

Spectral Control in Concatenative Speech Synthesis

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

Center for Spoken Language Understanding (CSLU)
OGI School of Science & Engineering
Oregon Health & Science University (OHSU)
20000 NW Walker Road, Beaverton, OR 97006, USA

Abstract

We report on research in which we increased the degree of spectral control in concatenative synthesis by controlling the formant frequencies of the synthetic speech, as well as the energies in four spectral bands. In addition, we eliminated “points” of concatenation in favor of “regions” of concatenation, by *cross-fading* between the end and the beginning of two speech segments that are part of a concatenation operation. We hypothesized that these approaches would decrease the frequency and severity of audible discontinuities in the synthetic speech and thus also increase the perceived quality of the speech. A listening test determined that stimuli created with the proposed methods resulted in significantly increased quality.

1. Introduction

In the process of generating audible speech from a textual representation, a text-to-speech (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. These speech synthesis models vary in their complexity. The first intelligible synthesizers 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 rule [1] or from a database [2]. Most aspects of speech are controllable, including the degree of articulation and characteristics of the speaker. The resulting speech is highly intelligible, but is often judged as not very natural. In an effort to increase naturalness without decreasing flexibility, researchers have increased the complexity of the speech synthesis model to take into account more physiological and physical details about the speech production process; this approach is called *articulatory synthesis* [3]. Unfortunately, it is proving 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*.

The most successful TTS approach to-date is called *concatenative synthesis*; in this approach, natural speech utterances of a single speaker must first be recorded and stored in an *acoustic inventory*. During synthesis, individual portions of speech are retrieved from the inventory, optionally modified, and then concatenated in the desired sequence. In the *unit-selection*

approach, the relationship between linguistic units from text-processing and acoustic units of the acoustic inventory is established by means of a search, which, given a sequence of target linguistic units, optimizes (1) the fit between chosen linguistic units and the target linguistic units, also known as *target cost*, and (2) the fit between the chosen consecutive units, usually in the acoustic domain, known as *concatenation cost*. Intelligibility and naturalness are very high in the concatenative synthesis approach [4]. However, output speech is limited by the contents of the acoustic inventory (not just the linguistic content, but also the emotional state of the speaker, degree of articulation, etc.), and inevitable concatenation errors can lead to audible discontinuities. To overcome the problems of limited content and discontinuities, researchers either significantly increase the size of the database to include more variability, or introduce additional modeling to modify and thus control the natural speech signal. In the latter case, models that include prosodic control of pitch and duration are common [5]. In addition to prosodic modifications, researchers have also proposed spectral modifications, for example smoothing spectral balance discontinuities at concatenation points, expressed as energies in four bands [6], smoothing formant discontinuities [7, 8, 9], and controlling the degree of articulation [10].

It is our long-term goal to combine parametric and concatenative synthesis methods to achieve highly flexible and natural speech, by researching data-driven speech models and high-quality speech modification algorithms. In this paper, we report on experiments involving modification of formant frequencies, spectral balance, and time-domain waveforms, using speech units selected by the concatenative approach. We eliminated “points” of concatenation in favor 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. We hypothesized that this approach would decrease the frequency and severity of audible discontinuities in the synthetic speech and thus increase the perceived quality of the speech.

Section 2 introduces the methods to analyze and construct formant frequency trajectories, and to implement the necessary changes to the speech signal. Section 3 describes a perceptual test designed to validate our hypothesis, and we conclude in Section 4.

2. Methods

The key goal of our proposed approach is to decrease the negative effects of unnatural discontinuities between two speech segments that are part of a concatenation operation. We aimed to achieve this decrease by explicitly controlling the first three formant frequencies and the energies in four spectral bands

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

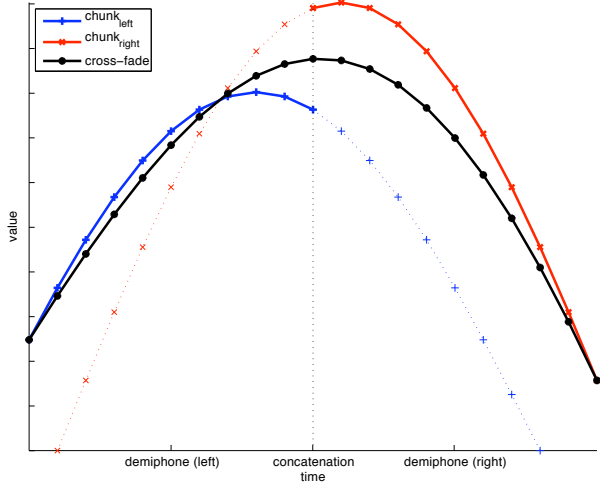


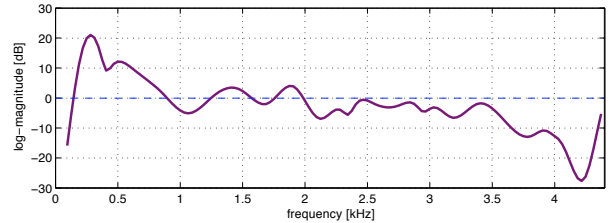
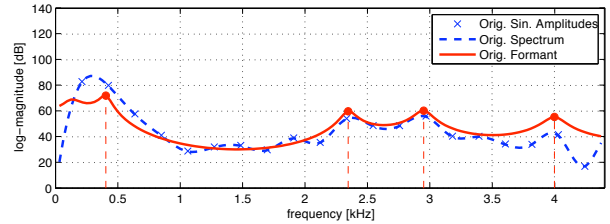
Figure 1: Cross-fading across a region of concatenation, using a fictitious 1-dimensional example feature. Without cross-fading, the final trajectory would be the concatenation of the solid left half-curve with the solid right half-curve, resulting in a large discontinuity. With cross-fading, the following 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.

of the synthetic speech. Although a predictive model of how these features, given linguistic targets, may evolve over time exists [11], we initially chose a cross-fading approach of natural formant frequencies and natural four-band spectral energies in the acoustic inventory to achieve smooth synthetic trajectories. We then modified the selected speech segments in accordance with the cross-faded feature trajectories. Even after controlling for formant frequencies and spectral balance, we expected remaining, unaccounted-for differences in the two speech segments to be joined. Therefore, a final time-domain cross-fade was employed, which cross-faded the waveforms of the two modified segments.

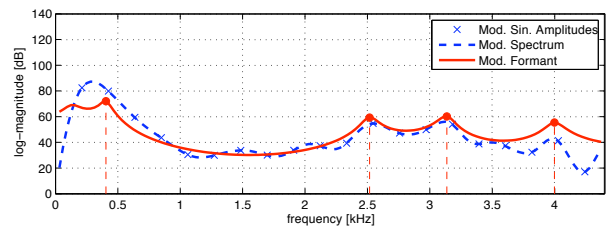
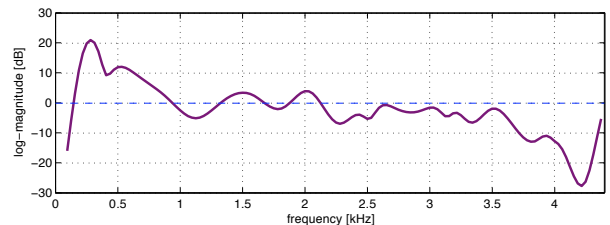
2.1. Acoustic Inventory and Feature Analysis

Our acoustic inventory consisted of the CSLU’s TLL diphone database, used in related previous experiments [12]. Typically, a diphone database only contains “chunks” (contiguous speech segments containing one or more speech units) of the type $a_R - b_L$, where a_R is the demiphone corresponding to the right-hand side of a phoneme a , and b_L is the demiphone corresponding to the left-hand side of the following phoneme b . To accommodate cross-fading in the formant frequency, spectral band, and time domains, we extended our analysis one demiphone to the left and one to the right, analyzing and storing chunks of the type $a_L - a_R - b_L - b_R$, equivalent to two full phonemes for each possible phoneme combination.

For each of the 1733 two-phoneme chunks in the acoustic inventory, we automatically extracted formant frequencies, as well as amplitudes and phases of harmonic sinusoids. We calculated energies in four discrete spectral bands (0–800 Hz, 800–2500 Hz, 2500–3500 Hz, and 3500–8000 Hz) by integrating the corresponding harmonic amplitudes [13, 6]. Formant frequency trajectories in vowel regions (the focus of the perceptual experiment in Section 3) were manually verified and corrected when



(a) Removing the frequency response of vocal tract and glottal source from the original speech signal. Top pane shows the original sinusoidal frequencies, the spectral envelope, and the model fit. Bottom pane shows the resulting residual.



(b) Creating a new spectrum. Top pane shows the warped residual. Bottom pane shows the frequency response of the new model, as well as the recombination of that model with the warped residual.

Figure 2: Formant frequency modification.

necessary, using a standard labeling tool in conjunction with a pen input device.

2.2. Feature Trajectory Construction

As mentioned previously, we aim to reduce concatenation errors by constructing smooth feature trajectories in the formant frequency and spectral balance domains, and then 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 tra-

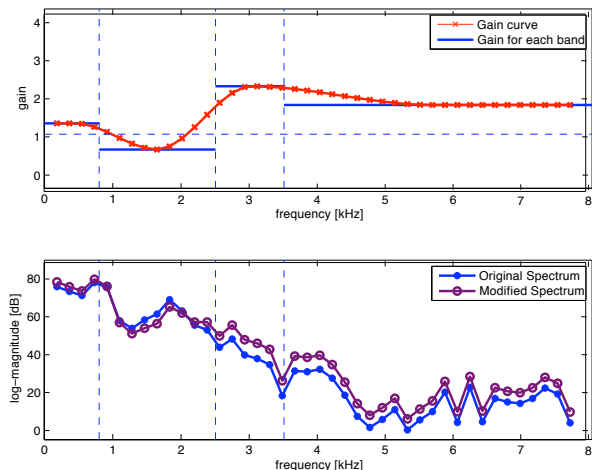


Figure 3: Spectral band modification. Top pane shows the desired amplitude gains for each individual band. To avoid discontinuities, a smooth gain curve is calculated. The bottom pane shows the original and modified sinusoidal harmonics and spectral envelopes.

jectories $\mathbf{s}(t)$ 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 $t = 1 \dots N$ of the last demiphone of the left chunk and the first demiphone of the right chunk, respectively, N denotes the total number of datapoints in the cross-fade region, and α is the cross-fade function given by $\alpha(t) = t/(N + 1)$.

Figure 1 illustrates the concept using a fictitious trajectory. We implemented both formant domain cross-fading on the first three formant frequencies, and spectral balance cross-fading, using the energies in four spectral bands.

The approach just described has some parallels with a “fusion unit” strategy researched previously [14]; however, the differences are that our proposed approach modifies formant frequencies instead of line spectral frequencies, does not require a fusion unit, and operates on features directly, instead of on their derivatives.

2.3. Speech Modification and Synthesis

Speech was synthesized using a pitch-synchronous, frame-by-frame, overlap-add, harmonic sinusoidal system. During synthesis, both left and right natural segments from the acoustic inventory are modified in accordance with the smooth feature trajectories constructed as described in the previous section, first in the formant frequency (FFXF) and then in the spectral band domains (SBXF). Finally the two modified speech segments are cross-faded in the time-domain (TDXF), to smooth any remaining acoustic differences.

2.3.1. Formant Modification

The modification of formants has attracted attention by many researchers. Most studies focus on the so-called “pole interaction” problem, which refers to the problem of correctly associating formants with the roots of linear prediction coefficients (LPC). Once formants are identified, modification is carried out by changing LPC poles’ angles and radii [15, 16], or direct modification of line spectral pairs [17, 18], usually followed by LPC synthesis. Researchers also proposed modifications in the mel-

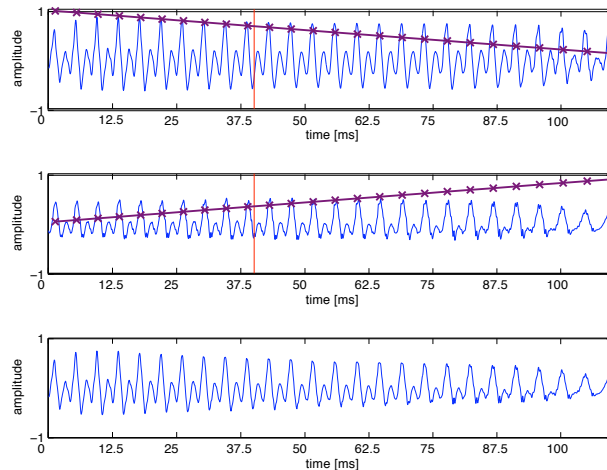


Figure 4: Time-domain cross-fade. Top pane shows left speech segment and middle pane shows right speech segment. Lines with \times markers represent cross-fade weights α and $1 - \alpha$. The traditional cutpoint is displayed as vertical lines. The bottom pane shows the cross-faded waveform.

cepstrum domain using the STRAIGHT analysis and synthesis method [19]. Finally, another modification approach is based on a joint all-pole and sinusoidal model, wherein residual harmonics are warped in accordance with changes to the all-pole model, leading to improved speech quality [14].

In our work, we used a variation on the last approach. The pole-interaction problem did not exist in our case since reliable formant information was available. Figure 2 shows an example of increasing F2 and F3 formant frequencies. In a first step, we constructed an estimate of the frequency response of the speech signal by linearly combining the effects of the vocal tract and the glottal source. The vocal tract was modeled as an all-pole formant filter using the original, manually verified formants F1–F3 information from Section 2.1; in addition, we added higher formants with constant frequency and bandwidths. The glottal source was modeled using a frequency domain representation of a standard glottal flow model [18], with global glottal source parameters that were tuned for the TLL voice. Next, we subtracted the resulting frequency response from an upsampled and smoothed envelope of the individual harmonic sinusoids, as illustrated by Figure 2(a). Then, we frequency warped the resulting residual in accordance with the desired formant frequency changes, and recombined the modified residual with a new formant filter that reflects the desired changes to formant frequencies and bandwidths. Finally, we sampled the new spectral envelope at harmonic intervals to obtain the new sinusoidal parameters, as illustrated by Figure 2(b).

2.3.2. Spectral Band Modification

After formant modification, we calculate the 4-band spectral energies of the modified spectrum, compare the energy values to the desired cross-faded spectral band trajectories, and compute the required gains. For each frame, sinusoidal amplitudes are multiplied with the resulting gain function, after appropriate smoothing to avoid energy discontinuities at the band edges (see Figure 3).

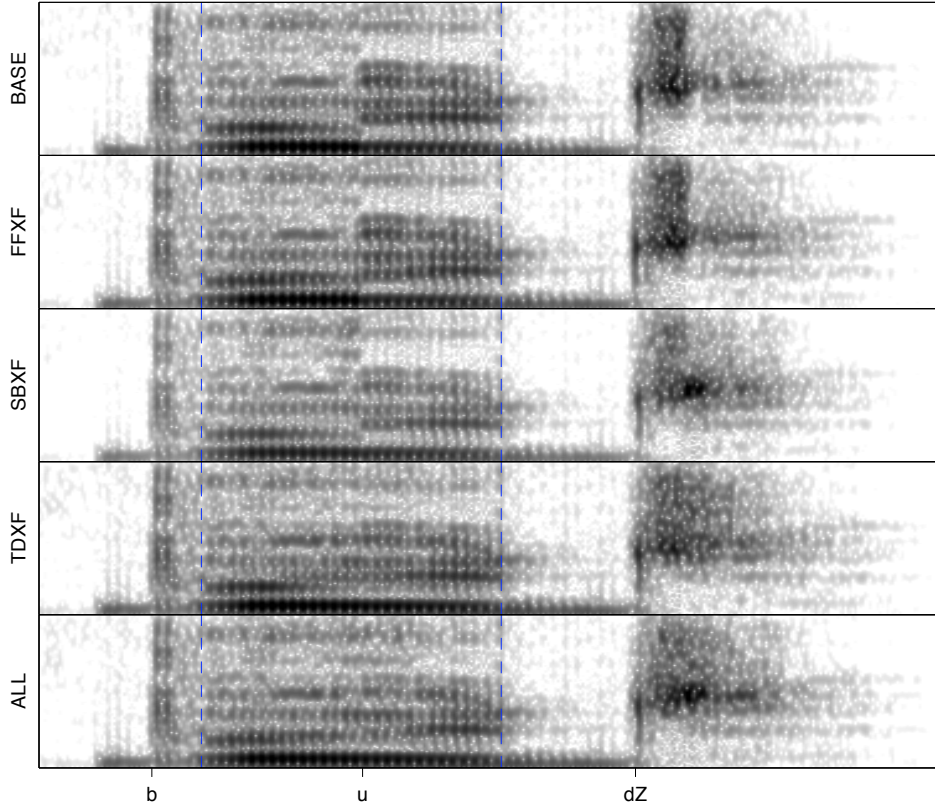


Figure 5: Spectrograms of the word /b u dZ/ in the five conditions. Dashed lines denote phoneme boundaries.

2.3.3. Time-domain Cross-fade

Despite best efforts to explicitly control speech parameters, in our case formant frequencies and spectral band energies, there are likely to be aspects of speech that remain unmodeled. During concatenation, a mismatch of those aspects may be heard as audible discontinuities. To address this problem, we used a time-domain cross-fade approach to make a smooth transition from one (already modified by the methods described above) chunk to the next. This approach required the synthesizer to produce parallel frames of speech with identical features, but from two distinct chunks, during regions of concatenation. We then linearly interpolated between these (pitch-synchronous) segments in the time domain, according to a cross-fade function similar to the one in Section 2.2 (see Figure 4). It should be noted that the TDXF approach implements global energy smoothing inherently.

3. Perceptual Experiment

3.1. Stimuli and Administration

To test the expected quality improvements over a baseline system (BASE), we ran a comparative mean opinion score (CMOS) listening test, using just one of the proposed approaches in isolation (FFXF, SBXF, and TDXF) or all of them jointly (ALL). Stimuli consisted of six vowels (two diphthongs /aI/ and /aU/, and four tense vowels /i:/, /@/l, /u/ and /A/) in a consonant-vowel-consonant (CVC) context.

For each of the six vowels, we were interested in the interactions between given formant frequency or spectral band distances, and the approaches designed to smooth them. There-

fore, we selected stimuli based on two distance types at the concatenation points, namely the formant distance, D_{FF} , and the spectral balance distance, D_{SB} . For each possible vowel concatenation in a $C_1 - V - C_2$ context in the acoustic inventory, we calculated the distances by applying equations

$$D_{FF}(V_L, V_R) = \sqrt{\sum_{k=1}^3 (FF_{k,V_L} - FF_{k,V_R})^2}$$

and

$$D_{SB}(V_L, V_R) = \sqrt{\sum_{k=1}^4 (SB_{k,V_L} - SB_{k,V_R})^2}$$

where V_L represents the left half of a vowel in a $C_1 - V$ context, V_R represents the right half of a vowel in a $V - C_2$ context, FF_{k,V_L} and FF_{k,V_R} represent the k^{th} formant frequencies (in Bark) at the concatenation point of V_L and V_R , and SB_{k,V_L} and SB_{k,V_R} represent the energies in the k^{th} spectral band (in dB) at the concatenation point of V_L and V_R .

After determining both D_{FF} and D_{SB} distances for all possible vowel concatenations, we normalized their values, and selected concatenations at the extremes of these distances, using the Euclidean distance to the four corners of the square spanned by the candidate data. This resulted in four stimulus types: large D_{FF} and large D_{SB} , large D_{FF} and small D_{SB} , small D_{FF} and large D_{SB} , and finally small D_{FF} and small D_{SB} , using the top and bottom 50% of the data for large and small, respectively. We repeated this process for all six vowels, using two concatenations per distance type, resulting in 48 (2 concatenations \times 4 types \times 6 vowels) different CVC words, some of them nonsensical.

| Listener | FFXF | SBXF | TDXF | ALL |
|----------|-------|-------|-------|-------|
| 1 | +0.04 | +0.17 | +0.40 | +0.50 |
| 2 | +0.40 | +0.63 | +0.77 | +1.15 |
| 3 | +0.08 | +0.33 | +0.67 | +0.58 |
| 4 | +0.10 | +0.27 | +0.38 | +0.65 |
| 5 | -0.08 | +0.06 | +0.31 | +0.23 |
| 6 | +0.10 | +0.15 | +0.58 | +0.23 |
| 7 | 0.00 | +0.27 | +0.38 | +0.42 |
| 8 | +0.19 | +0.54 | +0.60 | +0.67 |
| Mean | +0.10 | +0.30 | +0.51 | +0.56 |
| SD | 0.13 | 0.18 | 0.16 | 0.28 |

Table 1: Comparative mean opinion scores for the modified conditions, as compared to the BASE condition. Scores are averaged over all vowels with results shown for individual listeners, as well as the mean and standard deviation for averaged listener responses.

The selected CVC words were generated by an implementation of the proposed approaches in Section 2. We synthesized the selected CVC words under five different conditions: (1) no modifications were applied (BASE), (2) only formant frequency trajectories were cross-faded (FFXF), (3) only spectral band energy trajectories were cross-faded (SBXF), (4) only time-domain cross-fading was applied (TDXF), and (5) all cross-fading operations were applied (ALL). Note that the BASE condition performed a very short version of TDXF as part of the standard procedure of overlap-adding synthesis speech frames.

Each CVC word consisted of four chunks from the acoustic inventory (pause-C → C-V → V-C → C-pause), requiring three concatenation operations. Smoothing operations took place in all three concatenations, except that FFXF was not used when consonants were involved that lacked reliable formant information (such as unvoiced fricatives). We set vowel durations to their median values, as calculated from the acoustic inventory (130 ms for /i:/, 185 ms for /@/, 125 ms for /u/, 175 ms for /A/, 175 ms for /aI/, and 170 for /aU/). We used a naturally falling pitch contour with an average of 220 Hz for each CVC word.

Figure 5 shows spectrograms of the word /b u dZl¹ in all five conditions. The following observations can be made: the vowel and the final consonant are quite discontinuous in the BASE condition. The FFXF condition “connects” the formants of the vowel smoothly (especially F2), but large energy differences remain. The SBXF condition smooths the energy transition in the vowel (this can be seen clearly for F3 and F4), but formant discontinuities remain; however, this condition smooths the final consonant very successfully. The TDXF condition can be seen to smooth the vowel transition by fading one speech unit out as it is fading another unit in; however, formants do not truly connect this way, and at the middle of the cross-fade there are formant “duplicates” (as can be seen by the presence of two F2 tracks towards the middle of the vowel). For the final consonant, TDXF performs an adequate smoothing. Finally, the ALL condition connects formants, equalizes the energy in the four spectral bands, and cross-fades any remaining discrepancies.

The final test stimuli contained pairs of identical CVC words in two different conditions, with a 200 ms separating pause. We compared all 4 modified conditions against the BASE condition, but ignored ordering effects, which resulted in 4 pos-

¹An atypical concatenation inside the frication of the /dZ/ unit was forced for purposes of illustration.

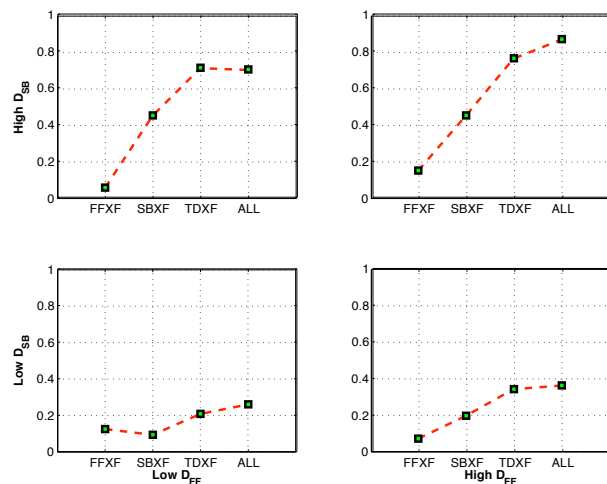


Figure 6: Comparative mean opinion scores for the four modified conditions as compared to the BASE condition, separated into the four stimulus types described in Section 3.1. Scores are averaged over all vowels and listeners.

sible condition pairs and a total of 192 stimuli (48 CVC words × 4 condition pairs).

We recruited 8 normal-hearing (self-reported) listeners, whose native language was American English. Listeners heard stimuli over circumaural headphones. Upon hearing the two words, they were asked to compare them based on quality and processing artifacts, using a scale of -2 (A is much better than B), -1 (A is slightly better than B), 0 (A and B are about the same), +1 (B is slightly better than A), and +2 (B is much better than A). The order of the conditions in a stimulus pair was randomized.

3.2. Results and Discussion

The CMOS values (preference scores) were first transformed to take into account the order of presentation. Table 1 shows the preference scores averaged over all words and listeners. We observed that all individual modifications improved quality, with FFXF yielding the least amount of improvement, followed by SBXF and then TDXF. The combined ALL condition led to the highest overall score. Individual *t*-tests (one-sided) showed significant ($p < 0.05$) differences between the following condition pairs: BASE-FFXF ($p = 0.04$), BASE-SBXF ($p = 0.002$), BASE-TDXF ($p < 0.001$), and BASE-ALL ($p < 0.001$). However, TDXF-ALL ($p = 0.31$) did not show significant differences.

Figure 6 illustrates the relationship between quality scores and conditions, when separated by the four stimulus types defined in Section 3.1. We observed that the ordering of conditions remained mostly invariant across all types. However, we noted that the SBXF condition resulted in a relatively low score for stimulus types for which D_{FF} and D_{SB} was small, and that the ALL condition did not improve upon the TDXF condition for two of the four stimulus types.

To further investigate the relationships between distances and scores of various conditions, we performed a linear regression with D_{FF} , D_{SB} , and $D_{FF} + D_{SB}$ as independent variables and scores Q for various conditions as dependent variable. Table 2 shows correlation coefficients for four relationships of interest, for all available data, and for data for which either D_{FF} or D_{SB} was large or small, respectively. For all data, correla-

| Correlation Coefficient | All | $D_{FF} \uparrow$ | $D_{SB} \uparrow$ | $D_{FF} \downarrow$ | $D_{SB} \downarrow$ |
|----------------------------------------|-------|-------------------|-------------------|---------------------|---------------------|
| $D_{FF} \rightarrow Q_{FFXF}$ | 0.11 | 0.18 | 0.14 | -0.17 | 0.06 |
| $D_{SB} \rightarrow Q_{SBXF}$ | 0.52* | 0.48* | 0.38 | 0.55* | 0.46* |
| $D_{FF} + D_{SB} \rightarrow Q_{TDXF}$ | 0.50* | 0.50* | 0.17 | 0.54* | 0.36 |
| $D_{FF} + D_{SB} \rightarrow Q_{ALL}$ | 0.52* | 0.45* | 0.23 | 0.60* | 0.34 |

Table 2: Correlations between distances and scores, for all data and large (\uparrow) and small (\downarrow) distances. Starred correlations are significant.

tion coefficients were significant at $r = 0.5$, with the exception of predicting Q_{FFXF} from D_{FF} . The latter relationship was not significant for any stimulus types. Predicting Q_{SBXF} from D_{SB} resulted in significantly positive coefficients, except for when D_{SB} was large. Predicting Q_{TDXF} and Q_{ALL} from $D_{FF} + D_{SB}$ resulted in significantly positive coefficients for all data, and for data with large or small D_{FF} ; however, when using data with large or small D_{SB} , coefficients were smaller, and not significant.

4. Conclusion

We proposed two approaches that increase the degree of spectral control in concatenative speech synthesizers, by controlling formant frequencies and energies in four spectral bands. We used the proposed methods (FFXF and SBXF) and one additional time-domain cross-fading technique (TDXF) to smoothly connect from one unit of the acoustic inventory to the next. A comparative mean opinion score listening test showed that all three methods significantly improved perceived quality, to varying degrees. Using all three methods in combination (ALL) was not significantly different from using TDXF alone. We speculate that this is so because (1) even though formants are not continuous in frequency, the human auditory system resolves cross-faded formants with small frequency differences into smoothly varying formants, and (2) a global energy smoothing takes place simultaneously. However, TDXF cannot implement other types of spectral changes, such as controlling the degree of articulation or modeling reduction phenomena due to changes in phoneme duration.

Even though we considered the whole phoneme region for cross-fading in this work, the approach could also be used for smaller regions centered around the point of concatenation.

In the future, we plan on exploring additional capabilities. One example is the application of formant parameters predicted by an explicit model that considers input parameters such as phoneme durations and degree of articulation. Another example is the transformation of formants by a mapping function for voice transformation.

5. References

- [1] D. Klatt, "Review of text-to-speech conversion for English," *JASA*, vol. 82, no. 3, pp. 737–793, Sept. 1987.
- [2] R. H. Manell, "Formant diphone parameter extraction utilising a labelled single-speaker database," in *ICSLP*, Sydney, Australia, 1998.
- [3] C. Shadle and R. Damper, "Prospects for articulatory synthesis: A position paper," in *Proc. of the fourth ISCA Tutorial and Research Workshop*, Perthshire, Scotland, 2001.
- [4] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T Next-Gen TTS System," in *Proc. Joint Meeting of ASA, EAA and DEGA*, 1999.
- [5] P. Taylor, A. Black, and R. Caley, "The Architecture of the Festival Speech Synthesis System," in *Proc. of the third ESCA workshop on speech synthesis*, Jenolan Caves, Australia, 1998.
- [6] Q. Miao, X. Niu, E. Klabbbers, and J. van Santen, "Effects of prosodic factors on spectral balance: analysis and synthesis," in *Speech prosody*, Dresden, Germany, 2006.
- [7] H. Mizuno, M. Abe, and T. Hirowaka, "Waveform-based speech synthesis approach with a formant frequency modification," in *ICASSP*, 1993, pp. 195–198.
- [8] 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.
- [9] 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.
- [10] J. Wouters, *Analysis and Synthesis of Degree of Articulation*, Ph.D. thesis, Oregon Graduate Institute, Portland, OR, 2001.
- [11] D. Broad and F. Clermont, "A methodology for modeling vowel formant contours in CVC context," *JASA*, vol. 81, no. 1, pp. 155–165, Jan. 1987.
- [12] E. Klabbbers, J. van Santen, and A. Kain, "The contribution of various sources of spectral mismatch to audible discontinuities in a diphone database," *IEEE Transactions on Audio, Speech, and Language Processing Journal*, vol. 15, no. 3, pp. 949–956, 2006.
- [13] Agaath Sluijter, *Phonetic Correlates of Stress and Accent*, Ph.D. thesis, Holland Institute of Generative Linguistics, 1995.
- [14] J. Wouters and M. Macon, "Control of spectral dynamics in concatenative speech synthesis," *IEEE Trans. Speech and Audio Proc.*, vol. 9, no. 1, pp. 30–38, Jan. 2001.
- [15] Y.-S. Hsiao and D.G. Childers, "A new approach to formant estimation and modification based on pole interaction," in *Thirtiethasilomar conference on signals, systems and computers*, 1996, vol. 1, pp. 783–787.
- [16] E. Turajlic, D. Rentzos, S. Vaseghi, and C.-H. Ho, "Evaluation of methods for parametric formant transformation in voice conversion," in *ICASSP*, 2003, pp. 724–727.
- [17] R. W. Morris and M. A. Clements, "Modification of formants in the line spectrum domain," *IEEE Signal Processing Letters*, vol. 9, pp. 19–21, Jan. 2002.
- [18] Boris Doval, Christophe d'Allesandor, and Nathalie Henrich, "The voice source as a causal/anticausal linear filter," in *VOQUAL*, Aug. 2003.
- [19] T. Toda, A. W. Black, and K. Tokuda, "Spectral Conversion Based on Maximum Likelihood Estimation Considering Global Variance of Converted Parameter," in *ICASSP*, Mar. 2005, vol. 1, pp. 9–12.