Perceptual Cost Function for Cross-fading Based Concatenation

Qi Miao, Alexander B. Kain, Jan P. H. van Santen

Center for Spoken Language Understanding (CSLU),
Division of Biomedical Computer Science (BMCS),
Oregon Health & Science University (OHSU), Oregon, USA 97006
{miaoqi, kain, vansanten}@cslu.ogi.edu

Abstract

In earlier research, we applied a linear weighted cross-fading function to ensure smooth concatenation. However, this can cause unnaturally shaped spectral trajectories. We propose context-sensitive cross-fading. To train this system, a perceptually validated cost function is needed, which is the focus of this paper. A corpus was designed to generate a variety of formant trajectory shapes. A perceptual experiment was performed and a multiple linear regression model was applied to predict perceptual quality ratings from various distances between cross-faded and natural trajectories. Results show that perceptual quality could be predicted well from the proposed distance measures.

Index Terms: perceptual score, formant frequency, cross-fading function, concatenation errors

1. Introduction

The most common Text-To-Speech (TTS) approach to-date is *concatenative synthesis*. In a variant, unit-selection system, the system performs a search in a pre-recorded speech database, which finds the best matched sequence of units for target speech by optimizing a two-part cost function: (1) target cost and (2) concatenation cost. The selected units are concatenated together to generate the final speech. 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. To reduce these concatenation errors, researchers (1) increase the size of the speech corpus to cover all possible combinations of the target unit sequences [1] or (2) apply additional modeling to modify prosody and speech spectrum. The first approach is usually time consuming and expensive in most cases. Furthermore, the voice quality of the speaker changes over time [2]. It's also not very practical in cases where personalized TTS is required. The concatenation errors occur both in speech prosody and the spectral domain. To eliminate the errors in prosody, building global pitch and duration models is a common method [3, 4]. To reduce spectral discontinuities, researchers have done studies on smoothing spectral balance discontinuities at concatenation points, expressed as energies in four bands [5], smoothing formant discontinuities [6, 7, 8], and applying a fusion-unit approach during the concatenation [9].

All of these studies try to achieve the goal of generating high-quality synthetic speech with smooth transitions at unit concatenation points. However, a smooth transition does not guarantee that the speech sounds natural. In our previous study [10], a linear cross-fading weight function was used to remove spectral and time domain discontinuities during concatenative speech synthesis. In this method, smoothing is performed by cross-fading across a "region" of concatenation, instead of the traditional "points" of concatenation. 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 the reason might be 1) the corpus used in the experiment was recorded in a constant phonemic and prosodic context, thus the original formant distance in the corpus is 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 could generate unnaturally shaped formant frequency trajectories. At the same time, TDXF cannot implement the impacts 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 unit.

Figure 1 shows one example of applying a linear weighted cross-fading function to two concatenated units trajectory and their target unit trajectory. The blue and red line represents the formant trajectories of two units to be concatenated. These trajectories not only have large distances in the frequency domain, but also their shapes are sharply divergent. The black curve is the cross-faded trajectory and the green curve is the trajectory from the target unit. However, even though the cross-faded trajectory is perfectly smooth, the overall shape is quite different from the natural curve. The cross-faded trajectory would result in unnatural sounding speech output.

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.

In this paper, we focus on solving the problem in the first part of the algorithm.

2. Hypothesis

Previous studies [11, 12] used perceptual data to predict the relationship between concatenation cost and audible distortions.

This work was supported by NSF grant #0313383 "Objective Methods for Predicting and Optimizing Synthetic Speech Quality".

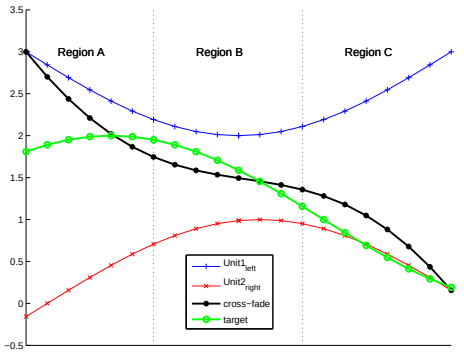

Figure 1: *Cross-fading between Two Trajectories.*

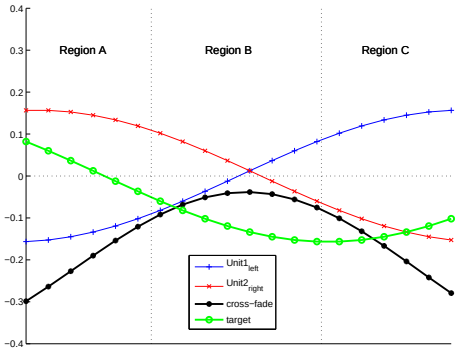

Figure 2: *First Derivative of the Cross-faded Trajectories.*

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 whole 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. 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 distance in region 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 coarticulated). Figure 1 and Figure 2 show an example of two concatenated trajectories, the cross-faded trajectory, and their first derivatives. The blue and red lines are trajectories from the concatenated units. The black lines are the cross-faded trajectory and its first derivatives. The green lines are the natural trajectories from the target unit. We can see that in region B, even though the absolute distance between the cross-faded and natural is small, the distance in the first derivatives is big. A large distance in formant trajectory shape may produce unnatural transitions during concatenation.

3. 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 1)[1] which covers 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.

Table 1: *Vowels in the Corpus*

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

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 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.

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 formants were corrected and used in the perceptual experiment.

4. Perceptual Experiment

4.1. Stimuli

We performed a search procedure to select the unit pairs which were used to synthesize the CVC words in the experiment. We calculated the distances between two candidate units in six regions. For each region, we calculated the maximum absolute distance and maximum first derivative distances between two candidate units. Then we performed z-transformations across the units in each region.

Since the second formant trajectory usually has the strongest dynamics for the vowels, we applied the following criteria on the z-transformed 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,
- large distances in all distance measures, or
- a relatively large distance in one distance measure

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

Table 2: *Definition of Distance Measures for Each Formant Trajectory.*

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

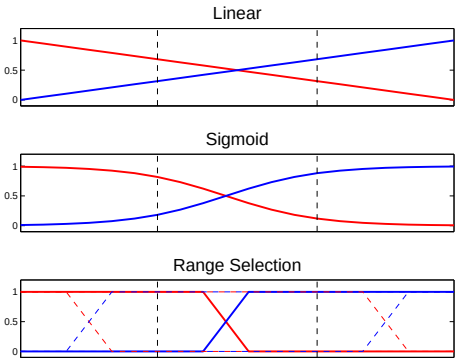

Figure 3: *Three Types of Cross-fading Weighted Function.*

For each vowel, two target units were selected. For each target unit, we used eight types of concatenations. For each pair of candidate units, three types of cross-fading models were applied. The first model is the linear weighted function as used in our previous study [10]. The second model is the sigmoid function. The third model is a *range selection function* where we put the cross-fading area always in the region where the largest discontinuity between two concatenated units occurs. Figure 3 shows the three cross-fading models. The red line shows the weight function that applies to the left unit, the blue line represents the weight function to the right unit. In combination, the total number of stimuli is 6*2*8*3 = 288.

In order to eliminate effects from other features, such as pitch, duration, and energy, we resynthesized 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.

4.2. Procedure

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" 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.

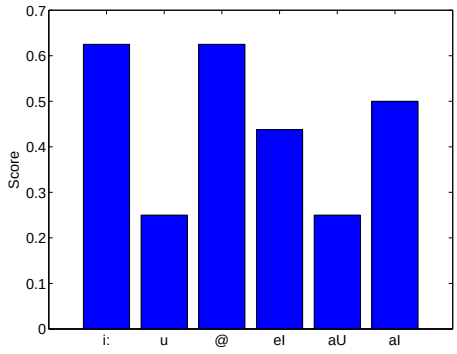

Figure 4: *Mean CMOS score for each vowel.*

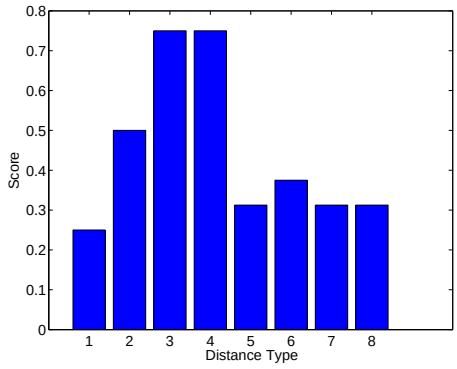

Figure 5: *Mean CMOS score for each distance type.*

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.

4.3. Results

Figure 4 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.

Figure 5 shows the mean CMOS scores for each distance type, as defined in Table 3. On average, larger distance in any distance measure is 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 subject, we first transformed the scores into z-transformed scores, then into a weighted final score by applying principal component analysis (PCA) [5]. This analysis eliminates the effects of different individuals using larger rating ranges, and also assigns larger weights to subjects more in agreement with other sub-

Table 3: *Definition of Distance Types for Selected Units*

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

jects. 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. Table 4 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.

Table 4: *Multiple Linear Regression Model for Each Vowel.*

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

5. Conclusions

We noted earlier that 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, a perceptually-validated cost function is necessary. This paper reports a study on the feasibility of developing such perceptual cost functions. 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 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.

Future work includes training the cost function based on other spectral distance measures as suggested in earlier work about perceptual prediction model based on the spectral distances [11, 12]. Next, we plan to train the function for more phonetic classes. Finally, we will train the optimal cross-fading models for each phoneme classes using these perceptual cost functions.

6. Acknowledgements

The authors would like to thank John-Paul Hosom and Esther Klabbers-Judd for the insightful discussions and all the subjects for participating the perceptual test.

7. References

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

[2] H. Kawai and M. Tsuzaki, "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.

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

[4] J. P. H. . van Santen, A. Kain, E. Klabbers, and T. Mishra, "Synthesis of prosody using multi-level unit sequences," *Speech Communication*, vol. 46, no. 3-4, pp. 365–375, 2005.

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

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

[7] D. T. Chappell and J. H. L. Hansen, "A comparison of spectral smoothing methods for segment concatenation based speech synthesis," *Speech Communication*, vol. 36, no. 3, pp. 343–373, 2002.

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

[9] J. Wouters and M. Macon, "Unit fusion for concatenative speech synthesis," in *International Conference on Spoken Language Processing (ICSLP)*, Oct. 2000, pp. 83–86.

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

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

[12] E. Klabbers and R. Veldhuis, "Reducing audible spectral discontinuities," *IEEE Trans. on Speech and Audio Proc.*, vol. SAP-09, no. 1, pp. 39–51, Jan. 2001.