theories in intonation, only few of them practical but all valid, there is no one right way to go about modeling intonation. The key questions to consider when attempting to model intonation for TTS are— (i) Are there efficient analysis and synthesis strategies with minimal reconstruction error for the model? (efficiency), (ii) Can it predict natural and appropriate intonation contours from text only input (expressiveness) (iii) Is the model amenable to transformation to a target speaker or style? (flexibility), (iv) Does it work in limited data scenarios, and does it improve as more data is available? (robustness) and finally a more theoretical desirable, (v) Does it have direct relevance to existing intonation theories—can it benefit from or validate progress in related areas like phonetics, phonology, psycholinguistics etc.

In Text-to-Speech synthesis, text is the only input from which the low-level intonation contour needs to be estimated. This is not straight-forward since intonation encodes a lot more information in the form of structure and type into an utterance than conveyed by the words in the sentence [Taylor, 2009]. The scope of this information may well be beyond words, as broad prosodic phenomena like emphasis, or at the frame level, as microprosody. Intonation models must efficiently capture this complex information and map it from the high-level linguistic input. Current techniques include frame-level decision trees with questions about phonetic/linguistic features from text analysis, and pitch statistics value at the leaf nodes [Black, 2006]. Zen et al. [2007] use phoneme level context-dependent Hidden Markov Models and [Schröder and Trouvain, 2003] use word-level Pitch target estimation and interpolation. While understandable intonation can be generated by
these intonation models, for the current goal of an affective intonation model with all the desirables mentioned above, it is necessary to design a new model.

2.1 Intonation: Its Parts and Uses

Physiologically, the fundamental frequency (or F0) is the rate at which a speaker’s vocal folds open and close in producing speech. The acoustic analogue of F0 is called pitch, the frequency at which speech is perceived by a listener. In practice, pitch is taken as the intonation contour of an utterance. Given that F0 exists only when the vocal folds vibrate, it follows that unvoiced sounds (like /s/, /ch/, /f/ etc.) do not have an F0 value. However, listeners perceive as if there is an interpolation through the unvoiced regions [Taylor, 2009]. The same is done in practice where simple interpolations through unvoiced regions (except pauses) are done to better model intonation as a contour. Fig. 2.1 illustrates a natural F0 contour of a read speech utterance of an American adult male speaker, interpolated through unvoiced regions and smoothed over a 5-point window. In this work, these are the kinds of contours which we try to learn from, and predict for unseen text. It is to be noted that pitch detection is still not perfect and often there are errors like halving (observe F0 around 200’th frame in Fig. 2.1), which pose more challenges to modeling. There may be other artefacts that are perceptually irrelevant and microprosody, the local fluctuations in F0 that arguably render some naturalness to the speech [van Santen and Hirschberg, 1994]. Pre-processing routines may be used to approximate F0 contours with perceptually identical piecewise linear or quadratic spline curves to remove the un-modelable fluctuations. In this work however, we use the smoothed F0 contours as shown in Fig. 2.1, and leave it to a robust intonation model to learn the right level of detail at which to model the contour. We exploit statistical approaches that offer the advantage of learning an optimal model, that explains the salient and recurring phenomena encountered in the training data.

The two notable phenomena studied in intonation contours are the Phrase boundaries (regions preceding pause regions) and Tones (or pitch accents), which are the ‘highs’ and ‘lows’ of the contour.

Functionally, we’ve seen in Chapter 1 that intonation in speech serves many purposes—often disambiguating the semantics but also adding complementary information making it sound more affective by attributing emphasis and style to the content delivered [Bolinger, 1989]. An expressive or emotional utterance has a high dynamic range within the contour and propositional utterances have
2.2 A Review of Existing Intonation Models

One view of intonation broadly categorized as auto-segmental metrical phonology [Ladd, 1996] aims to qualitatively describe the intonation contour as a series of known shapes (eg., ToBI, [Silverman et al., 1992]) that complement the underlying linguistic structure of a sentence. These shapes are studied as being unique to a language, or a dialect or to a particular linguistic form (questions, affirmatives etc.). Other phonological models, specifically for explaining prominence patterns include work by [Liberman and Prince, 1977] on metrical feet. However, as phonological theories, these approaches are descriptive of intonation but do not provide predictive knowledge [Xu, 2012] needed for Text-to-Speech synthesis. Recent attempts by [Rosenberg, 2010] include automatic annotation of speech data with ToBI labels that we exploit in later chapters.

From a speech production (physiological) perspective, the Fujisaki model of intonation [Fujisaki, 1983] is a mathematical model that tries to approximate the

Figure 2.1: An illustration of a smoothed, natural F0 contour, interpolated through unvoiced regions within each phrase.

contours without wide excursions (low variance). Similarly declarative/affirmative utterances are usually marked with a fall towards the phrase boundary in contrast to interrogative/exclamatory sentences that are likely to show a phrase final rise.

2.2 A Review of Existing Intonation Models

One view of intonation broadly categorized as auto-segmental metrical phonology [Ladd, 1996] aims to qualitatively describe the intonation contour as a series of known shapes (eg., ToBI, [Silverman et al., 1992]) that complement the underlying linguistic structure of a sentence. These shapes are studied as being unique to a language, or a dialect or to a particular linguistic form (questions, affirmatives etc.). Other phonological models, specifically for explaining prominence patterns include work by [Liberman and Prince, 1977] on metrical feet. However, as phonological theories, these approaches are descriptive of intonation but do not provide predictive knowledge [Xu, 2012] needed for Text-to-Speech synthesis. Recent attempts by [Rosenberg, 2010] include automatic annotation of speech data with ToBI labels that we exploit in later chapters.

From a speech production (physiological) perspective, the Fujisaki model of intonation [Fujisaki, 1983] is a mathematical model that tries to approximate the
logarithm of F0 as a sum of the outputs of two critically damped second order filters, that each produce a single shape (one for a long range phrase command, and another for local accent command). While the model is appropriate for declarative sentences, it cannot handle the phrase final rise in intonation like in English interrogatives. It also requires careful hand labelling to identify the start of the accent commands. There have been subsequent attempts to automate the process [Mixdorff, 2000]. Other approaches based on the principle to model intonation as superposition of intermediate contours [van Santen et al., 2005, Bailly and Holm, 2005] were also successful in different languages.

As a practical engineering model that can analyze and synthesize arbitrary intonational contours, the Tilt model [Taylor, 2000] describes the contour as a sequence of rise-fall events connected by simple linear interpolations. Each rise-fall event can be parametrized by a 4 valued vector, including duration, peak position, amplitude and shape of the event to approximate any arbitrary contour as illustrated in Fig. 2.2.

![Figure 2.2: An illustration of Tilt features for an event (taken from Taylor [2009].)](image)

While formally, each event can be associated with an underlying accented syllable, prediction of a speaker’s accented syllable is not trivial since speakers may not conform with canonical models of accent placement [Pan and Hirschberg, 2000]. Besides, the uniqueness of a speaker lies in the stress patterns he employs in a sentence, which we intend to model as his speaking style in this work.
Several other models describe the form of intonation from different perspectives and this section highlighted a few that are relevant to the discussion in this work. A more comprehensive survey of Intonation models may be found in [Taylor, 2009, Chapter 9].

### 2.3 Preliminary Work: The SPAM Model

In proposing a novel framework that suits our need for speaker style modeling, we draw upon strengths of existing representations to design an intonation model for SPSS. Our model, referred to as the Statistical Phrase/Accent Model (SPAM) of intonation is described in [Anumanchipalli et al., 2011]. It comprises two independent clustering and regression trees (CART, one for modeling long term phrase; and another for accents, the short term F0 excursions). This is illustrated in Fig. 2.3. To parameterize the accents, we use a variant of the Tilt representation where *every* syllable’s $\log(F0)$ shape is described as a 4-tuple Tilt vector, in contrast to only accented syllables being modeled. This gives complete control over the generated contour to synthesize multi-syllable events and also removes the requirement of intonation labels. Additionally, we use the Tilt model not to represent the actual values of $\log(F0)$ but the ‘residual’, after appropriately subtracting the phrase component.

![Figure 2.3](image_url)  
Figure 2.3: An illustration of the model components of the SPAM intonation model. In contrast, conventional intonational models employ a single decision tree [Black, 2006].

Existing statistical models for F0 usually generate ‘averaged out’ intonation contours with reduced dynamic range making them sound monotonous and robotic.
Previous solutions to this problem in spectral include the use of global variance of the reference natural data and imposing it on the generated parameters to simulate naturalness [Toda and Tokuda, 2007]. In our work, we deal with the problem by splitting the logarithm fundamental frequency with the assumption that it has two underlying additive components (motivated by [Fujisaki, 1983]). If not separated, these components can nullify each other and corrupt the final model (e.g., the down-drift phenomenon reduces the height of the contour in later regions of a phrase, causing two qualitatively similar accents in different regions of the phrase to be treated differently). This decomposition improves the dynamic range in the synthetic contours and also makes optimal use of the data. However, we refrain from predefining the shapes of the components, instead letting them be learned from data.

As for the number of accent shapes allowed, we use the hypothesis from phonological intonation theories that there are only a finite number of ‘known’ shapes that describe excursions in the F0 contour (similar in motivation to [Möhl er and Conkie, 1998]). We also lay additional constraints on the Phrase tree to have only long range contextual questions and the accent tree to have only short range questions. This is done to improve the discriminability of the underlying components further.

The next section describes the component decomposition method to train a SPAM intonation model from a speech database.

2.3.1 The Iterative Phrase/Accent Decomposition Method

To train a SPAM intonation model from a speech database, it is necessary to extract the phrase and accent components from the log($F_0$) contour of each utterance in the training data. To accomplish this, we employ an iterative, constrained algorithm to split each contour to its respective phrase/accent components.

An initial estimate of phrase command is used to start the procedure. We use the minimum value of $F_0$ over a syllable as an approximation of the phrase component. Figure 2.4 illustrates a phrase component initialization as the minimum value of the contour over each syllable.

The residual (i.e., log($F_0$) – phrase) is then obtained as the contribution of the accent components and some random noise. To parameterize the accents, the residual is analyzed and coded as a 4-valued Tilt tuple over each syllable. Intermediate CART tree models of phrase and accents are trained over the syllables using the estimates of phrase and accent components for each utterance. These models are
evaluated by resynthesizing the $\log(F0)$ contours and computing the average root mean squared error (rmse) and correlations with respect to the reference contours. The process is repeated for several iterations to optimize over all training data. The following constraints are applied at each stage –

- For the phrase components, at each iteration the phrase 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 estimated phrase component at each phoneme (done at the phoneme level for a sharper resolution).

- For the accent components, the constraint is that they should be limited in number. A $k$-means clustering is performed to model the representative shapes of accents over all syllables.

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. The constraints are chosen to be minimally assuming and are generic across languages, speakers or speaking styles, giving the model more degrees of freedom. Using the intermediate models built (phrase CART tree and accent codebook), a new estimate for $\hat{\log}(F0)$ is reconstructed. The mean reconstruction error over each syllable is added to the estimate of the baseline and residuals are recomputed. This
procedure is repeated till the objective criterion is met, here it is the minimum \( \log(F_0) \) reconstruction error on a held out test set. The parameters that give the best reconstruction error are chosen as the optimal phrase and accent components. A pseudocode of this method is provided as Algorithm 1.

Algorithm 1: Constrained Component Extraction

1: for all utterances do  
2:   for all syllables do  
3:     set phrase to min \{\log(F_0)\}  
4:     set accent to tilt(\log(F_0) - phrase)  
5:   end for  
6: end for  
7: while error \geq \epsilon \ do  
8:   train an accent codebook of size \( k \) over all accents  
9:   train a codebook CART tree using local features  
10:  train a phrase CART tree using long range features  
11:  for all utterances do  
12:     Generate \( \hat{\log}(F_0) \) using phrase & accent codebook  
13:     for all syllables do  
14:       accumulate error (\( \hat{\log}(F_0) - \log(F_0) \))  
15:       update phrase to (phrase + error)  
16:       update accent to tilt(\log(F_0) - phrase)  
17:     end for  
18:   end for  
19: end while

2.3.2 Speech Databases

We evaluate the proposed model and the training algorithm using several speech databases. We choose 8 speakers from 3 distinct speaking styles. Three sources are used: 2 speakers (\( \text{rms}, \text{slt} \)) from ARCTIC [Kominek and Black, 2003], a read speech database of short declarative sentences selected from a collection of stories; 5 speakers (\( f1a, f2b, f3a, m1b, m2b \)) from BURSC [Ostendorf et al., 1996], a radio broadcast corpus and 1 female speaker's digital audio book (\textit{emma}) of Jane Austen's \textit{Emma} [Prahallad and Black, 2011]. The databases are automatically segmented, aligning the speech with the transcription at a phonetic level. Pitch contours are
extracted using the `get_f0` tool of ESPS software [Talkin, 1993], smoothed and interpolated through unvoiced regions. 8 SPSS voices are built, one for each speaker using the Clustergen framework [Black, 2006].

### 2.3.3 Phrase/Accent Component Training

The iterative F0 decomposition algorithm described in Section 2.3.1 is used to extract the phrase and accent components of the F0 contours of all utterances of each speaker. Table 2.1 shows some features used in training the component models.

<table>
<thead>
<tr>
<th>Phrase features (Global trend)</th>
<th>Accent features (Local excursion)</th>
</tr>
</thead>
<tbody>
<tr>
<td>word POS</td>
<td>word POS</td>
</tr>
<tr>
<td>phrase number</td>
<td>syllable category</td>
</tr>
<tr>
<td>word position in phrase</td>
<td>predicted accent</td>
</tr>
<tr>
<td>#syllables in phrase</td>
<td>lexical stress</td>
</tr>
<tr>
<td>content words in phrase</td>
<td>prev/next values</td>
</tr>
<tr>
<td>normalized values of above</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.1: Example features used to train Phrase/Accent models

Note that conventional statistical F0 models use all features together in the model training. But in the proposed method, we separate them to appropriately deal with the phrase and accent components separately. Table 2.2 presents a trace of the training algorithm on one speaker f1a. It can be seen that the overall root mean squared error (RMSE) decreases and correlation (CORR) increases on the training data before converging over iterations.

<table>
<thead>
<tr>
<th>#Iter</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>0.457</td>
<td>0.384</td>
<td>0.237</td>
<td>0.186</td>
<td>0.182</td>
<td>0.181</td>
<td>0.181</td>
<td>0.180</td>
</tr>
<tr>
<td>CORR</td>
<td>0.488</td>
<td>0.549</td>
<td>0.641</td>
<td>0.705</td>
<td>0.714</td>
<td>0.717</td>
<td>0.718</td>
<td>0.719</td>
</tr>
</tbody>
</table>

Table 2.2: Average RMSE/Correlations of log (F0) contours per iteration on task f1a

The average resynthesis error is about 1-1.2 Hz for all speakers, which is perceptually insignificant. An example component split is shown in Figure 2.5 where the `log(F0)` is plotted along with the derived phrase and accent components and
the resynthesized contours. It can be seen that the final derived phrase is a gradual falling contour (though no such constraint is explicitly enforced in the model) and the accents are sequences of what look like metrical feet spread over multiple syllables.

![Image of F0 contour split into optimal phrase and accent commands](image.png)

**Figure 2.5:** Example F0 contour split into optimal phrase and accent commands

The trained phrase and accent components can be used as the intonation model and used for predicting F0s of unseen test sentences. At test time, the phrase and trees are traversed to predict the best long-term trend curve and local excursion sequence and added to generate an F0 contour for a novel sentence. Figure 2.6 compares the predicted contours of an unseen sentence generated by default frame-based F0 prediction model in Clustergen and the proposed SPAM intonation model. The durations of the reference sentence are used to enable the comparison in the illustration. It is easy to see that the F0 generated by the proposed approach has better variance and is seemingly more affective than the default frame-based F0 model.
2.3 Preliminary Work: The SPAM Model

2.3.4 Preliminary Evaluation

Table 2.3 objectively compares the contours generated by frame-based F0 model and the proposed statistical Phrase/Accent model on all 8 voices in terms of mean error and correlation. The proposed model scores comparably, yet worse than the default model in most cases. However, RMSE and CORR measures are computed at the frame level (5-10ms) which may be unsuitable for comparison of affect of intonation contours that have a much higher resolution. This observation is in keeping with earlier studies showing that these measures for comparison of synthetic F0 contours may not be ideal [Clark and Dusterhoff, 1999]. More discussion on objective metrics for evaluating intonation is presented in Chapter 5.

To get a more reliable comparison of the two approaches, we conducted subjective $AB$ listening tests, where human listeners were presented with a pair of speech stimuli, same in all respects except the intonation. 3 tasks sit, f2b and emma from each speaking style were used for the listening tests. 10 unseen sentences from each task are synthesized using the default and proposed F0 models. 11 American English speakers were presented the stimuli in a random order and asked to judge which sample they prefer to hear. Fig 2.7 summarizes the user preferences. It can...
be seen that the SPAM generated intonation contours were preferred by listeners in over 80% of cases conclusively showing that the proposed model generates more natural intonation contours than the default model, in all speaking styles.

In understanding these results, subjective tests must take precedence over objective measures since better human perception is the ultimate goal of TTS. However, the mismatch between the results of objective and subjective tests may seem counter-intuitive, because the objective metrics are the criteria we are optimizing on. We should note here that augmentative parts of prosody are idiosyncratic to each utterance and are dependent on the speaker's cognitive state. In other words, intonation cannot be 'exactly' trained or predicted for unseen text without access to the speaker's neural processes, which is of course beyond the scope of this research. The take-home for us, however is the reinforced need for better objective metrics that correlate well with human judgments.

### Table 2.3: Objective comparison of voices

<table>
<thead>
<tr>
<th>Task</th>
<th>Frame-based F0</th>
<th>SPAM F0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>CORR</td>
</tr>
<tr>
<td>rms</td>
<td>10.50</td>
<td>0.66</td>
</tr>
<tr>
<td>slt</td>
<td>11.15</td>
<td>0.63</td>
</tr>
<tr>
<td>f1a</td>
<td>29.85</td>
<td>0.44</td>
</tr>
<tr>
<td>m1b</td>
<td>14.80</td>
<td>0.45</td>
</tr>
<tr>
<td>f2b</td>
<td>28.23</td>
<td>0.57</td>
</tr>
<tr>
<td>m2b</td>
<td>23.49</td>
<td>0.42</td>
</tr>
<tr>
<td>f3a</td>
<td>27.38</td>
<td>0.35</td>
</tr>
<tr>
<td>emma</td>
<td>41.58</td>
<td>0.09</td>
</tr>
</tbody>
</table>

2.4 Proposed Work

We have argued in preliminary work that conventional frame-based approaches to F0 generation are sub-optimal since i) linguistic features do not have such low resolution to discriminate F0 values at the level of a frame, thereby generating implausible and repetitive F0 contours, ii) synthesis of prosodic functions requires greater flexibility, to control the contour at higher prosodic levels like syllables edges and intonational phrases etc.

By employing phrase/accent decomposition and syllable level accents, the SPAM
framework is shown to generate more natural intonation contours that are perceptually preferred over the default models. We also found that the proposed computational techniques can be used to extract phonological phenomena like global declination and metrical feet. This dimension of the SPAM framework is yet to be exploited. There are no restrictions inherent in the model to represent accents at the syllable level or the phrase at the segment level. Within the proposed framework, it is possible to empirically determine the best representational level for intonation. This is a key aspect in the design of intonation models that affects the quality of the linguistic → prosodic mapping, essential for prediction from text.

Given that the proposed framework models phonologically salient trends as realizable parameters, it is possible to qualitatively characterize data through the SPAM model parameters. Some planned characterizations include speaker expressiveness vs. the number of the accent codebooks; linguistic types vs. slope of declination; dialect vs. peak position in the accent etc.

Another interesting direction in TTS currently is the synthesis of multi-paragraph text, like Audiobooks. In our initial experiments, we did not change the setup for modelling intonation within this task. But in such prosodically rich domains, there is overwhelming evidence to support the view that prosodic cues go well beyond
the sentence. We are yet to exploit such high-level, discourse features within TTS intonation models. In view of these observations, we propose the following extensions within the SPAM framework for this thesis—

- Investigation of the right level of representation for phrases and accent shapes of intonation—including the frame, phoneme, syllable and metrical feet.
- Automatic identification of intonational correlates of prosodic functions like word prominence and speaking style within the SPAM framework.
- Investigating features beyond the sentence for intonation modeling in audio-books.
- Application of the SPAM intonation model for expressive TTS in different languages.

2.5 Summary

In this chapter we have introduced a statistical framework for intonation modeling that will be used for the rest of the thesis. A method for decomposition of F0 contours into long-term trend and short-term excursions. These are appropriately parameterized at the level of syllables and modelled for prediction from linguistic input. We have proposed future extensions to the models that will be pursued in the remainder of the thesis.
Chapter 3

Intonation Transformation

In Chapter 2, we presented a framework for statistical intonation modeling that can generate expressive F0 contours for unseen text. Another of the desirables we listed for an intonation model was flexibility. Flexibility here refers to the ability to transform a speaker's intonation contour to sound like that of another. The primary application for such a transformation is in voice conversion.

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. [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.

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. Figure 3.1 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 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 the speakers and slt, the only female speaker understandably 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, prominence and stress patterns. These speakers also happen to be the only non-American speakers within this set (awb is a Scottish English speaker and ksp is an Indian English speaker).

Figure 3.1: F0 contours of 5 arctic speakers for the phrase “Will we ever forget it.”

This illustration throws some light on the problem of intonation transformation and the challenges associated with such an attempt. Note that the example given is a from a read speech task. The style and range of expression through intonation is much higher as the task is prosodically more complex, 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 to train a mapping from the source speaker’s intonation patterns.
3.1 Related Work

Most voice conversion techniques use frame level representations of F0 in the order of about 5 milliseconds of speech. A popular F0 conversion method is pitch mean/range adaptation (as used in [Toda et al., 2007]) where the source speaker’s F0 is converted to the target speaker by employing a $z$-score transform, which is an affine transformation at the frame level —

$$F0^\text{tgt}(t) = \frac{\sigma^\text{tgt}}{\sigma^\text{src}} (F0^\text{src}(t) - \mu^\text{src}) + \mu^\text{tgt} \quad (3.1)$$

where $F0(t)$ is the fundamental frequency at time instant $t$ and $\mu$, $\sigma$ denote the mean and standard deviation of F0’s from the training and adaptation data for the source and target speakers respectively. [Gillett and King, 2003] employ the above at three levels of the contour for a better resolution in transformation. In these approaches, while the pitch range is 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 as illustrated in Fig. 3.1. For example, lexical stress is at the level of the syllables and prominence patterns can be explained at the level of metrical feet.

Some interesting attempts for F0 conversion include [Helander and Nurminen, 2007], where syllable level codebooks were trained and CART trees 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 initial experiments 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.
3.2 Preliminary Work: GMM Mapping of Pitch Accents

In our preliminary work, we propose some directions towards capturing the speaking style of a target speaker. While the correlates of speaking style also exist in duration and phrasing, in these experiments we focus specifically on the F0 contour. Within the contour, the speaking style is manifested in the sequence of shapes called accent tones. The uniqueness of a speaker or a task lies in the shapes that he employs—the length of the shape, the position of peak, the overall shape (amount of rise or fall or rise+fall etc.) We've seen that intonational phonology suggests that there are only a finite number of shapes that characterize the F0 contour for a language, dialect or a linguistic type. Given this hypothesis, its then logical to talk about conversion between these shapes.

Here, we attempt an informed conversion of a source speaker’s accent tones to those of a target speaker. As a first step in modeling such a conversion, we need to decide the scope within which we want to analyze the accent tones of the two speakers. While Helander and Nurminen [2007] used syllables for anchoring the accent tone and simulating parallel data, the style that we intend to capture in this work could be spread over multiple syllables, phonologically referred to as a stress group or metrical foot. Believing in the phonological hypothesis that the F0 is structured for conveying linguistic meaning of the underlying text, our goal is to convert an unseen F0 contour of the source speaker and predict the likely contour (with the appropriate accents) that the target speaker may have produced in delivering that sentence. For capturing speaking style, it makes practical sense to consider an accent in context of its neighboring syllables that exist within the same accent tone. In section 3.2.1, we propose a method to detect such regions, which we refer to as ‘feet’. We use the iterative phrase/accent decomposition method (Sec 2.3.1) described in the earlier chapter to separate the phrase and accent components. Whereas we used syllables in Chapter 2 for the SPAM model, we use the method at the level of metrical feet, so that the discussion of conversion of accent tones across speakers is more tangible.

3.2.1 Automatic Extraction of Metrical Feet

In order to model the correspondences and learn a mapping between the tone accents of two speakers, it is necessary to analyze the F0 contours within a phono-
3.2 Preliminary Work: GMM Mapping of Pitch Accents

logically identical context. As mentioned before, in order to model multi-syllable prosodic phenomena, we analyze each accent at the level of the metrical foot, where a foot may be defined as consisting of an accented syllable, followed by all un-accented syllables that precede the next accented syllable or a phrase boundary. In principle, this modeling is similar to the one used in [Klabbers and van Santen, 2006], however the data used in the latter is manually annotated and doubly checked for errors. Since we are interested in analysis and conversion to/from arbitrary speaker pairs, we propose an automatic method to detect foot like regions within the F0 contour and parameterize them within the SPAM framework.

To evaluate the effectiveness of the parameterization and the representational level, we compute reconstruction errors and correlations. An optimal representational scheme gives a small reconstruction error and preferably has some theoretical basis. We have seen in chapter 2 that syllable is more suited as a unit for intonation than the frame. In the work described in this chapter, we use the same strategy for the phrase components but the accent residuals are parameterized at the level of metrical feet.

To detect feet like regions, we run the iterative component extraction method to decompose all the $\log(F0)$ contours into their globally optimal phrase and syllable level accent residuals. We then run another pass of analysis over all the syllables, re-analyzing groups of syllables (referred in the algorithm as a stress group, having only a single primary stressed syllable in the group), so as to obtain the least reconstruction error. We also impose an additional constraint that a feet can end only at a word boundary (and now referred to as a sense group). This is done due to practical concerns arising from feet ending within a word during voice conversion. The procedure is presented as Algorithm 2 and an illustration of the example reconstruction from the detected foot regions using automatically analyzed parameters is presented in Fig. 3.2.

The output of the algorithm for each utterance is a sequence of Tilt vectors, one per each foot that is detected on the contour. It also outputs the linguistic context, i.e., the words underlying each foot. These vectors are easily synthesizable as is shown in the Fig 3.2. It can be seen that the foot based representation is quite good in that it can parameterize the salient peaks and the overall shape with minimal loss in error and correlation, and removes some unnecessary fluctuations within the contour that would be inevitable with a lower level representation. Table 3.1 presents the mean error and correlations for natural and resynthesized contours of the parameterized feet regions. 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 feet level representations, in addition to giving a simple parameterization of the speaker style.

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 feet. So, the algorithm given here is run on the source speaker and the feet are detected. It should be noted that different speakers may have a different set of metrical feet they choose to employ. For the arctic speakers, on an average there is only about 38\% of the feet matching per utterance for any 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 feet 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 feet analyzed in each utterance pair. The contours over each feet 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 3.2.1 shows the correlation matrix for the shape parameter (tilt) of corresponding feet in each speaker pair. Note that this matrix is not symmetric because the feet 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.

3.2.2 2-Level F0 Conversion Technique

The SPAM intonation model represents $\log(f_0)$ as a sum of two components, the phrase, that models the long term trend of the contour and accents that model the local detours. In this chapter, we use the z-score transform given in Equation 3.1 on the phrase contour because the phrases roughly capture the mean pitch and and a simple affine transformation is sufficient to approximate the target speaker’s pitch level. Accents are however complex since they are described by many parameters that bear information about the speaker style. To transform accents, we train a mapping between the two speakers’ accent vectors using parallel data as described below. An illustration of the proposed 2 level F0 conversion is shown in Fig 3.3.

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) based 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].

Let $x_t$ and $y_t$ be the Tilt accent vectors for each metrical foot. The joint probability density of the source and target vectors is modelled as the following GMM -

$$P(z_t | \lambda(z)) = \sum_{m=1}^{M} w_m N(z_t; \mu_m^{(z)}, \Sigma_m^{(z)})$$
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 $w_m^{(z)}$, $\mu_m^{(x)}$ and $\Sigma_m^{(x)}$ 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, 2000] and are not independent.

At conversion time, given an accent shape $x_t$ of the source speaker, we want to predict the most likely accent $y_t$ of the target speaker as follows—

$$
\hat{y}_t = \sum_{i=1}^{M} p(m_i|x(t), \lambda^{(z)}) E(y_t|x_t, m_i, \lambda^{(z)}),
$$

$$
E(y_t|x_t, m_i, \lambda^{(z)}) = \mu_i^{(y)} + \Sigma_i^{(yx)} \Sigma_i^{(xx)^{-1}} (x_t - \mu_i^{(x)}),
$$

$$
p(m_i|x(t), \lambda^{(z)}) = \frac{w_i \mathcal{N}(x_t; \mu_i^{(x)}, \Sigma_i^{(xx)})}{\sum_{j=1}^{M} w_j \mathcal{N}(x_t; \mu_j^{(x)}, \Sigma_j^{(xx)})}.
$$

### 3.2.3 Preliminary Experiments

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 to detect metrical feet as described in Sec 3.2.1. 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 metrical foot, the accent residual of the target speaker is analyzed within the same linguistic context as the source speaker, to obtain a parallel set of accents for the speaker pair. GMM Joint densities are trained and the mapping function described in Sec. 3.2.2 is computed. Also the means and standard deviations of the phrase components are computed to learn the $z$-score transformation for the phrase components of the two speakers.
Figure 3.3: Schematic diagram of proposed F0 conversion technique for phrase and accent components

For a test set of 100 sentences, the source speech is analyzed, the feet 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 (\( \text{rmse} \)) and correlation (\( \text{corr} \)) metrics for each utterance. The phrase model of the source speaker’s SPAM intonation model is used to predict a phrase curve, and the phrase level z-score transform is applied to estimate the appropriate phrase contour for the target speaker. GMM transform is applied on the Tilt parameterizations of the accents over each foot, 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 \( \log(F0) \) 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 \( \log(F0) \).

The predicted contours of the baseline and the proposed approaches are evaluated against the reference target speaker \( \log(F0) \)s, using \( \text{rmse} \) and correlation
measures. Table 3.3 shows the average 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 better than the baseline z-score mapping approach on the F0 contour. Fig. 3.4 compares the proposed 2-level transformation with the default method and compares them against the reference $F_0$ contour of the target speaker. It can be seen that the peaks are much closer in their position to the reference using the proposed approach relative to the baseline z-score mapping.

![Figure 3.4: F0 contours of z-score transformed and proposed techniques against reference](image)

Figure 3.4: F0 contours of z-score transformed and proposed techniques against reference

### 3.3 Proposed Work: Identity and Style Transformation

We have been able to achieve objective improvements using the 2-level F0 conversion technique described for SPAM parameterization over several speaker pairs. We are
yet to conduct large scale listening tests and apply the technique on many other pairs of speakers with varying styles of speech. We intend to extend the work in the following directions

- Evaluation of various phonological anchor units for conversion.
- Application of proposed transformation technique for identity/style/dialect transformation.
- Training intonation conversion from non-parallel speech data from speakers.
- Large scale perceptual evaluations to verify the proposed transformations.
- Design of objective metrics for speaker identity and style transformation.

3.4 Summary

In this chapter, we have proposed a phonologically sensitive approach for F0 transformation within voice conversion. An automatic method is designed to extract metrical feet within F0 contours. Corresponding feet 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 foot-based 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. Preliminary evaluations show that the method is better than the baseline frame-level z-score mapping technique for F0 conversion. Parts of this research are communicated to Interspeech 2012. Further directions are proposed to improve the techniques and evaluations within intonation transformation.
Algorithm 2: Automatic Metrical Feet Extraction Method

1: for all Phrases do
2:   foot initialized
3:   predict phrase contour using a statistical model
4:   $F_{0\text{residual}} = F_0 - \text{phrase}$
5:   for all Words do
6:     for all Syllables do
7:       add syllable to stress group
8:       $\text{syl}_{\text{accent}} = \text{tilt}_\text{analyze}(F_{0\text{residual}})$ over syllable
9:       $\text{syl}_{\text{error}} = F_{0\text{residual}} - \text{tilt}_\text{resynth}(\text{syl}_{\text{accent}})$
10:      $\text{foot}_{\text{accent}} = \text{tilt}_\text{analyze}(F_{0\text{residual}})$ over stress group
11:     $\text{foot}_{\text{error}} = F_{0\text{residual}} - \text{tilt}_\text{resynth}(\text{foot}_{\text{accent}})$
12:     if ($\text{foot}_{\text{error}} \geq \text{prev}_{\text{foot}_{\text{error}}} + \text{syl}_{\text{error}}$) AND (word boundary found) then
13:       foot = stress_group - { current syllable}
14:       foot ended on previous syllable
15:       output $\text{prev}_{\text{foot}_{\text{accent}}}$
16:      stress_group = current syllable
17:      $\text{prev}_{\text{foot}_{\text{error}}} = \text{syl}_{\text{error}}$
18:      $\text{prev}_{\text{foot}_{\text{accent}}} = \text{syl}_{\text{accent}}$
19:     else
20:       $\text{prev}_{\text{foot}_{\text{error}}} = \text{foot}_{\text{error}}$
21:       $\text{prev}_{\text{foot}_{\text{accent}}} = \text{foot}_{\text{accent}}$
22:     end if
23:   end for
24: end for
25: if stress_group $\neq \phi$ then
26:   foot ends at Phrase boundary
27:   output $\text{prev}_{\text{foot}_{\text{accent}}}$
28:   stress_group = $\phi$
29: end if
30: end for
### 3.4 Summary

Table 3.1: Errors and correlations on resynthesized contours using different representational levels

<table>
<thead>
<tr>
<th>Speaker label</th>
<th>Syllable</th>
<th>Foot</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>CORR</td>
</tr>
<tr>
<td>awb</td>
<td>0.13</td>
<td>0.77</td>
</tr>
<tr>
<td>bdl</td>
<td>0.09</td>
<td>0.82</td>
</tr>
<tr>
<td>ksp</td>
<td>0.08</td>
<td>0.79</td>
</tr>
<tr>
<td>rms</td>
<td>0.10</td>
<td>0.80</td>
</tr>
<tr>
<td>slt</td>
<td>0.07</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Table 3.2: Correlation matrix for the tilt shape parameter among speakers for corresponding feet

<table>
<thead>
<tr>
<th></th>
<th>awb</th>
<th>bdl</th>
<th>ksp</th>
<th>rms</th>
<th>slt</th>
</tr>
</thead>
<tbody>
<tr>
<td>awb</td>
<td>1</td>
<td>0.139</td>
<td>0.333</td>
<td>0.293</td>
<td>0.254</td>
</tr>
<tr>
<td>bdl</td>
<td>0.155</td>
<td>1</td>
<td>0.290</td>
<td>0.244</td>
<td>0.250</td>
</tr>
<tr>
<td>ksp</td>
<td>0.336</td>
<td>0.218</td>
<td>1</td>
<td>0.301</td>
<td>0.260</td>
</tr>
<tr>
<td>rms</td>
<td>0.256</td>
<td>0.202</td>
<td>-0.01</td>
<td>1</td>
<td>0.213</td>
</tr>
<tr>
<td>slt</td>
<td>0.230</td>
<td>0.162</td>
<td>0.137</td>
<td>0.158</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3.3: Objective comparison of frame level z-score transformation and GMM transformation of feet based accent vectors

<table>
<thead>
<tr>
<th>Speaker pair</th>
<th>Z-score transform</th>
<th>Foot based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>CORR</td>
</tr>
<tr>
<td>bdl-slt</td>
<td>0.494</td>
<td>0.377</td>
</tr>
<tr>
<td>bdl-ksp</td>
<td>0.264</td>
<td>0.450</td>
</tr>
<tr>
<td>bdl-awb</td>
<td>0.305</td>
<td>0.528</td>
</tr>
<tr>
<td>bdl-rms</td>
<td>0.593</td>
<td>0.461</td>
</tr>
<tr>
<td>ksp-bdl</td>
<td>0.324</td>
<td>0.557</td>
</tr>
<tr>
<td>ksp-slt</td>
<td>0.470</td>
<td>0.423</td>
</tr>
<tr>
<td>ksp-rms</td>
<td>0.493</td>
<td>0.339</td>
</tr>
<tr>
<td>ksp-awb</td>
<td>0.334</td>
<td>0.561</td>
</tr>
<tr>
<td>rms-bdl</td>
<td>0.216</td>
<td>0.565</td>
</tr>
<tr>
<td>rms-slt</td>
<td>0.628</td>
<td>0.247</td>
</tr>
<tr>
<td>slt-bdl</td>
<td>0.638</td>
<td>0.465</td>
</tr>
<tr>
<td>slt-rms</td>
<td>0.915</td>
<td>0.531</td>
</tr>
</tbody>
</table>
Chapter 4

Affect Transfer in Speech-to-Speech Machine Translation

In the previous chapters, we have seen some proposals to model and transform intonation. We have seen that intonation cannot be completely predicted from text or completely transformed to a target speaker is due to the lack of sufficient input information and the inherent randomness 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 ‘affect’ to broadly refer to such aspects of speech.

Affect in speech is the non-linguistic content in the speech signal through which the speaker conveys prosodic functions like focus. Aspects of affect are not merely stylistic choices, but also disambiguate between different semantic interpretations of a sentence. Prominence (word focus), for example, helps prioritize the concepts presented in the sentence by laying different emphasis on each content word. While this information is underrepresented in text, there are certain domains where this information may be accessible to the synthesizer. In this chapter, we propose to apply the models 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.

Apart from intonation, affect is also characterized by changes in the rhythm, duration and voice quality of the speaker. In preliminary work, we have looked only
at intonation. We intend to employ naïve transformations also in duration within this thesis. One desirable in S2SMT is to match the original speaker's duration at the sentence level. This can help the audio overlay on the video easier without having to shrink or stretch the synthetic utterance beyond perception.

Traditional approaches to S2SMT use the pipeline architecture where speech from a source language is passed through an automatic speech recognizer (ASR). The ASR hypothesis 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 translated text in the 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 in 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 thesis, we want to further exploit the source prosodic information by imposing it appropriately on the target side after translation, in essence transferring the affect 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 4.1 situates the problem within the framework of speech translation. We propose to address this problem from by learning how prosodic correlates of prominence patterns change across languages, considering the case of English<->Portuguese translation for our work. We have already created a database of parallel speech in these languages. We propose to use word alignment and accent prediction techniques, and map the prominence patterns across this language pair.
4.1 Preliminary Work: Parallel Speech Corpora

In order to learn the right mapping from the source language intonation to the target language, we need 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 explained. 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 easily available for speech. It is also challenging to setup control for the requirement of similar affective 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.

For the text prompts to record, we used the TAP-UP parallel text corpus which provides semantically parallel articles in both English and Portuguese. The corpus TAP-UP was built on the content of the UP Magazine\(^1\), the in-flight magazine of TAP Portugal airlines. The magazine covers a wide range of subjects including scenery, cuisines and places, from worldwide destinations, published in Portuguese and English. Since we intend to capture speaker prominence in as naturally as possible, we select paragraphs from the corpus instead of sentences to provide more context for the speaker for delivery.

\(^{1}\)http://upmagazine-tap.com/
We selected 90 paragraphs which have the same number of sentences on either language, to enable easy analysis. We used the text selection routines of the Festvox framework [Black and Lenzo, 2000] to select an optimal subset of phonetically balanced paragraphs from the TAP-UP corpus. These parallel articles were recorded by a native Portuguese speaker who is fluent in both English and Portuguese. The recordings were done at the paragraph level alternately in both the languages, so that the same paragraph is recorded in both the languages at once. The speaker first recorded in Portuguese since comprehension of the text is much easier in speaker’s the native language. The choice of recording at the paragraph level is deliberately made to allow the speaker enough context to lay appropriate and natural emphasis during his delivery. Also, recording semantically identical paragraphs in both languages facilitates as parallel delivery as possible (of prominence patterns etc.).

These paragraph level utterances were automatically chunked at the sentence level and are phonetically segmented using the \texttt{islice} module [Prahallad and Black, 2011] within Festvox. The total duration of the speech is approximately 1 hour in each language. The corpus statistics are presented in the table 4.1.

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Portuguese</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Paragraphs</td>
<td>89</td>
<td>89</td>
</tr>
<tr>
<td>#Sentences</td>
<td>420</td>
<td>420</td>
</tr>
<tr>
<td>#Tokens</td>
<td>8184</td>
<td>8211</td>
</tr>
<tr>
<td>#Words</td>
<td>2934</td>
<td>3283</td>
</tr>
<tr>
<td>#Tokens/sentence</td>
<td>19.48</td>
<td>19.55</td>
</tr>
<tr>
<td>Duration(mins)</td>
<td>60.36</td>
<td>59.47</td>
</tr>
</tbody>
</table>

To study the intonations in corresponding words, it is necessary to get word alignment information is necessary. We used GIZA++ [Och and Ney, 2000] to align the sentences in both the languages. A word alignment model trained on the Europarl data [Koehn, 2005] using the Portuguese and English language parallel data is used to find the alignments between the corresponding words in the two for the sentences in the TAP-UP database.
4.1 Preliminary Work: Parallel Speech Corpora

Table 4.2: Statistics of AuTOBI accent probability on different analysis units

<table>
<thead>
<tr>
<th>AuToBI analysis level</th>
<th>Prominent word Mean</th>
<th>Prominent word Std.dev</th>
<th>Non-prominent word Mean</th>
<th>Non-prominent word Std.dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>0.353</td>
<td>0.285</td>
<td>0.153</td>
<td>0.222</td>
</tr>
<tr>
<td>Primary stressed syl</td>
<td>0.498</td>
<td>0.271</td>
<td>0.331</td>
<td>0.269</td>
</tr>
<tr>
<td>Secondary/ unstressed syl</td>
<td>0.357</td>
<td>0.274</td>
<td>0.252</td>
<td>0.249</td>
</tr>
</tbody>
</table>

4.1.1 Analysis of Prominence

One of the main sources of affect in speech is word prominence. To detect prominence in the source speaker’s speech, we used the accent prediction modules of AuToBI [Rosenberg, 2010]. These modules find the likelihoods of having an accent of each word and syllable, given the speech. To evaluate the usefulness of AuToBI as a prominence predictor we used the data from speaker awb of the CMU Arctic databases. We used the prominence patterns for this data provided in [Yu et al., 2010]. The accent prediction module takes as input the speech, and the time boundaries within which analysis is carried out and outputs i) the likelihood of having an accent within the interval, and ii) the predicted ToBI shape of the accent. While accent is not a sufficient or necessary cue to word focus, in our preliminary work we used F0 alone to convey focus, and limit our study only to prominence conveyed through changes in F0.

Table 4.2 shows the statistics of the AuToBI detected accent probabilities in the database. The statistics are presented separately for the Prominent and Non-Prominent words. The human annotations of prominence are used as the references in each case. To determine the right level of analysis to model prominence, we tried both syllables and words as analysis units. Word accent probabilities are probably the most discriminative to tell if a word is accented or not. Within a word with the high accent probability, the primary stressed syllable is more likely to have an accent than the secondary stressed syllable. While these are rather basic assumptions of accents in the F0 contour, the numbers presented show the sanity of the automatic technique to determine prominence (as conveyed through F0). Such a score may be used as a feature in prediction of the right level of prominence as used in [Aguero et al., 2006]. While such features may be useful for duration models, F0 contours are much varied and the feature may not get selected at all during
training the models due to sparsity problems. But the experiment determines that words are more important when it comes to prominence, as they are relatively more discriminative than syllables, with respect to intonation. We used this knowledge to best exploit the parallel speech data we have.

The experiment also suggests that in natural speech, prominence is subtle and it may be sub-optimal make a binary decision as to weather a word is prominent or not. So rather than using only the prominent words, in preliminary work reported here, we learn from all pairs of content words. Even the transformation is done on all content words, so as not to introduce any additional errors due to misdetections in word prominence.

4.1.2 Cross-lingual F0 conversion

Using the SPAM intonation model (Chapter 2), both the English and Portuguese databases were used to create respective models of intonation. For each corresponding content word in both the languages, the accent components were analyzed and parameterized as Tilt vectors for the duration of the word. This gave us parallel data of word-level Tilt vectors to train a conversion function using Gaussian Mixture Models as seen in the preliminary speaker transformation experiments of Chapter 3.

The trained function was then used to transform the intonation contour of a novel utterance in the source language, given its translation and word alignment information to an intonation contour of the target language. We examined if this approach transferred the right affect that was used in the original speech to the target language. Using the phrase/accent decomposition method, the original speaker's intonation is decomposed into long term trend local accents components. For each content word, the residual accents were analyzed and assigned a 4-dimensional Tilt parameterization. The trained GMM mapping function is applied on the data parameters and the target Tilt values are predicted for the corresponding word after translation. At synthesis time in the target language, the default SPAM intonation models predict an F0 contour for the translated sentence. For the content words however, the converted Tilt parameterizations are resynthesized and overwriting the default predicted F0s. Appropriate smoothing was done to make the content-function word transitions coherent.

The idea is elegant in that we were able to model the continuous range of affect as opposed to categorical treatment that is usually employed for expressive or emotional speech synthesis with separate models for prominent and non-prominent
4.2 Proposed Work

4.1.3 Initial Experiments Transferring Prominence Across Languages

The phonetically segmented speech databases of the parallel data are used to create Clustergen SPSS voices in both the languages. This means that two distinct voices for the speaker in each language. SPAM intonation models were employed to effectively capture the distinct accents employed by the speaker.

A test set of 10 new sentences was prepared for evaluation of the proposed transformation technique. The same bilingual speaker was asked to record different instances of each sentence varying the focus word in each delivery. A total of 38 utterances of the test sentences were recorded. The sentences were automatically translated into Portuguese and intonation parameters of the original speech were analyzed to be used as the source vectors for transformation. The transformation function learnt on the parallel speech corpus was applied on the synthetic translated utterances. The resulting intonation contours were imposed and the sentences were synthesized. An illustration of one of the test sentences with varying concepts in focus is provided in the Table 4.3 of F0 contours. The text spoken is also shown, where _ and **bold face** are used to refer to the word assigned for the speaker to emphasize. It is clear that the transformation is doing reasonable transformations in most cases. Informal listening tests of the synthesized English sentences with the predicted intonation also reveal that the same words sound focused after translation as in the original Portuguese sentence.

4.2 Proposed Work

In the preliminary work presented, we have shown that we can learn systematic prosodic transfer functions from a corpus of parallel speech. Even with rather naïve transformation functions directly applied on the content words, we were able to produce understandable and appropriate transformations and realize a degree of affect transfer.

We hope to make develop more robust cross-lingual prosody transformation techniques within the scope of this thesis. One challenging issue with the S2SMT task is the cascading of errors through the pipeline. ASR hypotheses errors are
further exaggerated by SMT translations making the the TTS input malformed, both in syntax and meaning. Parlikar et al. [2010] have proposed a phrasing strategy to improve the understandability of SMT outputs that have broken syntax. Adell et al. [2012] have gone further proposing strategies for handling disfluencies filled pauses in the source speaker’s utterance. We intend to handle intonations in a similar way based on confidence scores from ASR/SMT modules etc. Some directions we wish to pursue include —

- Training transformation functions only on prominent words, as opposed to pairs of all content words.
- Investigating techniques for cross-lingual prosody transfer using non-parallel speech data.
- Appropriate handling of filled pauses/filler words for the case of English⇔Portuguese S2SMT.
- Affect sensitive automatic dubbing of TED talks.
- Graceful degradation of S2SMT in the presence of ASR/SMT errors.
- Application of the proposed techniques on another language pair.

4.3 Summary

In this chapter, we presented the problem of cross lingual transfer of affect in Speech-to-Speech Machine Translation taking the example of translation between English and Portuguese. A parallel speech database in these languages is created. The parallel speech corpus is used to learn a mapping function (as shown in the previous chapter) between F0 contour shapes over corresponding content words. As a proof of concept experiment, we were able to apply this function on unseen utterances of the source language and predict appropriate contours for translated utterances in the target language. Parts of this research are communicated to Interspeech 2012. Further evaluations are planned on automatic dubbing of affect in TED lecture videos within the English⇔Portuguese language pair.
Table 4.3: Input Portuguese F0 contours (on left) and predicted English F0 contours (on right) for the same sentence with varying prominence on the input side as indicated in bold face.

<table>
<thead>
<tr>
<th>Input Portuguese F0</th>
<th>Predicted English F0</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Primeiro</em> surgiu a água, depois o óleo de rosas.</td>
<td><em>First</em> came the water and then the rose oil.</td>
</tr>
<tr>
<td><img src="image1.png" alt="Graph" /></td>
<td><img src="image2.png" alt="Graph" /></td>
</tr>
<tr>
<td><img src="image3.png" alt="Graph" /></td>
<td><img src="image4.png" alt="Graph" /></td>
</tr>
<tr>
<td><img src="image5.png" alt="Graph" /></td>
<td><img src="image6.png" alt="Graph" /></td>
</tr>
<tr>
<td><img src="image7.png" alt="Graph" /></td>
<td><img src="image8.png" alt="Graph" /></td>
</tr>
</tbody>
</table>
Chapter 5

Evaluating Intonation

The annual Blizzard challenge [Black and Tokuda, 2005] evaluates several TTS systems built on the same speech corpus through human listening tests. Mean opinion scores (on a scale of 1-5) are used to rate competing TTS systems. ABX preference tests are also done to rate closeness of synthetic speech samples to reference natural speech. These evaluations are done on overall synthesis, i.e., human perception of both the spectral and prosodic quality. In the preliminary work reported in Chapter 2, we have used similar subjective measures to compare intonation models. But for evaluation of intonation transformation (Chapter 3) we used objective metrics for testing the performance.

While preference/identification tests by human listeners is the ultimate test for both TTS and voice conversion techniques, Clark and Dusterhoff [1999] find that listening subjects can distinguish between different intonation contours but cannot distinguish between the more subtle differences between non-identical intonation patterns. This means that perceptual tests are probably not useful to judge small improvements to localized F0 contours, as they are overshadowed by gross errors in intonation. It is hence difficult to setup listening tests where subjects identify who the speaker is, only from subtle differences in intonation. This calls for reliable objective measures to automatically evaluate intonation contours (and models). However, as suggested in [Clark and Dusterhoff, 1999] and the current work existing objective measures like root mean squared error and correlation do not correspond with human perceptual preferences.

The objective measures used in practice like root mean squared error and correlation are computed as averages over all the frames over an utterance, usually excluding the pause regions. In human listening, we perceive higher level features
like peak placement, pitch targets on prominent words etc. Most other artefacts related to microprosody and those on function words may be perceptually irrelevant, and perhaps should be excluded for comparing different intonation contours. An alternate way is to parameterize the contours and compute errors in the space of the new representation (like Tilt), that is perhaps more style and expression sensitive.

Another evaluation unique to this thesis will be that of cross-lingual affect transfer. Subjective evaluations in this part of the work will include human judgements of video segments using the proposed work and without it. Objectively, we intend to have automatic prominence detection techniques on the source and target speech to evaluate how close the prominence patterns are across corresponding concepts in both the languages.

Since we deal with all aspects of intonation, include its modeling, and transformation. We propose to re-examine and improve existing evaluation strategies for the different sub-problems we deal with in this thesis. Given the ease of conducting subjective evaluations through crowd-sourcing [Parlikar and Black, 2012], we want to conduct large scale listening experiments and design objective measures that correlate well with subjective human judgements.

The use of good objective measures is not merely to judge intonation models but in training models. Statistically trained intonation models benefit from a good optimization criterion and we hope the new objective metrics yield better models that will also generalize well for the unseen test set.

### 5.1 Proposed Objective Evaluations

Following are some directions we would like to pursue for objective measures—

- Using only the regions corresponding to the content words within the reference and testing F0s for RMSE/correlation measures.

- Computing distances between parameterized F0 contours rather than directly on the natural contours.

- Using automatic prominence detection techniques to detect compare prominence patterns in source and target speech.
5.2 Proposed Subjective Evaluations

We propose the following tests within subjective evaluations—

- Evaluation of expressiveness in synthetic voices.
- Subjective evaluation of identity for testing intonation transformation.
- Human classification of speaking style within style transformation.
- Evaluation of affect transfer in automatically dubbed lecture videos.
Chapter 6

Thesis Timeline

6.1 Estimated Timeline:

<table>
<thead>
<tr>
<th>Task</th>
<th>Estimated Completion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposal</td>
<td>May xx, 2012</td>
</tr>
<tr>
<td>Extensions to SPAM Model (Ch 2)</td>
<td>May – Jun 2012</td>
</tr>
<tr>
<td>Affect Transfer (Ch 4)</td>
<td>Jul – Oct 2012</td>
</tr>
<tr>
<td>Evaluation (Ch 5)</td>
<td>Nov – Dec 2012</td>
</tr>
<tr>
<td>Identity Transformation (Ch 3)</td>
<td>Jan – Mar 2013</td>
</tr>
<tr>
<td>Style Transformation (Ch 3)</td>
<td>Jan – Mar 2013</td>
</tr>
<tr>
<td>Thesis Writing</td>
<td>Apr – Jun 2013</td>
</tr>
</tbody>
</table>
### 6.2 Proposed Tasks

An assimilated list of the proposed tasks within this thesis—

| Intonation Modeling | Find the right level to model — phoneme, syllables or metrical feet?  
|                     | Joint training and estimation of phrase and accent components.  
|                     | Characterization of identity and style within model parameters.  
|                     | Incorporating features beyond sentence for Audiobook synthesis.  
|                     | Building SPAM intonational models for TTS in several languages.  
| Intonation Transformation | Evaluation of various phonological anchor units for conversion.  
|                          | Extensions for speaking style and dialect transformation.  
|                          | Learning intonation mapping from non-parallel speech data.  
| Cross-lingual Transfer | GMM-conversion only on prominent words in parallel speech corpora.  
|                          | Techniques for training from non-parallel speech data.  
|                          | Appropriate handling of filled pauses in speech translation.  
|                          | Automatic dubbing of TED talks with affect within English↔Portuguese.  
|                          | Testing the techniques on at least another language pair.  
| Evaluation | Perceptual tests for evaluating affect and expression.  
|             | Perceptual tests for identification of speakers through intonation.  
|             | Perceptual tests for speaking style identification through intonation.  
|             | Preference tests for evaluation of TED video segments dubbed with affect.  
|             | Considering only content/prominent words in error computation.  
|             | Computing errors on parameterized contours rather than raw F0 contours.  
|             | Objective metrics for identity and speaking style transformation. |


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. 2


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


