# Perceptual Cost Function for Cross-fading-based Concatenation

Qi Miao
B.S., Tsinghua University, Beijing, China, 2000
M.S., Tsinghua University, Beijing, China, 2003

A Master's thesis submitted to the faculty of the
Oregon Health & Science University
in partial fulfillment of the
requirements for the degree
Master of Science
in
Computer Science and Engineering

March 2012

The Master's thesis "Perceptual Cost Function for Cross-fading-based Concatenation" by Qi Miao has been examined and approved by the following Examination Committee:

Jan P. H. van Santen
Ph. D., Professor
Thesis Research Advisor

Esther Klabbers-Judd
Ph. D., Assistant Professor

Alexander Kain
Ph. D., Assistent Professor

# Dedication

To my mom and dad for their everlasting love.

# Acknowledgements

First and foremost, I offer my most sincere gratitude to my master's thesis supervisor, Dr. Jan van Santen, who has supported me over the years with his patience and knowledge. I would not have completed this thesis without his encouragements and efforts. Jan has given me so many inspirations about speech synthesis technology and has guided me through many successful and failed projects. I wish I could have learned more from him.

I would also like to thank many faculty members at the Center for Spoken Language Understanding. Thank you Esther Klabbers-Judd, Alexander Kain, and John-Paul Hosom for all your support and help in research discussion, experiment set up and perceptual experiment. Thank you Rachel Coulston for being my speaker for the speech corpus in this thesis. Thank you Peter Heeman to make this master's thesis to happen.

In my daily work I have been blessed with a friendly and cheerful group of fellow students and staff members. I would like to pass my gratefulness to Xiaochuan Niu, Akiko Kusumoto-Amano, Emily Tucker Prud'hommeaux, Taniya Mishra-Linger, Kristy Hollingshead Seitz, Fan Yang, Maider Lehr, Patricia Dickerson, and more.

I would like thank my husband, my best friend and my life partner, Jue Wang, for his love and support. Finally, I want to thank my son, Wesley, for filling my life with love and laughter.

# Contents

# Abstract

**Perceptual Cost Function for Cross-fading-based Concatenation**

**Qi Miao**

**Supervising Professor: Jan P. H. van Santen**

Concatenative synthesis is currently the most widely-used Text-to-Speech (TTS) framework. However, it suffers from the problem that it can not guarantee to minimize both the target cost and the concatenation cost at the same time. As a result, the selected units for concatenation may come from totally different phonemic and prosodic contexts, which can lead to audible discontinuities in the output speech at the concatenation points. Various speech modification methods have been studied and applied during concatenation. In most cases, they can create a locally smooth transition between two units, but the resulting speech may be far from the target. In a previous study, a linear cross-fading weight function was used to remove spectral and time domain discontinuities during concatenative speech synthesis. We learned that concatenation through a linear weighted cross-fading function can produce smooth, yet unnaturally shaped formant trajectories; in addition, we noted that the precise details of how to cross-fade a specific pair of units may be highly context dependent.

We propose a new algorithm that uses a unit-dependent parameterized cross-fading weight function to create more natural-looking formant trajectories and, it is hoped, better-sounding output speech. The proposed algorithm uses a perceptually-based objective function to capture differences between cross-faded and natural trajectories across

the whole region of the phoneme, and uses the phoneme identity, prosodic contexts, and acoustic features of the units to predict optimal cross-fading parameters. This thesis reports a study on the feasibility of developing such perceptual cost functions. A special corpus was designed to produce a variety of shapes of formant frequency trajectories in different linguistic environments. A perceptual experiment was performed to determine whether we could predict perceptual quality of output speech from acoustic distance measures. We generated a range of synthetic/natural stimulus pairs, where the synthetic stimuli were generated using three types of cross-fading models, applied to different regions in the vowel. The results show that the perceptual cost function can be reliably predicted from the distance measures. Moreover, the results support our hypotheses that: a) the quality of the output speech is influenced by the shape of formant trajectories in the entire region across the vowel; and b) human perceptual scores are correlated to both the absolute distance and the first derivative of the absolute distance of the formant trajectories.

# Chapter 1

# Introduction

## 1.1 Speech synthesis overview

Speech is an important verbal communication skill between humans. For centuries, speech has been extensively studied by researchers and scientists. For more than fifty years, people have been working on trying to generate artificial speech that sounds exactly like the natural speech produced by human beings. Researchers and scientists have developed different types of speech synthesizers to generate speech sounds.

Since the development of computers in the 1970s, researchers have sought to create a fully automatic speech synthesizer that utilizes computers. A *Text-to-Speech (TTS)* system is a computer-based technology which automatically converts the input text into a speech signal. In order to generate audible speech from a textual representation, a TTS system first converts text into a linguistic representation, which is then used to generate an appropriate acoustic waveform. This second step is achieved by using a speech synthesis model that describes the relationship between linguistic units and acoustic features. Figure 1.1 shows the basic function blocks for a TTS system.

## 1.2 Speech synthesis methods

When generating audible speech from a textual representation, a TTS system first converts text input into a linguistic representation, which is then used to generate appropriate prosodic features and an acoustic waveform. This second step is achieved by using a speech synthesis model that describes the relationship between linguistic units and acoustic

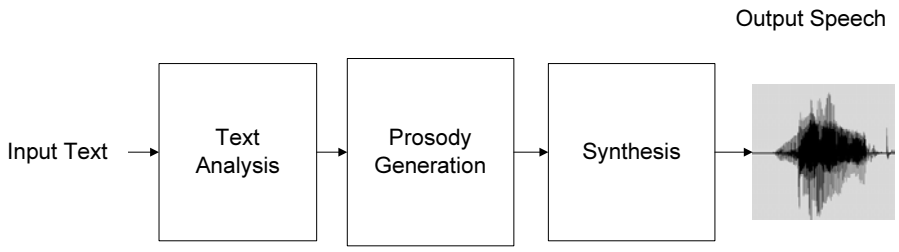


Figure 1.1: A basic function structure for a TTS system

features. These speech synthesis models vary in their complexity. In general, there are two types of synthesis models: *parametric synthesis* and *concatenative synthesis*.

### 1.2.1 Parametric synthesis

A parametric synthesizer usually consists of two major components: 1) a voice source and 2) a vocal tract model. In this method, multiple parameters are fed into the synthesizer to generate speech. A parametric synthesizer offers great control and flexibility on the output speech. On the other hand, this method requires a deep understanding of the human speech production procedure. A poor approximation of the voice source and vocal tract will result in robotic, buzzy, and unnatural voices. The very first device to generate speech sounds was called the *VODER* [6] which was demonstrated by Homer Dudley in the late 1930's. The first intelligible parametric synthesizer used an approach called *formant synthesis*, which utilizes relatively simple models of the glottal source and vocal tract. Model parameters can be generated either by rules [6] or from a database [8]. The most famous commercial formant synthesizer is *DECtalk* which is still widely used in many TTS applications. In parametric synthesis, most aspects of speech are controllable, including the degree of articulation and the characteristics of the speaker. The resulting speech is highly intelligible, but is often judged to be not very natural. In an effort to increase naturalness without decreasing flexibility, researchers have increased the complexity of the speech details in the speech production process in an approach called *articulatory synthesis* [13]. Articulatory synthesis attempts to generate the speech production system directly from the set of parameters for the human articulators. Unfortunately, it has proved difficult in practice to generate the high-dimensional parameter trajectories necessary to drive articulatory synthesis models, because the relationships between linguistic units and parameter trajectories are complicated and cannot be learned easily. Both formant and articulatory synthesis are examples of *parametric synthesis*.

### 1.2.2 Concatenative synthesis

*Concatenative synthesis* is the most common and successful TTS approach to date. In this method, the synthesizer generates speech by connecting pre-recorded speech units together

to form phrases and sentences. There are two types of concatenation methods as shown in Figure 1.2. One method is to record a *N-Phone* based inventory which consists of the minimal set of speech sounds. In this case, a diphone inventory is a popular choice. A diphone is a speech unit that contains the transition part of two phonemes. For example, the diphone "A-t" contains the second half of the phoneme "A" and the first half of the phoneme "t". The cut point inside each phoneme is considered to be the most stable part for each phoneme. Therefore, the resulting concatenation has fewer discrepancies. In these speech corpora, only one example of each N-phone unit is stored. During synthesis, prosodic features, such as pitch, duration and intensity, must be modified to match the target speech properties. Because of the extensive speech modification during synthesis, it is very difficult to achieve high quality output speech.

In another type of concatenation, known as *unit-selection*, the system performs a search in a pre-recorded speech database to find the best matched sequence of units for the target speech by optimizing a two-part cost function:

- *Target cost*: the difference between the selected units and the target speech
- *Concatenation cost*: the difference between the adjacent selected units during concatenation

The selected units are concatenated together to generate the final speech. In a perfect scenario, the selected units would have no concatenation errors and should sound exactly like the target natural speech. In reality, it is very hard to achieve because although the output speech is highly intelligible and natural in most cases, there are always concatenation errors due to the limited size of the speech corpus.

Speech modification is usually required to reduce these discontinuities. To overcome the problems of limited content and discontinuities, researchers have tried to (1) increase the size of the speech corpus to cover all possible combinations of the target unit sequences [2] or (2) apply additional modeling to modify both the prosody and the speech spectrum. The first approach is usually time consuming and expensive in most cases. Most commercial TTS systems record neutral flat speech to reduce the differences between units as much as possible. When highly expressive speech is desired, the recordings

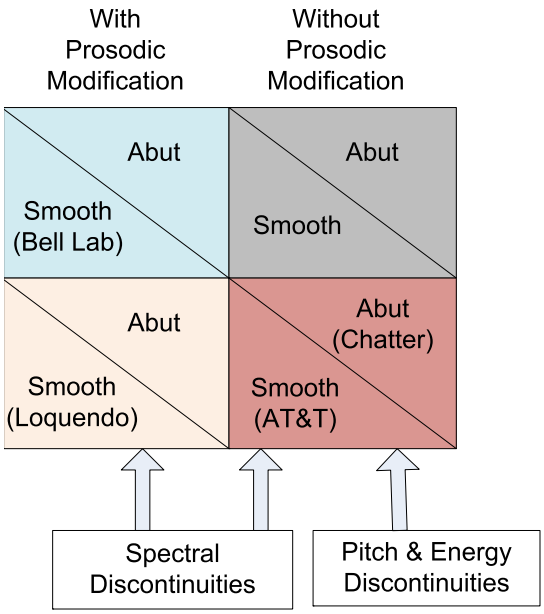


Figure 1.2: Current concatenation methods

have more dynamics and a larger search space for unit-selection. The voice quality of the speaker can also change over time [4] which degrades the consistency and the quality of the synthesized speech.

More and more people require a personalized TTS system. For example, a person with a speech or language disorder who uses an Augmentative and Alternative Communication (AAC) device with a TTS system might prefer using his own voice in order to preserve his identity. Unfortunately, it is not very practical in these cases to acquire a personalized TTS since not everyone can record a steady and clean corpus.

Concatenation errors occur both in speech prosody and in the spectral domain. To eliminate the errors in prosody, global pitch and duration models are commonly built [15, 16]. To reduce spectral discontinuities, researchers have studied smoothing spectral balance discontinuities at concatenation points, expressed as energies in four bands [9], smoothing formant discontinuities [10, 1, 7], and applying a fusion-unit approach during the concatenation [18]. All of these studies try to achieve the goal of reducing spectral discontinuities with smooth transitions at unit concatenation points without considering the natural global shape of the prosody and the spectral features. The resulting speech

sounds smooth and natural when the target speech has more dynamics and larger ranges in the feature space. For instance, when the synthesis of emotional and expressive speech is desired, the naturalness of the synthetic speech is largely influenced by the global shapes of the pitch contour, duration profile, and spectral features.

### 1.2.3 Hybrid method

Both parametric and concatenative synthesis have their advantages and disadvantages. They can both generate very intelligible speech. Typically parametric synthesis requires smaller storage space and runs more quickly because search operations are not needed. However, it is still difficult to provide accurate acoustic and linguistic parameters and synthesis rules for all purposes of synthetic speech. The result is a lack of naturalness. Concatenative synthesis, on the other hand, has the ability to generate highly natural speech as long as the corpus has good coverage of all acoustic and linguistic units.

In recent years, the *Hidden-Markov-Model (HMM) method* [19] has become popular research area in speech synthesis. In this method, spectral and excitation parameters are extracted and trained by a HMM model for each phoneme. Then, during synthesis, the model outputs a series of smoothed pitch, duration, spectral, and excitation parameter profiles for each phoneme that are used generate the final speech waveforms. This approach typically generates very smooth and intelligible speech, but the speech sounds can also be buzzy and unnatural.

Some researchers have also sought ways to combine the benefits of concatenative speech synthesis and rule-based system. Pearson emphet al. [12] showed that a synthesizer that combines concatenative and rule-based formant synthesis improved the speech intelligibility and naturalness. However, the discontinuity issue still remains in such systems.

## 1.3 Speech Modification through Concatenation

Concatenative synthesis is currently the most widely-used TTS method. The main problem facing concatenative synthesis is that units that are concatenated are generally recorded in different phonemic and prosodic contexts, which can lead to audible discontinuities at

the concatenation points. Various speech modification methods can be applied to the selected units to generate the prosody of the target speech and reduce the discontinuities. As described in the previous section, there are two types of discontinuities: 1) discontinuities in prosody (pitch and duration), and 2) discontinuities in the spectrum (spectral balance, formant frequencies, formant bandwidth, and overall intensity). Over the years, many methods have been developed to solve the problem. A successful speech modification method should be able to locally remove all of the audible discontinuities between units and produce a highly natural global shape of the prosody and spectrum relative to the target speech.

Although the existing unit-selection algorithm already takes the concatenation cost into consideration in the cost function, there are still discontinuities in the synthesized speech. Generally speaking, there are two types of discontinuities for concatenated units: 1) in the prosody domain where the pitch and duration are unmatched; and 2) in the spectral domain where the spectral structures are mismatched. All of these discontinuities will result in audible distortions in the synthesized speech. Many approaches have been proposed to reduce the discontinuities in prosody by building global pitch and duration models. In this dissertation, we focus on removing discontinuities in the spectral domain, i.e., spectral balance, formant frequency, formant bandwidth, and LPC parameters.

### 1.3.1 Concatenation through local smoothing

This method performs a local smoothing operation during concatenation to remove any sharp transitions between two units. But the problem is how to decide the appropriate overlap area between left and right units for each type of concatenation. If the overlap area is too small or there is no overlap area at all, it can produce very bumpy transitions at the concatenation point. Figure 1.3 shows an example of applying a simple local smoothing operation on the second formant trajectory of two units. The blue curve in the figure represents the left demiphone and he red curve in the figure represents the right demiphone to be concatenated. The black curve is the locally smoothed curve.

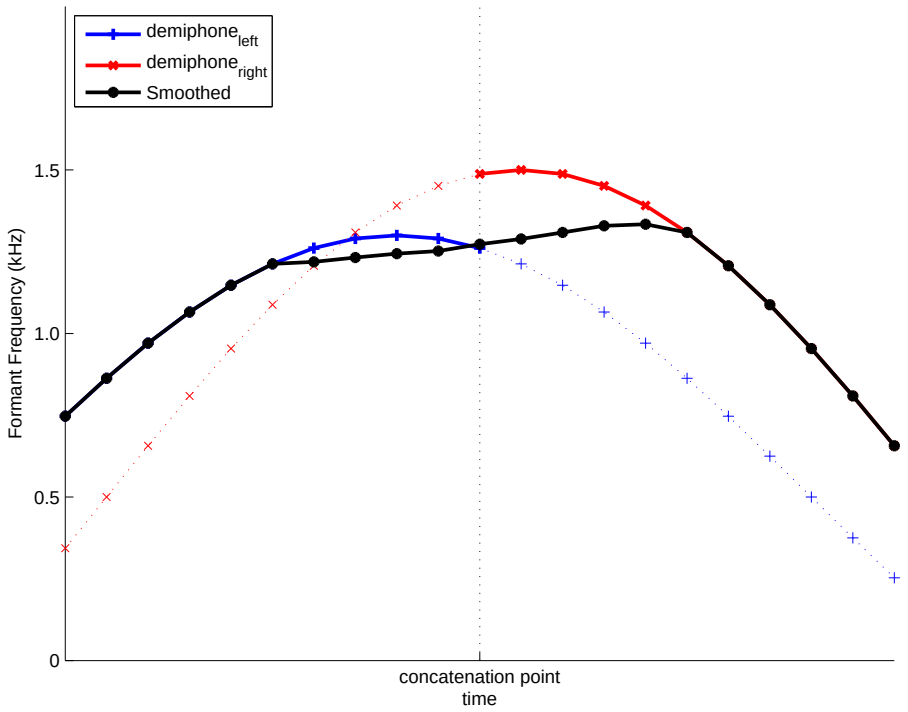


Figure 1.3: Concatenation with local smoothing

### 1.3.2 Concatenation through local interpolation

Olive [11] proposed to use straight line interpolation on the sets of LPC parameters in the transition between two phonemes. This interpolation was set by rules depending on the phoneme types, duration, and pitch information. The author also suggested that the straight line interpolation could be applied to other spectral features, such as formant frequencies. Figure 1.4 shows an example of applying straight line interpolation on the second formant trajectory of two units. The blue curve in the figure represents the left demiphone and he red curve in the figure represents the right demiphone to be concatenated. The black curve is the interpolated curve. Such a method works fine for units with small discrepancies but the quality is not satisfying for units with big discontinuities. It certainly raises the question of over-smoothing in the transition region. Moreover, when two units are too short, such as short vowels, there might not be enough data points to be interpolated.

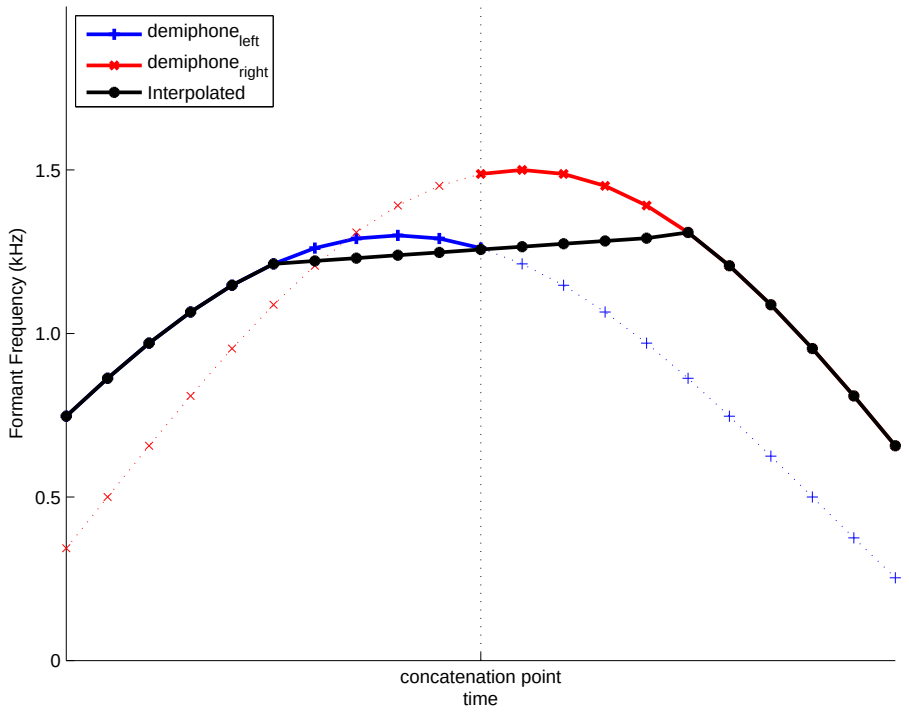


Figure 1.4: Concatenation with local interpolation

### 1.3.3 Concatenation with imposing target spectral dynamics

For both concatenation through local smoothing or local interpolation, a main requirement is that the original spectral mismatch should not be too large to maintain the natural spectral dynamics in the target speech during smoothing or interpolation. In this scenario, the selected units in the corpus should be close to each other and to the target speech. However, any unit-selection algorithm has to trade off between two cost functions: the concatenative cost and the target cost. As a result, the selected units may have large target costs with relatively small concatenation costs or have small target costs with large discontinuities between two units at the joining point. In either situation, a traditional spectral smoothing or interpolation method is not adequate. Therefore, some researchers apply spectral dynamics constraints from the target speech during smoothing to provide global spectral shape control while reducing the local concatenation errors.

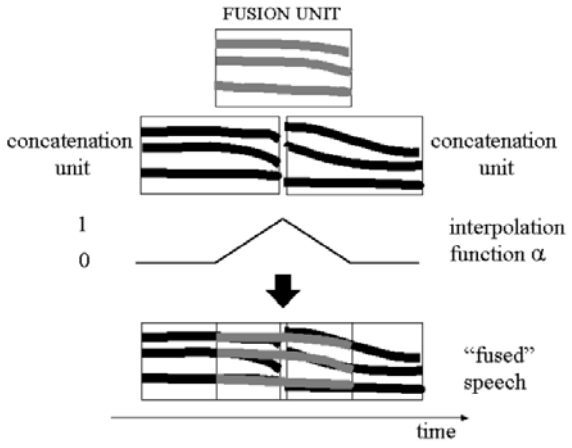


Figure 1.5: An illustration of applying a fusion unit during concatenation

Wouters [18] used a fusion unit (shown in Figure 1.5 [1]) to impose the natural spectral transition over the two concatenated units on the linear spectral frequency (LSF) trajectories. In this approach, a fusion unit is selected according to the phonetic and prosodic features of the target speech. The LSF trajectories are extracted from the joining units and the fusion units. During concatenation, the spectral dynamic constraints represented by the first derivative of the target LSF trajectories are linearly interpolated with the first derivatives from the joining units. This approach combines the spectral information from the target speech and concatenated units to impose a natural spectral transition during concatenation, result in in improved speech quality.

### 1.3.4 Concatenation with a linear weighted cross-fading function

In previous work [3], we applied a linear weighted cross-fading function to different speech features during concatenation. This method eliminates "points" of concatenation in favor

[1] This figure is taken from paper [17].

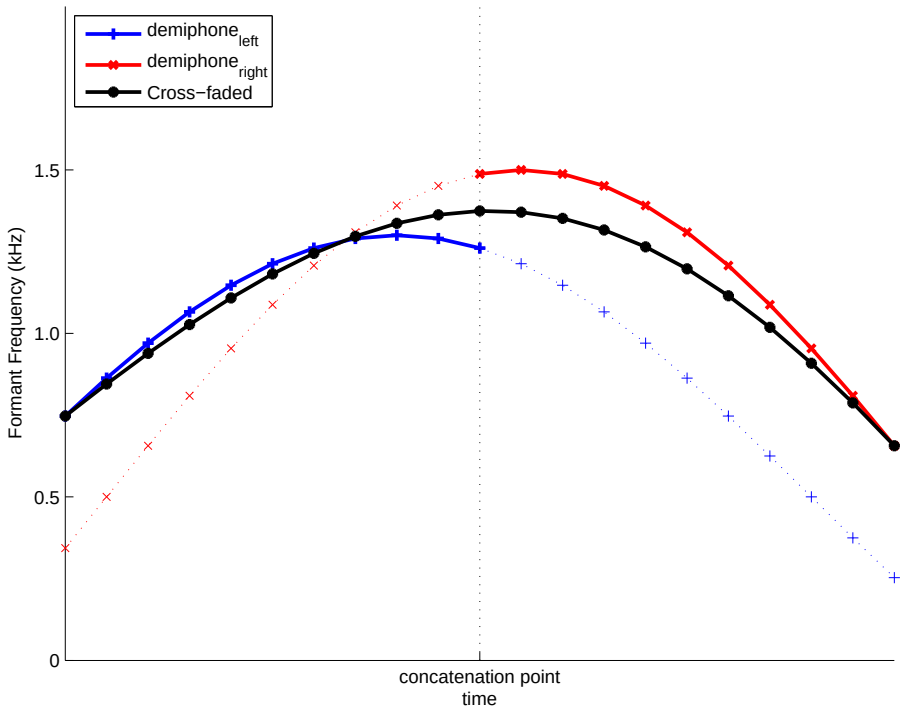


Figure 1.6: Concatenation with a linear weighted cross-fading function

of "regions" of concatenation by cross-fading (i.e. fading out one signal while fading in another) in various domains between the end and the beginning of two speech segments adjoining a concatenation. In this study, we first performed a linear weighted cross-fading operation on the spectral balance trajectories, the first three formant frequency trajectories, and the time-domain waveforms. Figure 1.6 shows an example of applying a linear weighted cross-fading function to the second formant frequency trajectory of two concatenated units. A perceptual test showed that all cross-fading operations significantly improved the perceived quality. However, using all three methods together is not significantly different from using time-domain waveform alone. We speculate that the shape of the cross-faded trajectories is not natural in some cases. For example, when two trajectories are sharply divergent as in Figure 1.7, the shape of the cross-faded trajectory is not natural. Moreover, these problems raise the question of how to optimize the cross-fading function based on the characteristics of concatenated units and target unit.

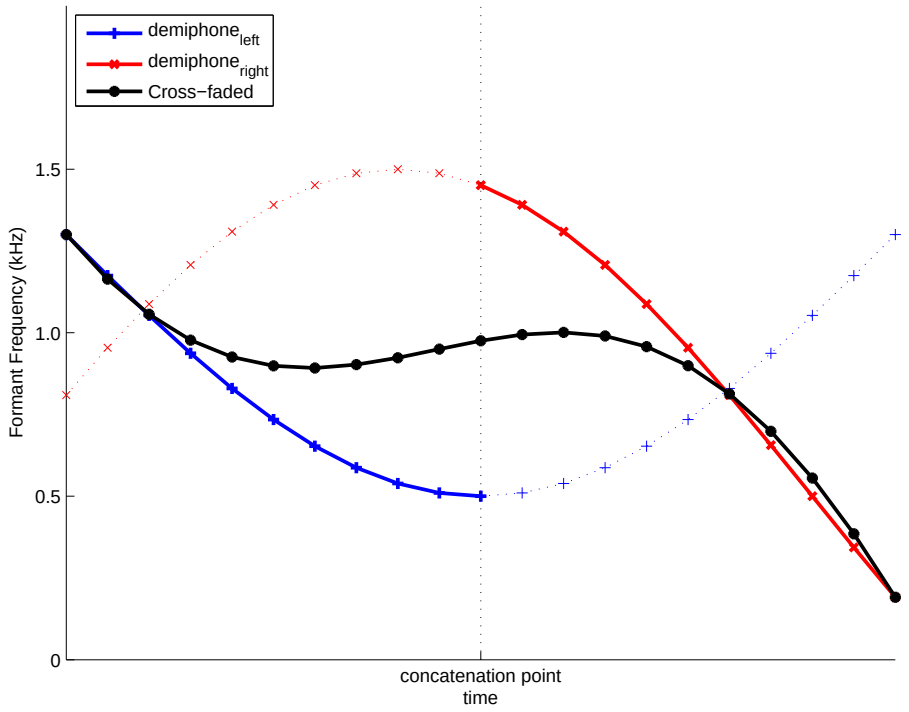


Figure 1.7: Concatenation with a linear weighted cross-fading function when two trajectories are divergent

## 1.4 The Methodology

The ultimate goal for a successful TTS system is to be able to generate perfect synthetic speech for any communication purpose for any individual. This goal requires a synthesizer that combines the strength of parametric and concatenative synthesis methods to achieve highly flexible, expressive, and natural speech. The main goal of our approach is to decrease the unnatural spectral discontinuities between two concatenated speech units, produce more naturally shaped feature trajectories for the target unit, and thereby increase the perceived quality of the speech by applying these trajectories during synthesis. Our motivation is to take advantage of all of the available information during synthesis, such as acoustic and linguistic features of the selected units and the target units, to provide both local concatenation error reduction and global spectral shape control. To that end, for any type of concatenation, our method can generate high quality speech as close to the target speech as possible.

We propose a new algorithm that uses a *unit-dependent trainable parameterized cross-fading weight function* to generate more natural-looking formant trajectories and, it is hoped, better-sounding output speech. The proposed algorithm:

- uses a perceptually-based objective function to capture differences between cross-faded and natural trajectories across the whole region of the phoneme, and
- uses phoneme identity, prosodic contexts, and acoustic features of the units to predict optimal cross-fading parameters to generate more natural formant trajectories.

This dissertation addresses the first part of the algorithm. Our hypotheses are:

- The quality of the output speech is influenced by the shape of formant trajectories in the entire region across the vowel.
- Human perceptual scores are correlated to both absolute distance and the first derivative of the absolute distance of the formant trajectories.

We designed a special corpus with a set of consonant-vowel-consonant (CVC) words that covers several dynamic ranges in speech prosody, such as duration and prominence.

The corpus also covers the most extreme areas of the vowel triangle. We chose three consonants (/k, b, l/) that have large coarticulation effects on the second formant. A perceptual experiment was performed and results were analyzed to confirm our hypothesis.

In Chapter 2, we explain our hypothesis and methodology to train a perceptually-based cross-fading function. Chapter 3 presents the speech corpus design and the experiment setup. Analysis of the experiment and results are discussed in Chapter 4. Chapter 5 summarizes the topic and discusses future work.

# Chapter 2

# Hypotheses and Methodology

In a previous study [3], a linear cross-fading weight function was used to remove spectral and time domain discontinuities during concatenative speech synthesis. In this method, smoothing was performed by cross-fading across a "region" of concatenation, instead of the traditional "points" of concatenation. In this chapter, we will first briefly introduce the experiment in which we applied a linear weighted cross-fading function to three speech feature domains. Then we focus on solving the problem in the first part of the algorithm.

Here we briefly review the previous work of applying a simple linear-weighted cross-fading function during concatenation for three aspects of speech: spectral balance (SBXF), formant frequencies (FFXF) and time domain speech waves (TDXF).

As mentioned previously, we aim to reduce concatenation errors by constructing smooth feature trajectories in the formant frequency and spectral balance domains, and then by modifying the natural speech signal accordingly. The construction of the feature trajectory was implemented by cross-fading the acoustic features of each speech frame across the entire phoneme that is involved in the concatenation operation. (We ignored atypical concatenations at phoneme boundaries.) Specifically, we considered the demiphone that followed the previous chunk and the demiphone that preceded the following chunk, giving us a double set of features over the entire phoneme region. Features were stretched or compressed by linear interpolation to match durations. The desired smooth feature trajectories $\mathbf{s}(t)$ were calculated by applying the equation $\mathbf{s}(t) = \alpha(t) \cdot \mathbf{r}(t) + (1 - \alpha(t)) \cdot \mathbf{l}(t)$, where $\mathbf{l}(t)$ and $\mathbf{r}(t)$ are feature vectors at time $t = 1 \ldots N$ of the last demiphone of the left chunk and the first demiphone of the right chunk, respectively, $N$ denotes the total number of data points in the cross-fade region, and $\alpha$ is the cross-fade function given by

$\alpha(t) = t/(N+1)$.

Figure 1.6 illustrates the concept of cross-fading using a linear weighted function on the second formant trajectories of two units. Without cross-fading, the final trajectory would be the concatenation of the solid left half-curve with the solid right half-curve, resulting in a large discontinuity. With cross-fading, the following demiphone of the left chunk and the previous demiphone of the right chunk are combined, resulting in a smooth final trajectory as indicated by the continuous curve. Cross-fading was implemented both in the formant domain on the first three formant frequencies and in the spectral balance domain. Comparing to concatenation through local smoothing (Figure 1.3) and concatenation through local interpolation (Figure 1.4), Figure 1.6 demonstrates the advantage of cross-fading over the traditional smoothing operation at the concatenation point. With a local smoothing or a local interpolation operation, the connected trajectory may have an unnatural transition and a strange trajectory shape when the original distance between two concatenated units is large.

From the experimental results we found that formant frequencies cross-fading (FFXF) alone was not as successful as time domain cross-fading (TDXF). We speculated that 1) the corpus used in the experiment was recorded in a constant phonemic and prosodic context, thus the original formant distance in the corpus was small, 2) the change of formant locations alone could introduce other artifacts during the signal modification procedure (SinLPC).

More importantly, we noticed that in some cases, the linear cross-fading weight function generated unnaturally shaped formant frequency trajectories. In addition, TDXF was not able to recover from the impact of spectral changes introduced by different prosodic contexts and acoustic features. These problems raise the question of how to optimize the cross-fade weight function based on the characteristics of concatenated units and target units.

Figure 1.7 shows one example of applying a linear weighted cross-fading function to two concatenated unit trajectories and their target unit trajectory. The blue and red lines represent the formant trajectories of the two units to be concatenated. These trajectories not only have large distances in the frequency domain, but also have sharply divergent

shapes. The black curve is the cross-faded trajectory. However, even though the cross-faded trajectory is perfectly smooth, the overall shape is quite unnatural. The cross-faded trajectory would result in an unnatural sounding speech output. A better cross-fading operation is needed.

We propose a new algorithm that uses a *unit-dependent trainable parameterized cross-fading weight function* to generate more natural-looking formant trajectories and, ideally, better-sounding output speech. The proposed algorithm:

- uses a perceptually-based objective function to capture differences between cross-faded and natural trajectories across the whole region of the phoneme, and
- uses phoneme identity, prosodic contexts, and acoustic features of the units to predict optimal cross-fading parameters to generate more natural formant trajectories.

Previous studies [14, 5] used perceptual data to predict the relationship between concatenation cost and audible distortions. These studies focused on discontinuities at the concatenation points. Our hypotheses are, similarly:

- *the quality of output speech is influenced by the shape of formant trajectories in the entire region across the vowel.*
- *human perceptual scores are correlated to both absolute distance and the first derivative of absolute distance between synthetic (cross-faded) and natural (target) formant trajectories.*

The specific goal of this study is to train a perceptual cost function for cross-fading based concatenation for formant frequencies. The cost function is determined by distance measures between the cross-faded trajectory and the target trajectory. As shown in Figure 2.1, we divided the formant trajectory for the vowel part into three regions: the first third (A), the second third (B) and the final third (C). The distances in regions A and C reflect the co-articulation influence caused by the surrounding phoneme. The distance in region B reflects the more steady-state part of the vowel formant trajectory (although, generally, this part is also co-articulated). We also define three regions in the first derivative of the formant trajectories.

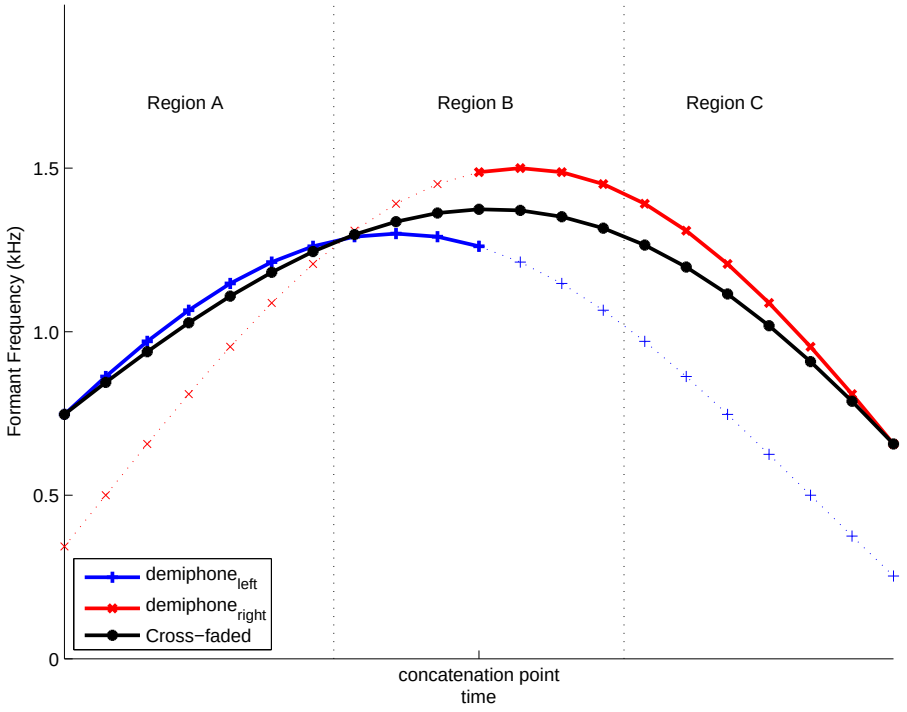


Figure 2.1: Cross-fading between two trajectories in three regions

# Chapter 3

# Data Collection and Experiment Set up

## 3.1 Speech Corpus

We recorded a corpus consisting of Consonant-Vowel-Consonant (CVC) words occurring in different prosodic contexts. The corpus was recorded by a female American English speaker. We selected six vowels (see Table 3.1)[1] which cover the most extreme areas of the vowel triangle and three consonants (/k, b, l/) which have large coarticulation effects on the second formant. The pre-vocalic and post-vocalic consonant in one CVC word could be the same.

| Vowels | Example |
|---|---|
| /i:/ | b*ee*t |
| /u/ | b*oo*t |
| /@/ | b*a*t |
| /ei/ | b*ay* |
| /aU/ | ab*ou*t |
| /aI/ | b*y*e |

Table 3.1: Vowels in the corpus

Each CVC word was put in two carrier sentences.

```
Please say the word /k i: k/ again.
Please DONT say the word /k i: k/ again.
```

In the first sentence, the CVC word is stressed and in the second sentence, the CVC word is unstressed. Both sentences were read at two different speaking rates: relatively

[1]Listed in Worldbet, an ASCII version of IPA.

slow and relatively fast. Therefore, each CVC occurs in four different prosodic contexts: 1) stressed and fast; 2) unstressed and fast; 3) stressed and slow; 4) unstressed and slow. Our intention is to generate different shapes of vowel formant trajectories caused by the linguistic context.

The corpus was recorded with an American female voice. There are a total of 3×6×3×4 = 216 CVC words in the corpus. Each CVC word was extracted from the original recordings. The formant frequencies for each CVC word were first calculated every 10ms using *Wavesurfer* plug-ins, then visually inspected and hand corrected by an expert. Only the first three formant frequencies were corrected and used in the perceptual experiment.

## 3.2 Perceptual Experiment

### 3.2.1 Stimulus Selection

We performed a search procedure to select the unit pairs which were used to synthesize the CVC words in the experiment. As we described in Chapter 2, for each formant trajectory, we define three regions: the beginning part (Region A), the middle part (Region B), and the ending part (Region C), as shown in Figure 2.1. Each region covers about one third of the trajectory. First we transform all of the formant frequencies to the Bark scale. Then we calculate the distances between two candidate units in the three regions by determining the Euclidean distance of the absolute distance of the formant frequencies ($D_{FF}$) and of first derivative distance ($D_{DF)}$) for the formant trajectories.The formant frequency distance ($D_{FF}$) was calculated as follows:

$$D_{FF}(unit_1, unit_2) = \sqrt{\sum_{k=1}^{n}(FF_{k,unit_1} - FF_{k,unit_2})^2}. \tag{3.1}$$

where $unit_1$ indicates the left unit and $unit_2$ indicates the right one, $FF_{k,unit_1}$ means the $k$th absolute formant frequency (in Bark) in the region of the left unit, $FF_{k,unit_2}$ means the $k$th absolute formant frequency (in Bark) in the region of the right unit.

The distance of the formant trajectories ($D_{DF)}$) was calculated as follows:

$$D_{DF}(unit_1, unit_2) = \sqrt{\sum_{k=1}^{n}(DF_{k,unit_1} - DF_{k,unit_2})^2}. \quad (3.2)$$

where $DF_{k,unit_1}$ means the $k$th first derivative formant frequency (in Bark) in the region of the left unit, $DF_{k,unit_2}$ means the $k$th first derivative formant frequency (in Bark) in the region of the right unit.

For each target unit, we calculated the maximum absolute distance and maximum first derivative distances in each region between two candidate units across all the concatenations available in the corpus. Then we normalized the distances in each distance measure to the range of (0, 1].

| Distance Measure | Description |
|---|---|
| 1 | Absolute distance in region A |
| 2 | First derivative distance in region A |
| 3 | Absolute distance in region B |
| 4 | First derivative distance in region B |
| 5 | Absolute distance in region C |
| 6 | First derivative distance in region C |

Table 3.2: Definition of distance measures for each formant trajectory.

Since the second formant trajectory usually has the strongest dynamics for vowel, we applied the following criteria on the normalized distances in the search on F2 to ensure good coverage of the different constellations of the 6 distance measures. Unit pairs were selected to have:

- small distances in all distance measures, or
- large distances in all distance measures, or
- a relatively large distance in each of six distance measures

For each vowel, two target units are selected. For one target unit, eight concatenation unit pairs are selected across the corpus. Every concatenation pair has one of the eight distance types. For every selected pair, we applied three types of concatenations: a) concatenation at the middle point of the trajectory with a linear weighted cross-fading

| Distance Type | Description |
|---|---|
| 1 | small distances in all distance measures |
| 2 | large distances in all distance measures |
| 3 | large distance in distance measure 1 |
| 4 | large distance in distance measure 2 |
| 5 | large distance in distance measure 3 |
| 6 | large distance in distance measure 4 |
| 7 | large distance in distance measure 5 |
| 8 | large distance in distance measure 6 |

Table 3.3: Definition of distance types for selected units

function, as used in our previous study [3]; b) concatenation at the middle point of the trajectory with a sigmoid weighted cross-fading function; c) concatenation through a *range selection function* where we always put the cross-fading area in the region containing the largest discontinuity between two concatenated units. Figure 3.1 shows the three cross-fading weighted functiosn we used during the concatenation. The red line shows the weighted function that applies to the left unit, and the blue line represents the weighted function that applies to the right unit. In the third concatenation type, we have six different cross-fading weighted functions. Each of these functions applies to one of the six regions we defined before. In combination, the total number of stimuli is 6 (vowels) $\times$ 2 (samples per vowel) $\times$ 8 (distance types) $\times$ 3 (three cross-fading weighted functions) = 288.

In order to eliminate effects from other features, such as pitch, duration, and energy, we re-synthesized the CVC words using a hybrid formant synthesizer with pitch, duration, and energy profiles imported from the target CVC word. The spectrum over 4KHz is copied from the target unit. Therefore, both utterances have highly similar acoustic features except for the first three formant trajectories. One CVC word was synthesized with the trajectories extracted from the natural target CVC word and the other one was synthesized with the trajectories generated by cross-fading models. The final test stimuli contained pairs of identical CVC words with a 200 ms-separating pause.

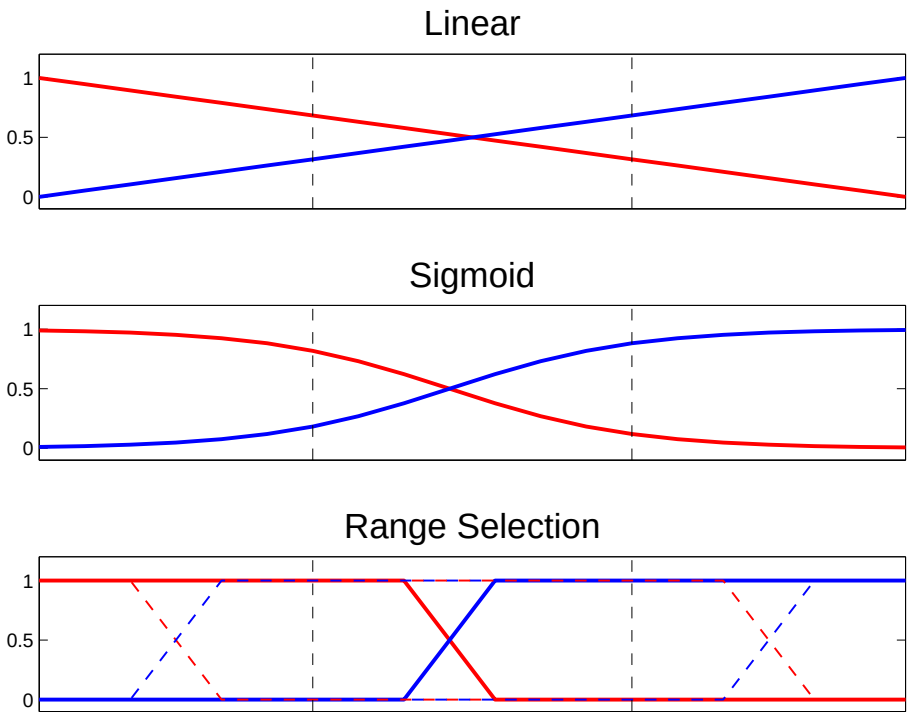


Figure 3.1: Three types of cross-fading weighted functions.

### 3.2.2 Experiment Set up

The experiment was set up as a Comparative Mean Opinion Score (CMOS) test. Eight expert subjects were asked to listen to pairs of CVC words and rate the quality of the CVC word "A" as compared to CVC word "B" on a five-point scale. A and B are the same CVC word synthesized by the same hybrid formant synthesizer. Quality was defined to include both naturalness and intelligibility of the word. The subject had to choose a score from: (-2) A sounds much better, (-1) A sounds better, (0) About the same, (1) B sounds better, (2) B sounds much better. The range of the scores is [-2, 2]. Only the voiced part of the CVC was synthesized. The unvoiced part was kept the same as the target CVC word. For the voiced part, the spectrum over 4KHz remained the same as the target CVC word. One word was synthesized using the formant trajectories from the natural speech and the other from cross-faded trajectories. The order of A and B was randomized. The experiment was performed in the CSLU Perception Lab with professional audio devices. During the experiment, subjects could repeat the stimuli as many times as they wanted

to make a selection. Subjects were allowed to take short breaks during the test if needed. The total time for one test was about 40 minutes.

# Chapter 4

# Results

Figure 4.1 shows the mean CMOS score for each vowel. The value of the score reflects how well the natural formant frequency trajectories compare with the cross-faded trajectories. On average, vowels /u/ and /aU/ have relatively lower scores.

To analyze the results, we define eight distance types for our experiment stimuli. Figure 4.2 shows the mean CMOS scores for each distance type, as defined in Table 3.3. On average, larger distances in any distance measure are expected to produce worse quality in the output speech, which is borne out by these results. The Figure shows that a larger distance in region A (the first third of the vowel formant trajectories) has the strongest impact.

To train the perceptual cost function for each experimental stimulus, we first transformed the scores from all subjects into normalized scores. Then we combined normalized score for all the stimuli into a data metric and applied principal component analysis (PCA) [9] to the scores. This analysis eliminates the effects of different individuals using larger rating ranges and also assigns larger weights to subjects more in agreement with other subjects. When transformed the normalized score into a weighted final score with the results from PCA. A multiple linear regression model between the PCA-based scores and distances in six different measures was trained for each vowel. The distance was calculated as the Euclidean distance between the cross-faded and natural trajectories in the frequency domain or in the delta-frequency domain, as appropriate. There are 288/6=48 data points per vowel and 19 parameters (six for F1, six for F2, six for F3, and one for the intercept) in the model. The degrees of freedom in the model are thus 48-19=29.

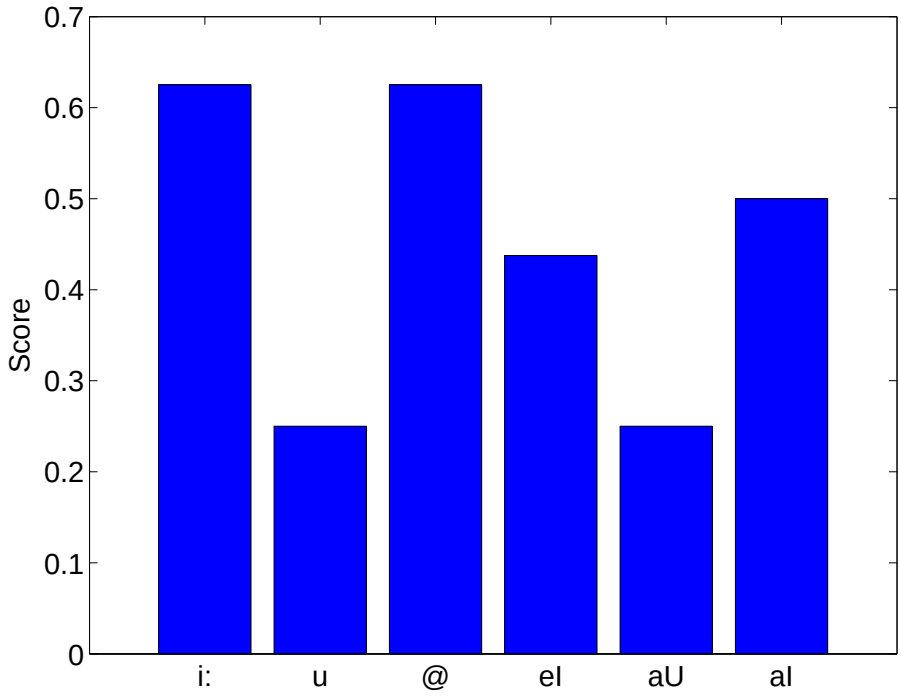


Figure 4.1: Mean CMOS score for each vowel.

Table 4.1 shows the goodness of the model fit ($R^2$ value), the variance of the PCA-based score, and the Root Mean Square Deviation (RMSD) between the observed ratings and the ratings predicted by the model. All models achieved good $R^2$ values. Vowels such as /i:, u, and @/ have larger correlations overall than diphthongs. However, the diphthongs have smaller variances and RMSD. We conclude that these distances indeed form a reliable predictor of perceptual speech quality, and thus can be used as a cost function for optimization of cross-fading.

| Vowel | $R^2$ | Variance | RMSD |
|---|---|---|---|
| /i:/ | .76 | 2.04 | .70 |
| /u/ | .62 | 1.06 | .63 |
| /@/ | .78 | 2.28 | .70 |
| /eI/ | .43 | 1.11 | .79 |
| /aU/ | .47 | .66 | .59 |
| /aI/ | .64 | .85 | .55 |

Table 4.1: Multiple linear regression with both linear and delta distances. Number of samples per vowel: 48, Degrees of freedom: 29, "*" significant with alpha = .05

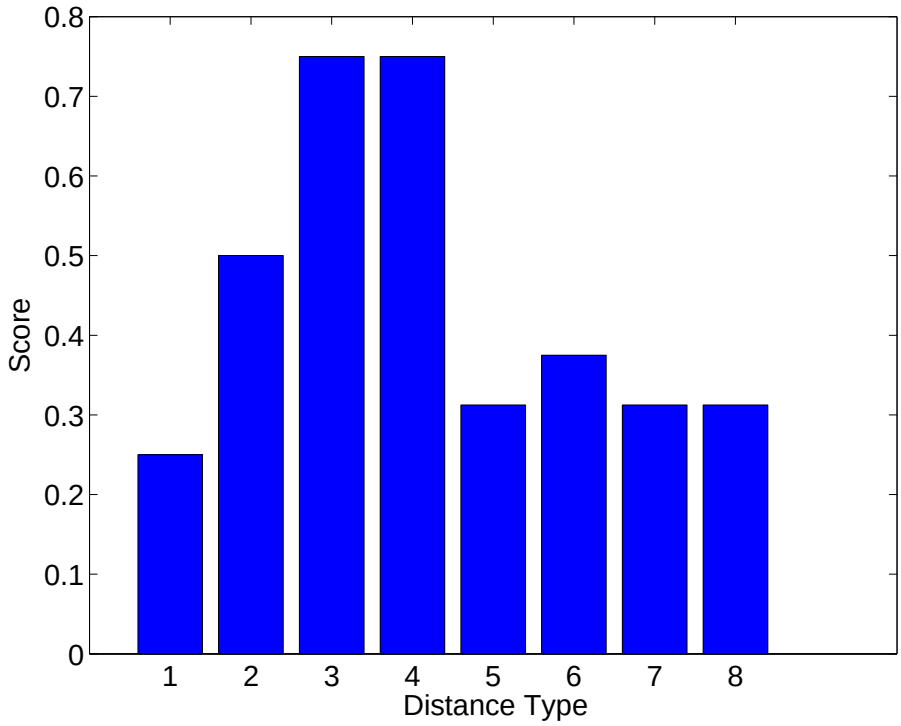


Figure 4.2: Mean CMOS score for each distance type.

To further understand how much the delta distances contribute to the perceived quality, we fit the linear model with linear distance and delta distance separately. Table 4.2 and Table 4.3 show that both linear and delta distances contribute to the perceived quality. To show the importance of the first derivative distances (delta distances) in the perceptual cost function, we calculated an $F$ score based on the correlation coefficients in Table 4.1 and Table 4.3. $F$ is calculated by the following equation:

$$F = \frac{29 \times (R^2_{all} - R^2_{linear})}{9 \times (1 - R^2_{all})} \tag{4.1}$$

We find that overall the perceptual quality is predictable from the distance measure we choose. The perceptual quality is predicted better for vowels than diphthongs, as shown in Table 4.4. The delta distances certainly help but only the vowel /@/ shows a significant improvement.

| Vowel | $R^2$ | Variance | RMSD |
|---|---|---|---|
| /i:/ | .64* | 2.04 | .84 |
| /u/ | .33 | 1.06 | .83 |
| /@/ | .64* | 2.28 | .90 |
| /eI/ | .25 | 1.11 | .90 |
| /aU/ | .28 | .66 | .68 |
| /aI/ | .31 | .85 | .76 |

Table 4.2: Multiple linear regression with delta distances only. Number of samples per vowel: 48, Degrees of freedom: 29, "*" significant with alpha = .05

| Vowel | $R^2$ | Variance | RMSD |
|---|---|---|---|
| /i:/ | .66* | 2.04 | 1.34 |
| /u/ | .49* | 1.06 | .73 |
| /@/ | .38* | 2.28 | 1.17 |
| /eI/ | .16 | 1.11 | .95 |
| /aU/ | .36* | .66 | .64 |
| /aI/ | .40* | .85 | .71 |

Table 4.3: Multiple linear regression with linear distances only. Number of samples per vowel: 48, degrees of freedom: 38, "*" significant with alpha = .05

| Vowel | $R^2$(linear) | $R^2(all)$ | $F$ |
|---|---|---|---|
| /i:/ | .66 | .76 | 1.34 |
| /u/ | .49 | .62 | 1.10 |
| /@/ | .38 | .78 | 5.86* |
| /eI/ | .16 | .43 | 1.53 |
| /aU/ | .36 | .47 | 0.67 |
| /aI/ | .40 | .64 | 2.15 |

Table 4.4: The contribution of delta distances to the perceived quality. "*" significant with alpha = .05

# Chapter 5

# Discussion and Future Work

We noted earlier that a linear weighted cross-fading can produce smooth yet unnaturally shaped formant trajectories; in addition, we noted that the precise details of how to cross-fade a specific pair of units may be highly context-dependent. We thus proposed to use trainable parameterized cross-fading, in which these details are provided by context-sensitive parameters. For this reason, a perceptually-validated cost function is necessary.

This paper reports a study on the feasibility of developing such a perceptual cost function. Toward this end, a special corpus was designed to produce a variety of shapes of formant frequency trajectories in different linguistic environments. A perceptual experiment was performed to determine if we could predict perceptual quality of output speech from acoustic distance measures. We generated a range of synthetic/natural stimulus pairs, where the synthetic stimuli were generated using three types of cross-fading models, applied to different regions in the vowel. We made sure that the synthetic stimuli covered a wide range of acoustic constellations, as measured by distances in the frequency and delta-frequency domains between the units in the first, second, and third thirds of the vowel region. We then applied these same six distance measures to compare synthetic (i.e., cross-faded) and natural (i.e., target) trajectories. A multiple linear regression model was trained for each vowel based on the perceptual score and these distance measures. The results show that the perceptual cost function can be reliably predicted from the distance measures. Moreover, the results support our hypotheses that: a) the quality of the output speech is influenced by the shape of formant trajectories in the entire region across the vowel; and b) human perceptual scores are correlated to both absolute distance and the first derivative of absolute distance of formant trajectories.

Concatenative speech synthesis is still the most widely-used TTS system in current speech technology research. We believe it will have many potential applications in the future. The biggest obstacle to generating natural sounding synthetic speech remains to minimize the overall unit selection cost without sacrificing either concatenation cost or target cost. A lot of work has been done to reduce prosodic and speech spectral discontinuities during concatenation. The study presented in this dissertation is just the beginning of a series of studies on developing an unit-dependent trainable parameterized cross-fading weight function to generate more natural-looking speech spectral feature trajectories and thus better-sounding output speech. The next steps for this study are to: 1) define parameterized families of cross-fading functions, i.e., use different types of cross-fading function for different concatenation units; 2) train the mapping between unit pair features and the parameters in the cross-fading function by minimizing a distance measure that compares the natural trajectories with the cross-faded trajectories; and 3) apply the trained cross-fading function during concatenation with proper speech signal processing methods.

Future work includes getting data from more speakers, including more phoneme classes, particularly consonants, in the study and training the cost function on other spectral distance measures as suggested in earlier work about perceptual prediction models based on the spectral distances [14, 5]. With an optimal distance measure for certain spectral feature, we can train perceptual cost functions for more phonetic classes and train the optimal cross-fading functions for each phoneme classes using these perceptual cost functions.

# Bibliography


[1] Chappell, D. T., and Hansen, J. H. L. A comparison of spectral smoothing methods for segment concatenation based speech synthesis. *Speech Communication 36*, 3 (2002), 343–373.

[2] Hunt, A., and Black, A. Unit selection in a concatenative speech synthesis system using large speech data. In *IEEE Int. Conf. Acoust., Speech, Signal Processing* (1996), pp. 373–376.

[3] Kain, A., Miao, Q., and van Santen, J. Spectral control in concatenative speech synthesis. In *6thISCA Workshop on Speech Synthesis* (Bonn, Germany, 2007).

[4] Kawai, H., and Tsuzaki, M. A study on time-dependent voice quality variation in a large-scale speech corpus for speech synthesis. In *IEEE Workshop on Speech Synthesis* (2002), pp. 15–18.

[5] Klabbers, E., and Veldhuis, R. Reducing audible spectral discontinuities. *IEEE Trans. on Speech and Audio Proc. SAP-09*, 1 (Jan. 2001), 39–51.

[6] Klatt, D. Review of text-to-speech conversion for English. *JASA 82*, 3 (Sept. 1987), 737–793.

[7] Low, P. H., Ho, C. H., and Yaseghi, S. Using estimated formant tracks for formant smoothing in text to speech synthesis. In *ASRU* (2003), pp. 688–693.

[8] Manell, R. H. Formant diphone parameter extraction utilising a labelled single-speaker database. In *ICSLP* (Sydney, Australia, 1998).

[9] Miao, Q., Niu, X., Klabbers, E., and van Santen, J. Effects of prosodic factors on spectral balance: analysis and synthesis. In *Speech prosody* (Dresden, Germany, 2006).

[10] Mizuno, H., Abe, M., and Hirowaka, T. Waveform-based speech synthesis approach with a formant frequency modification. In *ICASSP* (1993), pp. 195–198.

[11] Olive, J., and Spickenagel, N. Speech resynthesis from phoneme related parameters. *Journal of Acoustical Society of America 59*, 12 (1976), 993–996.

[12] Pearson, S., F., H., and Hata, K. Combining concatenation and formant synthesis for improved intelligibility and naturalness in text-to-speech systems. *International Journals of Speech Technology 1* (1997), 103–107.

[13] Shadle, C., and Damper, R. Prospects for articulatory synthesis: A position paper. In *Proc. of the fourth ISCA Tutorial and Research Workshop* (Perthshire, Scotland, 2001).

[14] Syrdal, A. K., and Conkie, A. D. Data-driven perceptually based joint costs. In *5th ICSA Workshop on Speech Synthesis* (2004), pp. 49–54.

[15] van Santen, J. P. H. . Assignment of segmental durations in text-to-speech synthesis. *Computer Speech and Language 8* (1994), 95–128.

[16] van Santen, J. P. H. ., Kain, A., Klabbers, E., and Mishra, T. Synthesis of prosody using multi-level unit sequences. *Speech Communication 46*, 3-4 (2005), 365–375.

[17] Wouters, J., and Macon, M. Effects of prosodic factors on spectral dynamics. ii. synthesis. *JASA 111*, 1 (2002), 428–438.

[18] Wouters, J., and Macon, M. W. Unit fusion for concatenative speech synthesis. In *ICSLP* (2000).

[19] Zen, H., and T., T. An overview of nitech hmm-based speech synthesis system for blizzard challenge 2005. In *Interspeech* (2005), pp. 93–96.