Effects of Prosodic Factors on Spectral Balance: Analysis and Synthesis

Qi Miao, Xiaochuan Niu, Esther Klabbers and Jan van Santen

Center for Spoken Language Understanding
OGI School of Science & Engineering
Oregon Health & Science University
20000 NW Walker Road, Beaverton, OR 97006, USA
{miaoqi, xiaochua, klabbers, vansanten}@cslu.ogi.edu

Abstract

In natural speech, prosodic factors such as accent, stress, phrasal position and speaking style play important roles in controlling several acoustic features, including segmental duration, pitch, and *spectral balance*, i.e., the amplitude pattern across different frequency ranges of the power spectrum. To synthesize speech that sounds natural, these effects need to be accurately modeled. In this study we describe and evaluate a synthesis method that mimics the effects of prosodic factors on spectral balance. We measure spectral balance by using the energy in four broad frequency bands that correspond to formant frequency ranges. An additive model is used to capture the effects of prosodic factors on spectral balance. A new sinusoidal synthesis module is implemented under Festival to predict the target spectral balance value for each band from analysis results and apply it to the amplitude parameters of the sinusoidal model during synthesis. In this study we evaluate an important strength of this system, which is its ability to reduce spectral discontinuities in unit concatenation.

1. Introduction

It is well-known that prosodic factors such as word stress, phrase accent, phrasal position, and speaking style, have systematic effects on the acoustic features of prosody. It has also become clear that these features are not confined to segmental duration, pitch, and loudness, but also include aspects of the spectral structure, such as spectral tilt [1, 2, 3], spectral dynamics [4], and spectral balance [1, 8].

In order to capture the effects of prosodic factors on acoustic features of prosody, one can either record a large corpus with speech units in all possible prosodic contexts, or develop computational methods to modify these features according to the target speech prosody. The second approach requires a relatively small corpus and the key procedure is to design good models to predict and modify the features. These numerical models provide a more efficient way to modulate acoustic features and spectral features based on prosodic factors. It also increases the flexibility of TTS systems to generate target features close to that of natural speech.

Speech can be synthesized using either concatenative or rule-based synthesis. Although concatenative synthesis is currently the most widely-used method, the problem with it is that two units that are concatenated are generally recorded in different phonemic and prosodic contexts, which can lead to audible discontinuities at the concatenation points. Many studies have been focused on predicting and eliminating the occurrences of these discontinuities [5, 6].

In this study, we extend our previous research on the analysis of effects of prosodic factors on the spectral balance of vowels [8] to include all the phonemes in American English. We implement a new sinusoidal synthesis module under Festival [7]. This module predicts the target spectral balance contours for each band from the analysis results and applies it to the amplitude parameters of the sinusoidal model during synthesis. The specific goals of this study are the following:

- To mimic the effects of prosodic factors on the spectral balance of speech by using a statistical model.
- To eliminate or reduce the discontinuities in spectral balance during synthesis so that the output speech has a smoothed spectral balance trajectory in each formant frequency band.
- To perform a perceptual experiment which shows that the modification of spectral balance during concatenation can reduce audible discontinuities caused by spectral balance mismatch, even when formant frequency mismatch exists.

This paper is organized as follows. Section 2 introduces the spectral balance analysis phase. Section 3 explains the procedure of applying predicted spectral balance values during sinusoidal synthesis. The perceptual experiment and results are demonstrated in Section 4. Final conclusions and future work are presented in Section 5.

2. Spectral Balance Analysis

2.1. Measurement

Spectral balance is measured in four broad frequency bands. We define the four bands according to the four formant frequencies. These four bands are generally phoneme independent, and contain the first, second, third and fourth formant for most of the phonemes. Formants contain the largest portion of energy in the frequency domain. Moreover, when some prosodic factors change, e.g., from unstressed to stressed, the energy near formants will change much more than those near other frequency locations. Choosing frequency bands according to formant frequencies has an important advantage for statistical analysis, because it will reduce interactions between phoneme identity and prosodic factors. For speech with $16k$Hz sampling rate, the

This research was conducted with support from NSF grants 0082718, "Modeling Degree of Articulation for Speech Synthesis", 0313383,"Objective Methods for Predicting and Optimizing Synthetic Speech Quality", and 0205731, "Prosody Generation for Child Oriented Speech Synthesis".

four bands are defined as: B1:100-800Hz, B2: 800-2500Hz, B3: 2500-3500Hz, B4: 3500-8000Hz.

We compute the four-band energy by performing pitch synchronous sinusoidal analysis in each pitch period. We calculate the sum of the amplitudes of sinusoidal parameters in each band, and then transform the energy into the Log domain as the four spectral balance values for this period. When there is no pitch available, a fake pitch value of 100 Hz is assigned. For each phoneme unit we calculate the spectral balance values at five different locations inside this unit, at positions 10%, 25%, 50%, 75% and 90% in the unit. This is sufficient to capture the trend of spectral balance change inside a unit.

We model spectral balance variation using an additive version of the Sums-of-Products model [9]. Specifically, Spectral balance (SB) is calculated as the sum of effects of several prosodic factors in the Log domain as

$$SB_j(C_0, C_1, \cdots, C_n) = \sum_{i \in K} S_{i,j}(C_i) \quad (1)$$

Where $C_0, C_1, \cdots, C_n$ represent the levels of prosodic factors, K is the set of factors, $S_{i,j}(C_i)$ represents the effect of factor i being at level C_i on band j.

2.2. Analysis

The spectral balance prediction model is trained on the same corpus as we used in a previous study [8]. This corpus contains 472 sentences spoken by a non-professional native American English speaker. These sentences are selected using a greedy algorithm, which covers the different combinations of prosodic factors. This corpus was segmented using CSLU's forced alignment system and was manually adjusted. Prosodic factor levels were labeled by hand. For instance, the factor stress was labeled as unstressed or stressed; the factor accent was labeled with a 0-3 scale of strength; the relative position to boundaries was labeled as the distance to the previous and next minor (i.e., separated by comma) or major (i.e., separated by utterance) phrase boundaries. Because of the interaction effects between stress and accent, these two factors are combined into one factor ("StressAccent").

Following [9], we created a *tree* as follows. The top branching consisted of "Nucleus" for vowels and diphthongs vs. consonants; next, consonants branched into "Onset"; "Coda" and "Intervocalic". These are further split into consonant classes (e.g., voiceless stops). For prediction of spectral balance, in each category an additive model is trained for four bands at five different analysis positions inside the phoneme. Note that in our previous study we only analyzed the spectral balance at the mid-point and confined the analysis to vowels [8]. In this study in order to apply the analysis results during synthesis, four additional points are added during analysis to capture the movement of spectral balance within a phoneme segment. The spectral balance curve for a phoneme can then be obtained by interpolating between these five points.

Due to the space limitations, we can only briefly review here the key analysis results. A typical result is shown in Figure 1, which shows that in accented words the effects of word stress are primarily on the higher frequency bands. Other results include effects of phrase boundaries and phonemic context. Broadly speaking, the results confirm the analysis results from other studies [1, 2, 3].

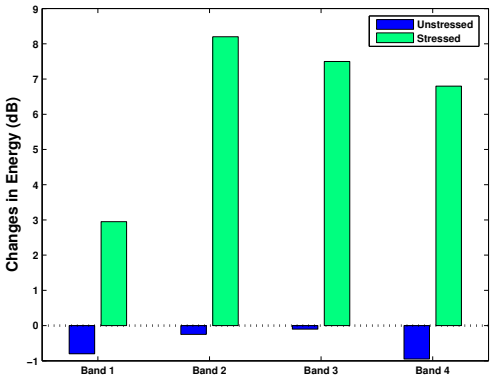

Figure 1: *Effects of word stress for accented words.*

3. Spectral Balance Synthesis

We have implemented the spectral balance predictor as a new module under Festival with the OGI plug-in modules [7]. This module takes the prosodic factor levels computed by the Festival text analysis system and predicts the spectral balance values for each band at each time point. The prosodic factors are the same as the factors used in the analysis.

The synthesis step uses a diphone concatenation method. We used CSLU's "as_diphone" corpus for synthesis which was designed to cover all diphones (3328 diphones), in a (constant) prosodic context defined by a carrier phrase ("Say CV-te again" for CV diphones, e.g., "Say BEEte again"; or "Say at-VC again" for VC diphones). During synthesis, the diphone synthesizer selects diphone units from the corpus and concatenates them together to generate the target utterance. The concatenation units are stored as sets of parameters, which, for sinusoidal synthesis, include pitch marks, the amplitude and the phase of sinusoidal parameters.

The diphone corpus came from a different speaker other than the analysis corpus. To apply our analysis results, we made two assumptions:

- The effects of phoneme identity are speaker dependent.
- The effects of stress, accent and relative position are speaker independent.

Therefore, spectral balance value for band j is predicted as

$$\begin{aligned} SB_j(Phoneme, StressAccent, Position) = \\ S_j(Phoneme)_{Synthesis} \\ + S_j(StressAccent)_{Analysis} \\ + S_j(Position)_{Analysis} \end{aligned} \quad (2)$$

The effects of phoneme identity are simply computed as the mean spectral balance values for each phoneme category in the diphone corpus. In the future, when a new corpus with a different speaker is installed, these mean values can simply be replaced by those for the new corpus.

A new sinusoidal synthesis module based on the OGIsinLPC synthesis module has been implemented. This module takes the predicted spectral balance values, interpolates between the five different analysis points for each band, generates a smoothed spectral balance trajectory within the segment and applies it to our amplitude parameters of the sinusoidal synthesis method. The spectral balance modification is an optional

feature of the OGIsinLPC module. The modification of spectral balance is only performed in sonorants.

4. Perceptual Experiment

The goal of the proposed spectral balance analysis approach is two-fold: to better mimic the effects of prosodic factors on spectral balance, and reduce spectral balance discontinuities during concatenation. In this study, we address the second of these.

4.1. Material

We used a corpus consisting of four nasal-vowel-nasal CVC's (*moon*, *mean*, *moan*, and *main*) occurring in 18 different contexts as defined in terms of "foot structure" [13]. Specifically, the CVC could be stressed or unstressed, it could occur as the first/medial/final syllable in the foot, and the foot could contain varying numbers of syllables. The speech was recorded by a female American English speaker. For this experiment, each target word was excised from the original speech waves and stored in a separate corpus. In this corpus, there are a total of 18*4=72 words. For the words "moon" and "mean", the unit concatenation point is put in the middle of vowel. For the words "moan" and "main", the unit concatenation point was labeled as the steady formant transition point for the diphthong. For each word, we performed pitch synchronous sinusoidal analysis and calculated spectral balance values. For each unit pair, both the formant frequency distance and the spectral balance distance are calculated at the concatenation point of the two units. Therefore, for each word, there are 18*17=306 possible concatenations.

Since it is plausible that differences in spectral balance between units are correlated with differences in formant frequency values, we performed a search procedure to select the pairs which were used to synthesize the words to be used in the experiment. We divided stimuli in two sub-groups: one with relatively high formant frequency distance (Hi_F) and the other with relatively lower formant frequency distance (Lo_F). We used the following criteria in the search:

- The selected unit pairs have a large spectral balance distance, to ensure that our method would have an effect.
- The Hi_F and Lo_F sub-groups should have a large difference in formant frequency distance.
- The two sub-groups of stimuli have a small difference in spectral balance distance.

They selection algorithm works as follows. Consider two unit pairs A and B, with formant frequency distances $D(F, A)$ and $D(F, B)$ and spectral balance distances $D(S, A)$ and $D(S, B)$, respectively. In the search algorithm, the (A, B) pair is rated by the following score:

$$Score(A, B) = \frac{(D(S, A) + D(S, B))(D(F, A) - D(F, B))}{|D(S, A) - D(S, B)| + 1} \quad (3)$$

We calculate the score for each unit pair and sort them based on their scores. 24 unit pairs with highest scores were selected for each word. Figure 2 shows the selected unit pairs for the word moon. The horizontal axis is the Euclidean spectral balance distance and the vertical axis is the Euclidean formant frequency distance. Each dot in the figure represents a unit pair and dots with circles are the selected samples. The two distances are calculated as

$$D_{FF}(m-V, V-n) = \sqrt{\sum_{k=1}^{3}(FF_{k,m-V} - FF_{k,V-n})^2}. \quad (4)$$

and

$$D_{SB}(m-V, V-n) = \sqrt{\sum_{k=1}^{4}(SB_{k,m-V} - SB_{k,V-n})^2}. \quad (5)$$

where $m-V$ indicates the left unit and $V-n$ indicates the right one, $FF_{k,m-V}$ means the kth formant frequency (in Bark) at the concatenation point in the left unit, $SB_{k,m-V}$ means the kth spectral balance value at the concatenation point in the left unit.

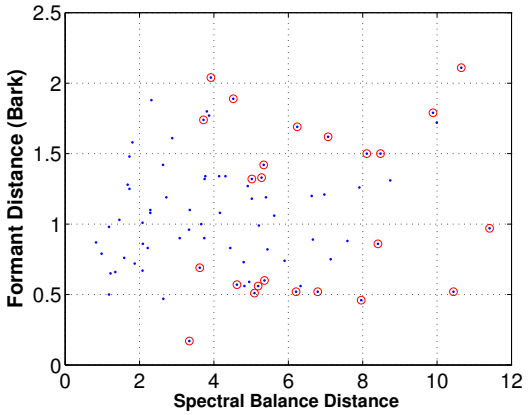

Figure 2: *Stimulus selection process.*

The selection algorithm succeeded in finding unit pairs that were well-separated in terms of formant frequency distances but quite close in terms of spectral balance distances. For the four words, the formant distances (in Bark) for the Hi_F vs. Lo_F were 1.42 vs. 0.40, 1.92 vs. 0.54, 1.32 vs. 0.48, and 1.66 vs. 0.58; for the spectral balance distances, the corresponding values were 6.23 vs. 6.27, 6.34 vs. 6.19, 5.63 vs. 5.64, and 6.52 vs. 6.54.

The selected unit pairs were synthesized by the OGIsinLPC module under Festival both with and without spectral balance modification in the vowel part. In order to reduce the influence of other prosodic factors, default duration and pitch values were used during synthesis.

4.2. Procedure

The experiment was performed in the same environment as in the previous research by Klabbers et al. [11]. The experiment was set up as a Comparative Mean Opinion Score (CMOS) test. Six expert subjects were asked to listen to the pairs of words and rate the quality of the word "A" compared to word "B" on a five-point scale. The subject had to choose one answer from: (-2) A sounds better, (-1) A sounds slightly better, (0) About the same, (1) B sounds slightly better, (2) B sounds better. Word A and B are the same word concatenated using the same unit pairs. One is concatenated without spectral balance modification and the other with modification. Meanwhile, the order of A and B is randomized. The experiment was performed in the

CSLU Perception Lab with professional audio devices. During the experiment, subjects could repeat the stimuli for many times until a selection was made. The total time for one test was about 20 minutes.

4.3. Results

We followed the same analysis method as described in Klabbers et al. [11]. The scores from subjects were first transformed so that a higher score represents a better quality after the spectral balance modification. If the score is larger than zero, the corresponding subject concludes that the stimuli with spectral balance modification has a better quality. Six out of six subjects had mean scores greater than zero for the Hi_F and Lo_F stimuli, which is significant at $p < 0.016$ using an exact sign test. Thus, the modification of the spectral balance significantly reduces the discontinuities during concatenation regardless of formant discrepancies.

The score for Lo_F stimuli was slightly higher (0.806) than for Hi_F stimuli (0.694), suggesting that the impact of reducing spectral balance discontinuities is larger when formant frequency discontinuities are smaller. However, this differential was not statistically significant.

To further understand the impact of spectral and formant frequency distances on the ratings, the following analysis was performed. For each stimulus a weighted final score was computed by using principal component analysis (PCA) applied to the per-listener z-transformed scores [12]. 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. A multiple linear regression between the PCA-based scores and two types of distances showed that spectral balance distance and formant frequency distance contribute differently to subjects' perceptions, with significant effects of the former ($t_{93} = 3.99$, $p < 0.001$) and the effects of formant frequency distances not being significant ($t_{93} = 1.25$, $n.s.$). We note that, of course, the concatenated units came from phonemically identical contexts. In a different study, Klabbers [11] found that in a diphone corpus, where units are obtained from different phonemic contexts, formant frequency distance is an important contributor to audible discontinuity.

5. Conclusion and Future Work

This paper presents a study on analysis and synthesis of effects of prosodic factors on spectral balance of speech. An additive model is used to model the effects of several factors (vowel identity, stress, accent and relative position) on the spectral balance of speech. By performing pitch synchronous sinusoidal analysis, the spectral balance is measured by the energy in four formant frequency bands in which energy is computed as the sum of amplitude parameters of sinusoidal model. In each phonetic category, we trained additive models for four bands separately.

A new sinusoidal synthesis module is implemented under Festival. This module predicts the target spectral balance value for each band from analysis results and interpolates between different analysis point to generate a smoothed spectral balance trajectory, and applies it to the amplitude parameters of the sinusoidal models.

To test if our method can reduce the audible discontinuities in spectral balance during units concatenation, we selected 96 pairs of diphone units ("m-V" and "V-n") and synthesized these words both with and without spectral balance modification for the vowel part. Six subjects were asked to listen to the stimuli and rate the relative quality of two versions. All the subjects agreed that after the spectral balance modification, the discontinuities during unit concatenation are reduced. In our experiment set, the stimuli with lower formant frequency distance had a higher score than the sub-group with higher formant frequency although the paired students' t-test showed there is no significant difference in mean between two sub-groups.

This paper reported on a subproject of a larger project that has as goal to control all aspects of the speech signal that are related to prosody, including not only pitch and duration, but also "degree of articulation" and spectral balance. The solution that we proposed to address spectral balance had two sub-goals, namely mimicking spectral balance dynamics in natural speech and elimination of spectral balance discontinuities. This paper showed that we succeeded in achieving the second of these goals. The first goal will be addressed in work currently underway.

6. References

[1] Agaath Sluijter. Phonetic correlates of stress and accent. Holland Institute of Generative Linguistics, 1995.

[2] G. Fant, A. Kruckenberg, S. Hertegard, and J. Liljencrants. Accentuation and subglottal pressure in Swedish. In *Proceedings of ESCA Tutorial and Research Workshop on Intonation: Theory, Models and Applications*, Athens, 1997.

[3] N. Campbell and M. Beckman. Stress, prominence and spectral tilt. In *Proceedings of ESCA Tutorial and Research Workshop on Intonation: Theory, Models and Applications*, Athens, 1997.

[4] J. Wouters and M. Macon. Effects of prosodic factors on spectral dynamics. I. Analysis. *Journal of the Acoustical Society of America*, vol. 111, No. 1, pp: 417-427, 2002.

[5] J. Wouters and M. Macon. Unit fusion for concatenative speech synthesis In *Proceedings ICSLP*, Beijing, China, 2000

[6] Esther Klabbers and R. Veldhuis Reducing audible spectral discontinuities. IEEE Transactions on Speech and Audio Processing, Vol. 9(1), pp: 39-51.

[7] P. Taylor, A. Black and R. Caley The architecture of the Festival speech synthesis system in *Third ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998.

[8] Jan P. H. van Santen, X. Niu. "Prediction and Synthesis of Prosodic Effects on Spectral Balance of Vowels", *4th IEEE Workshop on Speech Synthesis, Santa Monica, CA*, 2002.

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

[10] Esther Klabbers and Jan P. H. van Santen. Control and prediction of the impact of pitch modification on synthetic speech quality. *Proceedings EUROSPEECH-03, Geneva, Switzerland*, pp. 317-320, 2003.

[11] Esther Klabbers, Jan P. H. van Santen and Alexander Kain. The contribution of various sources of spectral mismatch to audible discontinuities in a diphone database. (Draft) 2005

[12] Jan P. H. van Santen. Perceptual experiments for diagnostic testing of text-to-speech, systems. *Computer Speech and Language* vol, 7, pp. 49-100, 1993.

[13] Esther Klabbers, Jan van Santen and Johan Wouters. Prosodic factors for predicting local pitch shape, *IEEE 2002 Workshop on Speech Synthesis, Santa Monica, CA*, September 11-13 2002.