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 pro-
spectra between
p"ose several methods for adjusting the waveforms selected for concatenation. Techniques
examined include optimal coupling, waveform interpola-
tion, 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 approx-
imate the desired speech characteristics and is continuous
in both the time domain and spectral structure.

1. PROBLEM
Many concatenative text-to-speech systems produce con-
tinuous 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 seg-
ments to include a close match for each desired segment,
but this technique is costly in terms of database collec-
tion, search requirements, and segment memory storage.

Other concatenative synthesis systems use a set of spe-
cially selected diphones with boundaries set at the cen-
ters of phonemes where formants are stable. In both ap-
proaches, 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 spe-
trally align the selected speech segments and avoid discord-
ance. In the absence of spectral smoothing, formants and

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

other spectral characteristics will change abruptly at the
transitions between concatenated speech segments. These
spectral discontinuities will be present in most concatena-
tive 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 fre-
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 sud-

"enly appear and disappear. Peaks will fade and appear
at nearby frequencies rather than shifting between frequen-
cies. 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 concate-
nated speech segments. These approaches include simple
spectral averaging, more complex formant-adjusting meth-
ods, and a psychoacoustic technique. We consider existing
techniques and improvements to demonstrate their appli-
cation 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 [16]),
the field remains fresh and typically only common existing
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), auto-correlation 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
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
Comprehensive auditory scene analysis

In the present work, we are focusing on the auditory scene analysis. An auditory scene is defined as a set of auditory events that are perceived as occurring simultaneously. Previous studies have shown that the auditory scene analysis is a complex process that involves the integration of multiple auditory cues, such as spectral, temporal, and spatial information. The goal of this study is to develop a model that can accurately predict the auditory scene analysis.

The model is based on a combination of statistical and neural network techniques. The statistical part of the model is used to estimate the probability of each auditory event occurring, given the observed auditory cues. The neural network part of the model is used to integrate these probabilities and make a final prediction of the auditory scene.

The model was trained on a large dataset of audios, and it was able to accurately predict the auditory scene in most cases. The results of this study suggest that the development of an accurate model for auditory scene analysis is possible, and this could have significant implications for a variety of applications, such as speech recognition, hearing aids, and virtual environments.