SPECTRAL SMOOTHING FOR CONCATENATIVE SPEECH SYNTHESIS *

David T. Chappell and *John H. L. Hansen*
Robust Speech Processing Laboratory
Duke University, Box 90291, Durham, NC 27708-0291
http://www.ee.duke.edu/Research/Speech d.chappell@ieee.org jhlh@ee.duke.edu

ABSTRACT

This paper addresses the topic of performing effective concatenative speech synthesis with a limited database by proposing methods to smooth the transitions between speech segments. The objective is to produce natural-sounding speech via segment concatenation when formants and other spectral features do not align properly. We propose several methods for adjusting the spectra between waveform segments selected for concatenation. Techniques examined include optimal coupling, waveform interpolation, linear predictive pole shifting, and psychoacoustic closure. Several of these algorithms have been previously developed for either coding or synthesis, but our application of closure for segment processing is novel. After spectral smoothing, the final synthesized speech can better approximate the desired speech characteristics and is continuous in both the time domain and spectral structure.

1. PROBLEM

Many concatenative text-to-speech systems produce continuous speech by selecting waveform segments from databases with a large number (i.e., +25,000) of segments with varied characteristics [5, 6]. Direct concatenation of segments from such a large database can yield high speech quality since the database contains enough sample segments to include a close match for each desired segment, but this technique is costly in terms of database collection, search requirements, and segment memory storage. Other concatenative synthesis systems use a set of specially selected diphones with boundaries set at the centers of phonemes where formants are stable. In both approaches, the formants may not align perfectly, but the spectral alignment is generally acceptable.

With a small database, however, direct concatenation of available speech segments sometimes produces a series of segments which fail to match the desired parameters. Time-domain techniques can easily adjust the prosodic characteristics, but additional processing is needed to spectrally align the selected speech segments and avoid discordance. In the absence of spectral smoothing, formants and other spectral characteristics will change abruptly at the transitions between concatenated speech segments. These spectral discontinuities will be present in most concatenative synthesis systems but will be more noticeable in a small-database environment.

Algorithms have been formulated to adjust the pitch to be continuous between segments; however, the remainder of the spectral structure is not so easily modified. The concatenation of segments is time-synchronized by pitch peaks, and the Pitch-Synchronous Overlap Add (PSOLA) algorithm [9] adjusts the duration and fundamental frequency. Merely averaging the signal in the time domain does not cause the formants to match correctly in the frequency domain, however, and thus spectral smoothing must be tackled as a special problem.

In the absence of spectral smoothing, unnatural spectral transitions will arise. For example, spectral peaks will suddenly appear and disappear. Peaks will fade and appear at nearby frequencies rather than shifting between frequencies. Studies have shown that smooth changes in frequency and spectrum are interpreted as changes within a single speaker, whereas sudden changes are interpreted as being a change in speaker [8]. A spectral smoothing scheme can eliminate these audibly unnatural transitions. The goal of this study is therefore to propose several spectral-based smoothing and adjustment algorithms to address spectral discontinuity in segment-based concatenative synthesis.

2. SPECTRAL SMOOTHING

We have developed several methods in both the time and frequency domains to smooth transitions between concatenated speech segments. These approaches include simple spectral averaging, more complex formant-adjusting methods, and a psychoacoustic technique. We consider existing techniques and improvements to demonstrate their application to spectral smoothing for concatenation.

We have implemented several approaches to spectral smoothing and detail here only those methods which have yielded the best results. Although a few researchers have studied smoothing techniques (e.g., audio morphing [10]), the field remains fresh and typically only common existing

*This study was supported in part by a contract from the U. S. government.

speech processing algorithms (e.g., linear prediction techniques described below) are employed. Several of these techniques were originally developed for other purposes, including interpolation for audio coding and voice modification, and they are not generally applied to spectral smoothing for concatenative synthesis. Here we describe only the spectral smoothing applications of the algorithms and do not discuss their original applications.

Our general approach to smoothing is to take one frame of speech from the edge of each segment and interpolate between them. It is thus important that the edge frames are good representatives of the sound (*see Section 3*). We perform linear interpolation in different domains between the two frames, though we also suggest cubic spline interpolation as an alternative. The frames are pitch-synchronous where one frame is two pitch periods long; this synchronization is important for some interpolation methods.

One important issue of spectral smoothing is determining in what circumstances the smoothing should be performed. If two segments have a sufficiently close spectral match, then the distortion introduced by smoothing techniques may outweigh the performance gain. Moreover, many smoothing techniques are inappropriate for use with unvoiced speech.

Another issue is determining the best time span over which to interpolate. The pitch will remain continuous if data is inserted equal to an integer number of pitch periods. Our experiments showed that three to five periods generally works well; however, more studies should be done to determine the proper number of pitch periods for different circumstances.

3. OPTIMAL COUPLING

It is common in concatenative synthesis that the boundaries of speech segments are fixed, but the optimal coupling technique allows the boundaries to move to provide the best fit with adjacent segments [2]. A measure of mismatch is tested at a number of possible segment boundaries until the closest match is found. While any form of measure may be used, for the sake of improving spectral quality, using a spectral discontinuity measure is appropriate. Measures considered include mel-frequency cepstral coefficients (MFCC) and the auditory-neural based measure (ANBM) [4]. It is not necessary to implement optimal coupling to perform spectral smoothing, but it does provide some improvement at a small cost.

4. WAVEFORM INTERPOLATION

Waveform interpolation (WI) is a speech-coding technique which takes advantage of the gradual evolution of the shape of pitch period waveforms. In WI, a waveform is interpolated in either the time or frequency domains. In order to conserve space in coding, a signal is typically transmitted as quantized frequency coefficients for separate rapidly and slowly evolving components [7]. The waveform is typically one pitch period long, but the length may be an integer number of periods.

Though developed for coding purposes, WI can also be used for spectral smoothing. In this case, the waveform is interpolated between the frames at the edges of speech segments to create smoothed data to insert between them. The concept is the same as for coding, but the end goal is different. For synthesis, the original waveform can be kept intact for interpolation rather than compressing the data via quantization. When the original waveforms are available, interpolating in either the time or the frequency domain yields identical results. A new pitch period of the desired length is constructed by averaging the amplitudes of the periods of natural speech at the same relative positions within the waveforms.

In addition to direct use for calculating smoothed speech frames, WI can also be applied for residual interpolation. Linear prediction methods (*see Section 5*) concentrate on interpolating the spectral envelope, but the residual signal must also be generated. Rather than using a generic pulsed excitation or a single residual appropriate for the speaker, we use WI to interpolate between the residuals of the bordering frames of natural speech.

5. LP TECHNIQUES

Linear prediction (LP) interpolation techniques are often used with the intention of smoothing LP-filter coefficients in LP coding (LPC). The basic strategy is to model the signal as spectral and excitation components and to adjust each component separately. Here, we discuss interpolating the LP spectral parameters in any of several domains, while the residual is interpolated using waveform interpolation (*see Section 4*).

There are various representations of the LP parameters: prediction coefficients (PC), reflection coefficients (RC), log area ratios (LAR), arcsine of the reflection coefficients (ASRC), cepstral coefficient, line spectrum pairs (LSP), line spectral frequency differences (LSFD), autocorrelation coefficients (ACF), impulse response (IR). In speech coding, these various LPC parameters can be interpolated so that less data can be transmitted, but in the case of spectral smoothing we interpolate so that we can construct new data to fill in the gaps between existing segments of speech. Research in coding shows that some representations perform better for interpolation than others. Some representations (e.g., LPC coefficient, cepstral coefficient, and impulse response) can yield an unstable LPC synthesis filter after interpolation, and thus they generally are not used for interpolation. The LSP representation is generally accepted as giving the best performance in terms of spectral distortion, and it always yields stable filters after interpolation [7].

In speech coding, the LP poles are rarely shifted directly in the z-plane because the parameters are usually stored and transmitted in another representation. The line spectrum pair (LSP) representation, also known as line spectral

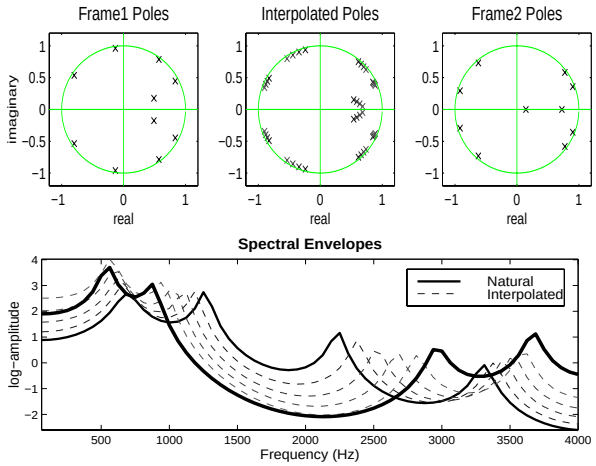

Figure 1: Example LP pole shifting scenario.

frequency (LSF), is often used for speech coding. Interpolation between LSPs has been used not only for coding but also for synthesis and even spectral smoothing. For waveform synthesis, LPSs lose the compression advantage over the direct use of poles or other representations.

When LP poles are shifted, pole positions should not be directly linearly interpolated in the complex plane. Instead, the magnitude and phase of the poles should be interpolated separately. Interpolating in the complex plane can produce values which are not truly intermediate between the original poles, but interpolating in the magnitude-phase domain produces more reasonable results. The magnitude of a pole relates to the bandwidth of a corresponding formant, while the angle relates to the frequency. Ideally, each LP pole would correspond to a single formant, but in practice multiple poles will affect the location and bandwidth of each formant. Thus, although pole shifting modifies formants, it can have undesired effects such as formant bandwidth spreading.

In pole shifting, a common problem arises when one frame of speech has more real poles than the adjoining frame. Figure 1 illustrates this scenario. One solution is to convert to a domain where each pole has a complex conjugate [3]. Our solution is to first perform a matching of conjugate pairs that results in the minimum total distance between matched pairs. For each remaining unmatched conjugate pair, we select the nearest single real pole as a match.

6. CONTINUITY EFFECT

The continuity effect is a psychoacoustic phenomenon. When two sounds are alternated, a less intense masked sound may be heard as continuous despite being interrupted by a more intense masking sound. The sensory evidence presented to the auditory system does not make it clear whether or not the obscured sound has continued. Psychologists call this effect "closure" [1, 8].

Perceptual closure occurs when a missing gap in sound is filled by a noise or other sound that masks the missing sound. The visual counterpart to auditory closure is looking at a scene while moving past a picket fence; the observer assumes that the scene continues uninterrupted behind the fence boards even though only part of the scene is visible at any one time. In auditory perception, illusory continuity requires that the masking sound be near enough in frequency to the missing sound for simultaneous masking to occur according to the neural response of the peripheral auditory system.

The continuity effect has also been shown to work for speech signals alternated with noise.A series of studies has shown that bursts of noise interrupting speech at the rate used in phone or diphone concatenation (about 6 per second) is near a minimum in the effects of the noise on speech comprehension. Moreover, with this interruption frequency and the desired fraction of time spent on speech vs. noise (91%), listener tests revealed a very high word articulation rate. In some circumstances, interrupting noise has been shown to actually increase intelligibility [1, 8].

In the case of spectral smoothing, the continuity effect can be employed by adding noise between speech segments. Although closure has not been previously applied to speech synthesis, the concept is not entirely foreign: in some audio systems, large burst errors are sometimes filled with white noise. We take the concept a step further by spectrally shaping the noise so that it contains only the spectral envelope necessary to possibly contain any intermediate sound. The listener's perception fills in any gaps so that it seems as though speech is being produced within the noise, and the perceived speech is continuous with the preceding and following existing speech.

Figure 2 shows an example of a frequency-domain filter for inserted noise. The spectral envelopes of the two original speech frames are compared. The filter is constructed to meet the maximum of the two envelopes at all points and to interpolate between any large peaks (presumably formants) between the two spectra. Gaussian white noise is passed through this filter to create shaped noise that will mask any hypothetical speech between the two natural frames without introducing more noise than necessary for the auditory masking.

7. RESULTS & EVALUATIONS

Initial testing and algorithm development was done with simple tests that involved concatenating sets of two phones. Later testing was done by integrating the promising schemes into a concatenative speech synthesis system. The waveform concatenation system used for evaluating these

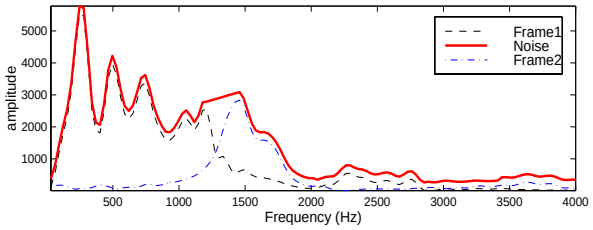

Figure 2: Example noise envelopes for continuity effect.

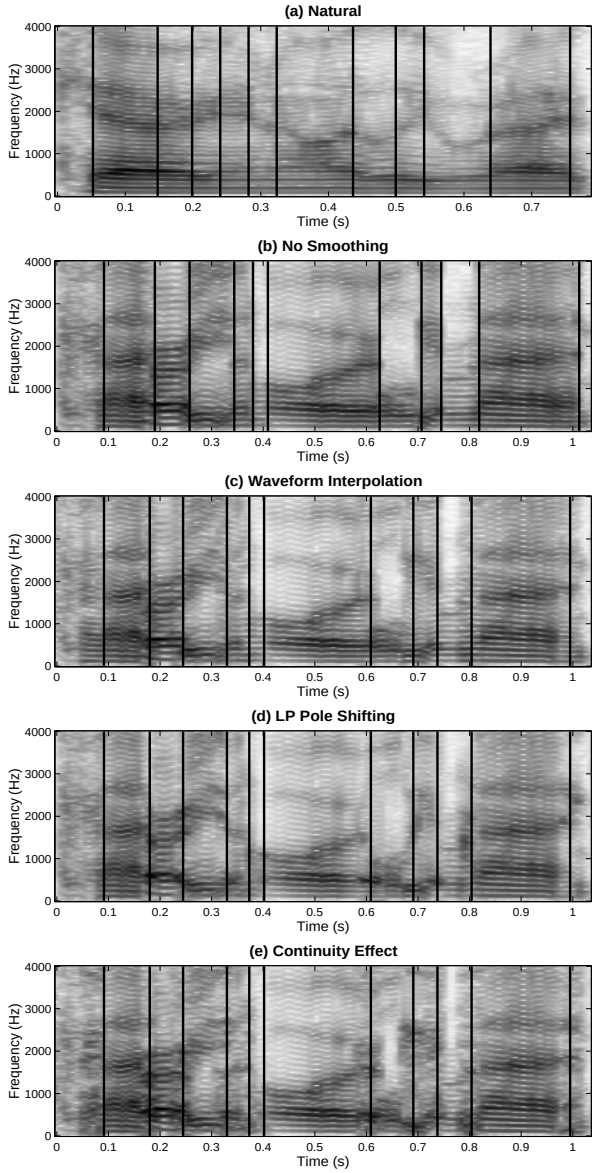

Figure 3: Spectrograms of the phrase "carry an oily rag": (a) naturally produced and (b)-(e) concatenated speech. Solid vertical lines mark segment boundaries.

smoothing algorithms is designed for synthesis when only a limited amount of training data is available. These methods assume a data corpus of only about four hundred phone segment waveforms per speaker as is generally found in the TIMIT database. In each of the reported cases, the synthesizer incorporates optimal segment coupling.

Figure 3 shows one example spectrogram from each technique for the same concatenated speech segments: "carry an oily rag." Note that the formants tend to be smoother and more continuous in several of the techniques, especially with LP pole shifting.

We are also taking objective measurements from listener tests. Table 1 shows the preliminary results from a larger formal listening test. Four expert listeners were asked to indicate their preferences in terms of naturalness and intelligibility for four different words and phrases. The listeners

Algorithm	Natural	Intelligible
Natural Speech	1.13	1.38
Raw Concatenation	2.75	2.66
Waveform Interp.	3.09	3.22
LP Pole Shifting	3.75	3.53
Continuity Effect	4.41	4.34

Table 1: Listener preferences (lower values are better).

ranked spectral smoothing algorithms algorithms as compared with raw concatenated speech (without smoothing) and with natural speech. In general, listeners seemed to dislike the noise of the continuity effect and the audible spectral components of LP pole shifting. The scores reinforce the conclusion that spectral smoothing will sometimes yield improvements yet sometimes make the resulting speech worse.

8. CONCLUSION

The net results of the proposed algorithms are mixed. Some resulting smoothed synthesized phrases demonstrate noticeable improvements over standard techniques applied to small databases, while other phrases are of lesser quality. The final speech has smoother, more continuous formants and is sometimes more natural-sounding than direct concatenation of selected segments without processing. Therefore, when properly employed, spectral smoothing techniques can improve the results of concatenative speech synthesis with a limited database.

9. REFERENCES

[1] A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound.* MIT Press, Cambridge, Massachusetts, 1990.

[2] A. D. Conkie and S. Isard. "Optimal Coupling of Diphones". J. P. H. van Santen *et al.*, eds., *Progress in Speech Synthesis*, pp. 293–304. Springer-Verlag, New York, 1997.

[3] V. Goncharoff and M. Kaine-Krolak. "Interpolation of LPC Spectra via Pole Shifting". *Proc. 1995 IEEE ICASSP*, vol. 1, pp. 780–783. 1995. Detroit, Michigan.

[4] J. H. L. Hansen and D. T. Chappell. "An Auditory-Based Distortion Measure with Application to Concatenative Speech Synthesis". *IEEE Transactions on Speech and Audio Processing*, 6(5):489–495, September 1998.

[5] T. Hirokawa and K. Hakoda. "Segment Selection and Pitch Modification for High Quality Speech Synthesis using Waveform Segments". *Proc. 1990 ICSLP*, vol. 1, pp. 337–340, Nov. 1990. Kobe, Japan.

[6] A. J. Hunt and A. W. Black. "Unit Selection in a Concatenative Speech Synthesis System using a Large Speech Database". *Proc. 1996 IEEE ICASSP*, pp. 373–376. May 1996. Atlanta, Georgia.

[7] W. B. Kleijn and K. K. Paliwal, eds. *Speech Coding and Synthesis.* Elsevier, Amsterdam, 1998.

[8] B. C. J. Moore. *An Introduction to the Psychology of Hearing.* Academic Press, New York, 4th ed., 1997.

[9] E. Moulines and F. Charpentier. "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis using Diphones". *Speech Comm.*, 9:453–467, 1990.

[10] M. Slaney, M. Covell, and B. Lassiter. "Automatic Audio Morphing". *Proc. 1996 IEEE ICASSP*, pp. 1001–1004.