D.T. Chappell, **J.H.L. Hansen**, "Spectral Smoothing for Speech Segment Concatenation," Speech Communication, Volume 36, Issues 3-4, March 2002, Pages 343-373.

A comparison of spectral smoothing methods for segment concatenation based speech synthesis

D.T. Chappell [†], **John H.L. Hansen,**

Robust Speech Processing Group
Center for Spoken Language Research
University of Colorado Boulder, Campus Box 594
(Express Mail: 3215 Marine Street, Room E-265)
Boulder, Colorado 80309-0594
303 – 735 –5148 (Phone) 303 – 735 – 5072 (Fax)
John.Hansen@colorado.edu (email)
[†] **Department of Electrical Engineering**,
Duke University,
P.O. Box 90291,, Durham, NC 27708-0291,
chappell@ee.duke.edu (email)

Speech Communication,
Volume 36, Issues 3-4, pp 343-373, March 2002.

ELSEVIER

Speech Communication 36 (2002) 343–374

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

A comparison of spectral smoothing methods for segment concatenation based speech synthesis ☆

David T. Chappell [b], John H.L. Hansen [a,b,*]

[a] *Robust Speech Processing Laboratory (RSPL), Center for Spoken Language Research (CSLR), Room E265, University of Colorado, 3215 Marine St., P.O. Box 594, Boulder, CO 80309-0594, USA*

[b] *Department of Electrical Engineering, P.O. Box 90291, Duke University, Durham, NC 27708-0291, USA*

Received 21 April 1999; received in revised form 24 May 2000; accepted 15 December 2000

Abstract

There are many scenarios in both speech synthesis and coding in which adjacent time-frames of speech are spectrally discontinuous. This paper addresses the topic of improving concatenative speech synthesis with a limited database by proposing methods to smooth, adjust, or interpolate the spectral 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 consider several methods for adjusting the spectra at the boundaries between waveform segments. Techniques examined include optimal coupling, waveform interpolation (WI), linear predictive parameter interpolation, and psychoacoustic closure. Several of these algorithms have been previously developed for either coding or synthesis, while others are enhanced. We also consider the connection between speech science and articulation in determining the type of smoothing appropriate for given phoneme–phoneme transitions. Moreover, this work incorporates the use of a recently-proposed auditory-neural based distance measure (ANBM), which employs a computational model of the auditory system to assess perceived spectral discontinuities. We demonstrate how actual ANBM scores can be used to help determine the need for smoothing. In addition, formal evaluation of four smoothing methods, using the ANBM and extensive listener tests, reveals that smoothing can distinctly improve the quality of speech but when applied inappropriately can also degrade the quality. It is shown that after proper spectral smoothing, or spectral interpolation, the final synthesized speech sounds more natural and has a more continuous spectral structure. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Speech synthesis; Speech coding; Spectral smoothing; Spectral interpolation

1. Introduction

When speech is produced naturally by a human, there is a measurable degree of continuity between phone segments. This degree of continuity is related to the physical movement and placement of the vocal system articulators. When speech is produced artificially, such as in segment-based synthesis or in low-bit-rate coding, the same phone-to-phone continuity may not exist.

Speech synthesis, coding, and voice transformation can benefit from improvements in spectral smoothing. There are a number of scenarios in which the spectral structure of speech at adjacent

☆ This work was supported in part by SPAWAR under grant No. N66001-92-0092.

* Corresponding author. Tel.: +1-303-735-5148; fax: +1-303-735-5072.

E-mail address: John.Hansen@colorado.edu (J.H.L. Hansen).

PII: S0167-6393(01)00008-5

time observations is not smooth. Listeners can detect abrupt changes in the timbre of the speech, that cause the speech to sound unnatural. Although there are several existing techniques to smooth or interpolate the spectral structure of speech (Slaney et al., 1996; Conkie and Isard, 1997; Plumpe et al., 1998), most studies consider only a single method with limited performance comparison. Hence, there is still much room for improvement.

Text-to-speech (TTS) systems based on concatenation produce continuous speech by selecting waveform units from speech databases. Many of these systems use databases with a large number (e.g., over 50,000) of available segments with varied characteristics. This approach can yield high-quality speech (Hunt and Black, 1996; Hirokawa and Hakoda, 1990), but such algorithms succeed primarily because of their large databases. These large-database synthesis schemes generally concentrate on segment choice and search algorithms since the corpora contain enough sample units to include a reasonably close match for each desired phoneme. In contrast, concatenative speech synthesis with a smaller database of only a few hundred to thousand phone unit waveforms per speaker will yield more discontinuities at segment boundaries. With a smaller database, each speech segment must be modified to fit the desired properties. The spectral characteristics of the beginning and ending of each phone must be molded to smooth the transition between adjoining units. It is also important to know when spectral smoothing is needed. For example, there should be a high degree of continuity between /aa/ and /r/ in the word *car*, but more discontinuity between /k/ and /aa/ in the same word. While spectral smoothing can benefit speech synthesis based on both large and small databases, smoothing is more valuable for smaller databases.

In this paper, our emphasis is on small database size concatenative synthesis, with a focus on contrasting the strengths and weaknesses of spectral smoothing algorithms. We point out, however, that spectral smoothing algorithms are useful in other scenarios as well such as LP based speech coding and voice transformation. The basic waveform unit used here is the phone, with multiple adjacent phones used when appropriate matches are found in the speech database.

The paper outline is as follows. First, we consider relevant background on topics in concatenative speech synthesis, spectral smoothing, and articulation. We also review a recently-proposed auditory-based distance measure that can aid the smoothing process. Next, we present several spectral smoothing algorithms. Each algorithm's description is accompanied by a discussion and evaluation. In Section 4, we describe the method used to determine which approach to spectral smoothing to use for various phone-to-phone concatenation cases. Finally in Section 5, we present formal listener tests to evaluate the quality of the various spectral smoothing algorithms considered. We also describe algorithm results and discuss our findings on the proper use of spectral smoothing methods.

2. Background

This section provides background on several topics important to the research discussed within this paper. Several sources discuss these topics in more detail (e.g., Deller et al., 2000; Flanagan, 1972; O'Shaughnessy, 1990; Witten, 1982).

2.1. Concatenative synthesis overview

There are a number of different techniques for synthesizing speech. The technique used in this work, concatenative synthesis, starts with a collection of speech waveform signals and concatenates individual segments to construct a new utterance. The concatenation process itself is straightforward, though large databases may require complex search algorithms, and signal processing may be used to modify the constructed signal to achieve some desired speaker characteristics. The final speech is more natural and the speaker is more recognizable than with other forms of synthesis since concatenative synthesis begins with a set of natural speech segments.

The database of recorded waveform segments is typically in one of two forms. Many concatenative speech synthesis systems produce continuous speech by selecting waveform segments from databases with a large number (i.e., +50,000) of segments with varied characteristics (Hirokawa and Hakoda, 1990; Hunt and Black, 1996; Huang et al., 1997; Breen and Jackson, 1998). These large databases are usually not recorded with concatenative synthesis in mind but instead are generic speech corpora. 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; however, such a technique is costly in terms of database collection, search requirements, and segment memory storage and organization. For databases that contain multiple instances of each speech unit, synthesizers commonly select segments based upon two cost functions: the *target cost* compares available segments with a theoretical ideal segment, and the *concatenation cost* measures the spectral continuity between potentially concatenated segments (Hunt and Black, 1996). A recent study has compared several spectral distance measures to determine which measures best predict audible discontinuities when used as concatenation costs (Klabbers and Veldhuis, 1998).

In comparison, other concatenative synthesis systems use a set of specially selected diphones with boundaries set at the phoneme centers where formants are stable. These databases are much smaller and contain only one example of each diphone in the language (about 1200 in English). Such diphone databases are typically recorded specifically for concatenative synthesis. In both database styles – generic corpora and diphone databases – the formants of concatenated speech segments may not align perfectly, but the spectral alignment is generally reasonable.

The limited amount of speech in any given database is unlikely to include segments that precisely match the desired reference segment, given the existence of any knowledge of reference segment characteristics. While most synthesizers simply take the nearest matching segment as it stands without additional processing, some systems will modify the segment before concatenation. Pitch-synchronous overlap and add (PSOLA) is often used to adjust the segment pitch and duration to match the desired reference. By manipulating pitch-synchronous analysis windows, PSOLA provides a simple mechanism for prosodic adjustment (Moulines and Charpentier, 1990; Moulines and Laroche, 1995). While a perfectly matched segment is desirable, modifying the available data is a practical method of achieving similar results. Modifying the three prosodic characteristics – pitch, duration, and power – allows a limited database to produce a wider range of speech segments for concatenation. Many implementations of PSOLA do not include spectral smoothing in order to minimize the computational complexity, but we have expanded upon the basic time domain PSOLA algorithm to incorporate smoothing.

2.2. Spectral smoothing

In both speech synthesis and audio coding, there are circumstances where subsequent data segments have audibly different spectra at their adjoining boundaries. Signal processing can be used to smooth the existing waveform or create new data to bridge the gap between segments resulting from compression or coding errors. Straightforward linear interpolation in the frequency domain does not yield acceptable results, and therefore alternative algorithms (see Section 3) are needed to provide more natural transitions. It is noted that *spectral smoothing* generally indicates modification of existing audio frames and *spectral interpolation* means the addition of frames; here we emphasize the addition of frames but do not distinguish between the two terms.

In the absence of spectral smoothing, unnatural spectral transitions will arise. Studies have shown that smooth changes in frequency are perceived as changes within a single speaker, whereas sudden changes are perceived as being a change in speaker (Moore, 1997). Other research has shown that formant discontinuities are audible in TD-PSOLA synthesis (Donovan, 1996). Spectral smoothing can eliminate these audibly unnatural transitions. Therefore, the goal of this study is to explore several spectral-based smoothing and adjustment algorithms to address spectral discontinuity for segment-based concatenative synthesis and to explore ways to determine when and where the smoothing should be applied.

At present, spectral smoothing is most commonly used for speech and audio coding. Similar methods are sometimes used for speaker transformation (Savic and Nam, 1991; Mizuno and Abe, 1995; Slifka and Anderson, 1995). In comparison, spectral smoothing is only sometimes used for speech synthesis (Mizuno et al., 1993). Even though our experiments have focused on TD-PSOLA synthesis, other researchers have successfully applied spectral smoothing to other synthesis algorithms such as the Harmonic/Stochastic (H/S) model and multi-band resynthesis PSOLA (MBR-PSOLA or MBROLA) (Dutoit and Leich, 1993; Dutoit, 1994) as well as the harmonic plus noise model (HNM) (Syrdal et al., 1998). In some cases, spectral smoothing of concatenated speech can degrade synthesis quality rather than yield improvement (i.e., produce various artifacts such as suddenly appearing/disappearing narrowband peaks, spectral peaks fading and rising versus shifting in frequency, and nonlinear peak frequency shifts (Goncharoff and Kaine-Krolak, 1995)). Spectral smoothing tends to perform best when the original spectra are similar to each other; such as in speech coding and concatenative synthesis with large or specially-designed databases.

2.3. Spectral distance measure

In a previous study, an auditory-neural based measure (ANBM) was proposed which aids in the selection of speech units for speech synthesis via segment concatenation (Chappell and Hansen, 1997; Hansen and Chappell, 1998). The ANBM measures the "distance" between the spectral characteristics of two adjacent time-slices in the speech signal. It differs from other spectral distance measures in being based upon a model of mammalian auditory perception.

The ANBM uses the output of a computational auditory model to generate one feature vector for each frame of speech. First, a computational model generates the average firing rates of synapses of auditory nerves. We use Carney's nonlinear auditory model, which is based upon and closely approximates measurements of auditory nerve (AN) fibers in cats (Carney, 1992). The auditory model calculates the time-varying spike rate for the synapse between an inner hair cell and an AN. Next, the analysis stage locates the primary, or modal, frequency at which each AN fires. To find the primary firing frequency for an AN channel, we first calculate the spectrum and then find the frequency for the corresponding peak absolute value. This dominant frequency is stored in the feature vector for that frame of speech. Both the auditory model and the measure's analysis stage operate on each AN channel separately; for each channel k for a given frame, the analysis stage stores the primary firing frequency value x_k within the feature data vector $\vec{x}$. Finally, the feature vectors are compared via the city-block metric shown below to estimate the perceived mismatch between frames of speech,

$$d_1(\vec{x},\vec{y}) = \sum_{k=1}^{N} |x_k - y_k|. \tag{1}$$

A lower ANBM score implies less perceived auditory difference, while a larger score implies greater perceived discontinuity.

This measure can therefore provide information on the amount of perceptual segment mismatch to direct additional signal processing to smooth any discontinuities or disfluencies. Here, we consider its use for determining whether a concatenated segment boundary is sufficiently smooth, though it may also be useful for determining the degree to which a speech signal sounds natural or concatenated.

While the original formulation of the ANBM did not specify the number of auditory-nerve channels, we chose to use 32. Using the known characteristic frequencies for cat inner hair cells, 32 channels cover characteristic frequencies from 100 Hz to 3587 Hz (Liberman, 1982). Phase locking, which is the mechanism assumed in the technique of finding the modal firing frequency, is known to occur only below 4–5 kHz (Moore, 1997).

2.4. Articulation

Articulation is rarely considered in spectral smoothing. Nonetheless, knowledge of manner of articulation and its acoustic correlates can aid in spectral smoothing. Speech articulation and its connection to acoustics are well-understood and described in several texts (e.g., Coker, 1976; Deller et al., 2000; Ladefoged, 1975; Moore, 1997; O'Shaughnessy, 1990; Pickett, 1980).

The effects of coarticulation on formant movement represent a major cause for the need for spectral smoothing. This influence represents a serious challenge for segment-based synthesis when segment codebook sizes are small since segments are more likely to be individual phones. To overcome these problems, many speech synthesis systems use diphone units, which are bordered at the relatively stable positions in the center of phonemes, rather than phones, which are bordered at the unstable transition positions.

Coarticulation is caused by articulators moving smoothly into position and gradually returning to neutral positions over the course of one or more phones. When an articulator's motion is not directly involved in the production of a phone, it is free to move according to previous and subsequent phonemes. For example, labial phonemes allow the tongue to move freely while lingual phonemes allow the lips to move. The limits of motion of the articulators used for speech production are at different rates of speed (O'Shaughnessy, 1990; Zemlin, 1968), which implies that the transition periods between different phonemes should have different durations. Research using an articulatory model has demonstrated the effects of movement of articulatory organs on segmental duration (Shiga et al., 1998).

Acoustics and articulation are important for spectral smoothing as well as general synthesis due to the effects of coarticulation on formant positions. Some phonemes can yield similar steady-state spectra but differ in phone transitions (e.g., /d/ and /g/ versus /b/) (Parthasarathy and Coker, 1992). In nasalization and rhotacization, a consonant colors the spectrum of adjacent vowels in a predictable way.

In English, the first three formants largely determine the phoneme. F_1 is high when the tongue constriction is nearer the glottis and when the mouth opening is large and unrounded. F_2 generally increases as the point of constriction moves forward from the glottis, as the tongue constriction narrows, and as the size of the mouth opening increases. F_3 increases as the constriction moves forward from the glottis and also as the mouth opening increase in size and becomes less rounded. Moreover, formant bandwidth depends upon the degree of constriction: open vowels have narrower formants than constricted vowels (Stevens and House, 1955). Fig. 1 illustrates the correlation between articulation and the first three formant positions for vowels (Ladefoged, 1981). Several sources (e.g., Fant, 1960; Witten, 1982; Deller et al., 2000; Ladefoged, 1975; O'Shaughnessy, 1990) cite average formant frequencies for various vowels and describe the spectral structure for consonants.

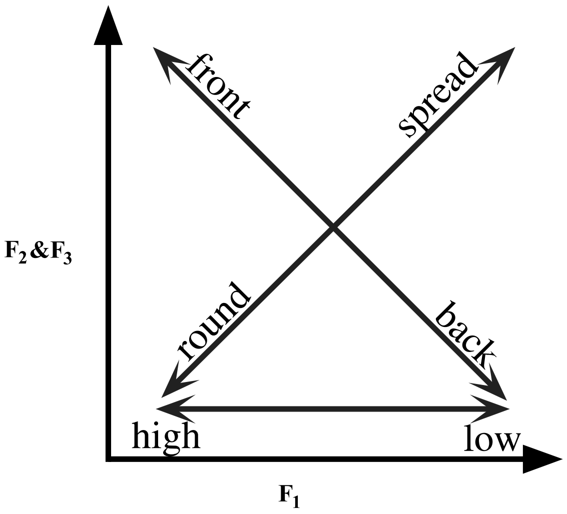

Fig. 1. Acoustic (formant) correlates of articulatory features for vowels.

Table 1 summarizes the relationships between articulation and acoustic features (Ladefoged, 1975). The described acoustic features are intended only as rough guides, and the actual acoustic correlates depend on the exact combination of articulatory features.

3. Spectral smoothing algorithms

Four approaches to spectral smoothing are considered here, with details provided for only those methods which show encouraging results.

Table 1
Acoustic correlates of articulatory features

Articulation	Acoustic features
Vowel	formant frequencies typically contained in each of freq. bands: 0–500, 500–1500, 1500–2500 Hz, etc.
Bilabial	F_2 and F_3 comparatively low
Alveolar	F_2 around 1700–1800 Hz
Velar	F_2 usually high; common origin of F_2 and F_3 transitions
Retroflex	general lowering of F_3 and F_4
Stop	sharp beginning of formant structure
Fricative	random noise pattern dependent on point of articulation
Nasal	formant structure similar to vowels with formants around 250, 2500, 3250 Hz; F_2 low amplitude; distinct antiresonance
Lateral	formant structure similar to vowels with formants around 250, 1200, 2400 Hz; higher formants reduced in amplitude
Approximant	formant structure similar to vowels; usually changing
Dental	F_2 around 1600–1800 Hz; F_3 around 2900–3050 Hz

Although several researchers have studied smoothing techniques (e.g., audio morphing (Slaney et al., 1996), HMM-based smoothing (Plumpe et al., 1998)), the field is emerging and typically only common existing speech processing algorithms (e.g., linear prediction techniques) are employed. Several of these processing techniques were originally developed for other purposes, including interpolation for audio coding and voice transformation, and in general are not typically applied for spectral smoothing in concatenative synthesis. Here we focus only on their application to spectral smoothing.

One approach to smoothing is to interpolate between boundaries of adjoining segments. Therefore, these anchor frames should be good representatives of the sound. The approach taken here is to perform linear interpolation in different domains between the two anchor 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 in spectral smoothing is to determine for which circumstances smoothing should be performed. If two segments have a sufficiently close spectral match, then distortion introduced by smoothing may negate the performance gain. Moreover, many smoothing techniques are inappropriate for use with unvoiced speech.

Another issue is to determine 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 have shown that three to five periods generally works well; however, further study is needed to determine the proper number of pitch periods for different circumstances.

The remainder of this section describes the smoothing algorithms in detail. We focus on LP algorithms since they are commonly used and can yield good results. We also devote special attention to the continuity effect since it is a new approach for smoothing. In addition, we mention other spectral smoothing algorithms to complete the discussion. The smoothing algorithms we examine are (i) optimal segment coupling, (ii) waveform interpolation, (iii) LP techniques (pole shifting and LSF interpolation), and (iv) the continuity effect.

In the following four sections, we illustrate examples of spectral smoothing using speech spectrograms at the end of each section. Fig. 18 summarizes all sample speech spectrograms for the phrase “carry an oily rag”. Fig. 18(a) represents the phrase produced naturally by a male speaker. Fig. 18(b) reflects the results from segment synthesis with no spectral smoothing from a codebook with a nominal size of 380 segments.

3.1. Optimal coupling

It is standard in concatenative synthesis that the boundaries of speech segments be fixed, but the optimal coupling technique allows the segment boundaries to move in order to improve the spectral match between adjacent segments (Conkie and Isard, 1997). At its simplest, the optimal coupling technique is rather straightforward. During synthesis, each segment's boundary for concatentation is chosen in order to fit best with the adjacent segments in the synthesized utterance. An objective measure of spectral mismatch is used to determine the level of spectral fit between segments at various possible boundaries. The measure of spectral mismatch is tested at a number of possible segment boundaries, and the minimum measure score indicates the location of the closest spectral match.

If two segments are to be concatenated, where the end frame of the first segment is in the range $x_i, \ldots, x_\mathrm{f}$, and the start frame of the second segment is in the range $y_i \ldots y_\mathrm{f}$, then the distance measure function $d()$ is evaluated at all possible boundary positions to find $\min_{a,b} d(x_a, y_b)$. For concatenation of this segment pair, the boundary frames x_a and y_b of the segments are selected such that the measured mismatch between frames is minimal. Fig. 2 shows an example scenario where moving the segment boundaries will noticeably change the spectral alignment of formants.

While any form of measure may be used to determine the amount of mismatch, for the sake of improving spectral quality, using a spectral discontinuity measure is appropriate. Measures considered here include the difference of mel-frequency cepstral coefficients (MFCC) and the auditory-neural based measure (ANBM) (Hansen and Chappell, 1998).

In simple frame mismatch, distance measures are calculated for frames ending at various possible segment boundaries. The distance measures take into account only the single audio frame from each speech segment which lies next to the boundary under consideration. More advanced variations of optimal coupling also consider the direction of spectral parameter motion. Although our studies used only a simple frame mismatch, more complex proposals include use of regression coefficients and linear coefficient fit (Conkie and Isard, 1997).

There are a number of advantages to the optimal coupling technique. The algorithm is conceptually simple and easy to implement. It can be combined with other spectral smoothing techniques and need not stand alone. Optimal coupling can successfully complement other smoothing techniques because it causes formants to be naturally closer to each other at segment boundaries. Since coupling does not modify the existing speech, it does not introduce additional artifacts. For spectral matching purposes, it effectively expands the number of speech segments in the database.

Despite these advantages, there are also several disadvantages to optimal coupling. Finding the optimal coupling point requires a search for each segment joint, and an exhaustive search is required for full optimal coupling. Moving the segment boundaries carries the risk of accidentally cutting an important part of a sound or adding an inappropriate sound. Errors often occur by leaving out too much of a sound, though heuristic rules can

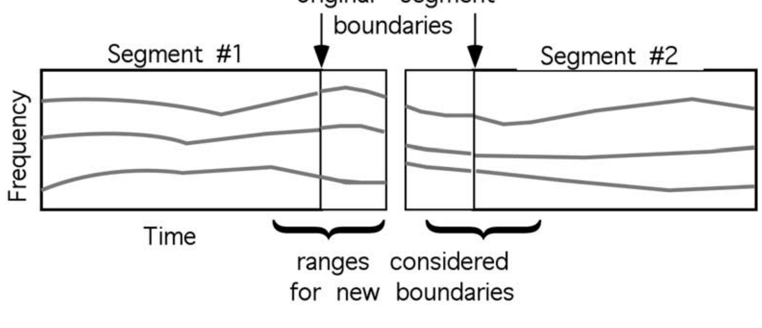

Fig. 2. Optimal segment coupling.

reduce or eliminate this effect. More importantly, optimal coupling is limited in its effectiveness since it only works with existing speech and does not actually modify formant positions.

In summary, optimal segment coupling is a relatively simple approach for a reasonable benefit. It is easy to use either by itself or in combination with a more direct smoothing algorithm. In some forms of concatenative synthesis, segments are specifically designed or pre-selected from a database such that formants match smoothly at their edges, and in these cases optimal coupling will provide little if any gain. In comparison, optimal coupling clearly has no application to spectral smoothing for speech coding since the original signal is already naturally smooth.

Fig. 18(c) shows a spectrogram of the phrase "carry an oily rag" with optimal coupling between segments. In Section 5.4 we discuss this spectrogram and compare it with similar spectrograms resulting from other spectral smoothing algorithms.

3.2. Waveform interpolation

Waveform interpolation (WI) is a speech-coding technique which takes advantage of the gradual evolution of the shape of pitch-period waveforms. The WI coder operates on a frame-by-frame basis. In each segment, the pitch track is calculated and *characteristic waveforms* are extracted. Each characteristic waveform is typically one pitch period long, but the length may be an integer number of periods. In coding, characteristic waveforms are extracted from the original signal at regular time intervals. In order to conserve space in coding, a WI-coded signal is typically transmitted as quantized frequency coefficients for separate rapidly and slowly evolving components. On reconstruction, intermediate pitch-cycle waveforms between transmitted waveforms are approximated by interpolation. To produce an interpolated waveform, both the pitch period and waveform signal are interpolated in either the time domain (at a common period of 2π radians) or the frequency domain (Kleijn et al., 1996; Kleijn and Haagen, 1995). WI is essentially a form of smoothing intended for speech and audio coding.

Though developed for coding purposes, WI can also be adapted for use in spectral smoothing. In this case, the waveform is interpolated between frames at the edges of speech segments to create new inserted smoothed data. 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. Such a scheme has been used on frames with constant pitch in MBROLA synthesis (Dutoit and Leich, 1993).

Fig. 3 shows an example of WI with two natural frames of speech (/aa/ and /ae/) and one interpolated frame. In addition, a performance example is shown in Fig. 18(d), which shows an example spectrogram of a phrase with smoothing performed via waveform interpolation.

We conclude that WI is generally better than no smoothing, but has difficulty producing consistent results. In the simplified version, WI is conceptually simple, computationally fast, and easy to implement in the time domain. When spectral envelopes are similar, WI can give good results. However, it does not perform actual formant smoothing and thus yields only small improvements. WI generally produces smoother results for a large number of interpolated pitch periods, and works best on vowel-to-vowel transitions. Although the results often sound smoother than with no interpolation, there are artifacts, and the general quality is not as good as smoothing techniques that directly manipulate formant positions.

In addition to direct use for calculating smoothed speech frames, WI can also be applied for residual interpolation in the linear prediction (LP) methods (Kleijn and Haagen, 1995). LP methods 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.

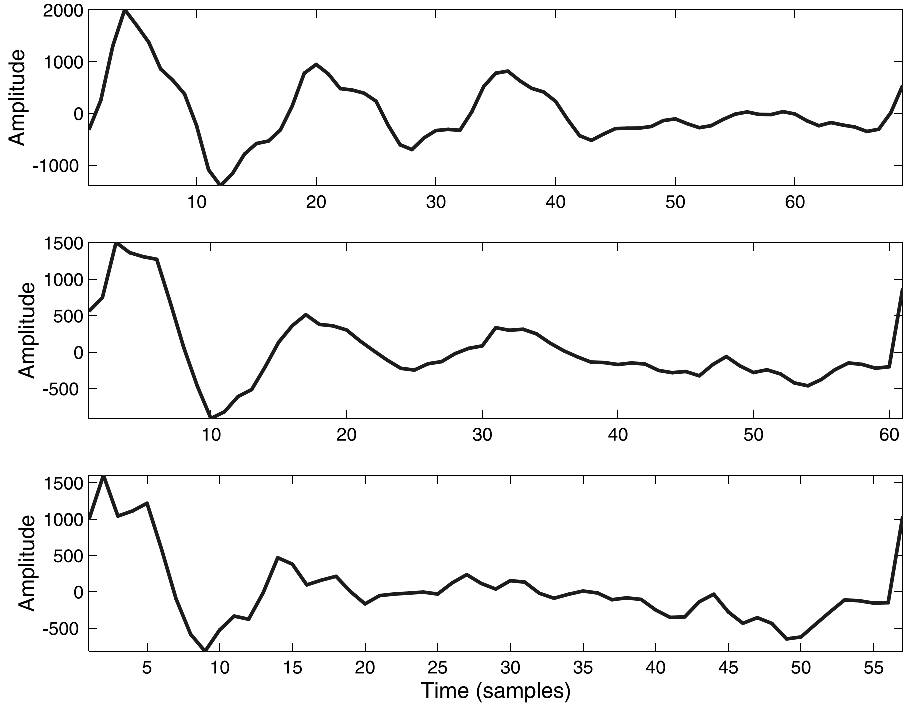

Fig. 3. Example of waveform interpolation for a single inserted frame from /aa/ (anchor Frame 1 on top) to /ae/ (anchor Frame 2 on bottom).

Fig. 4 illustrates the application of WI to the residual in LP smoothing. This use of WI with LP follows a recent trend in speech synthesis towards mixing deterministic and noise components.

We performed evaluations using WI with LP on regions of speech. WI generated the residual to go along with the original LP spectrum. The resynthesized speech was in some cases practically indistinguishable from the original, while there were usually some small artifacts. With longer interpolated regions, the level of noticeable distortion was greater. When the interpolated region differs from the speech on either side (e.g., near silence or a short phone in the interpolated region), then the distortion is more noticeable. As expected, the frames differ more from the original natural speech in the center of the interpolated region. The most common distortion was that the timing of pitch pulses was off from the original, giving a somewhat artificial quality to the speech. It is believed that this artifact is due at least in part to using a constant-length frame size for this test, and that a

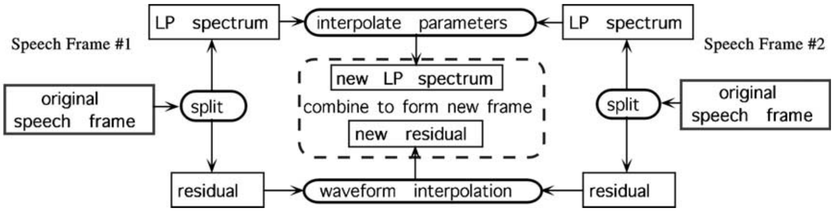

Fig. 4. Waveform interpolation applied to residual of LP coding.

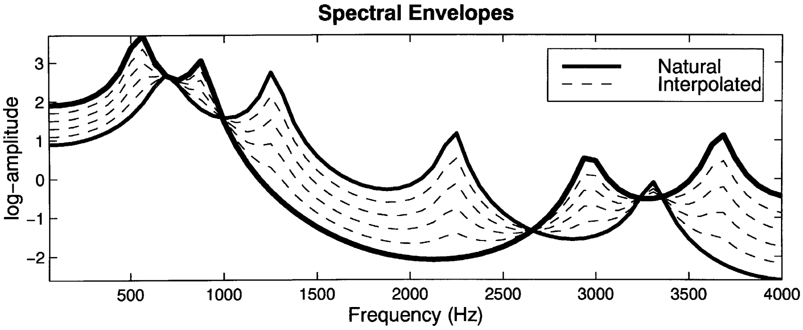

Fig. 5. Example of directly interpolating LP spectra.

pitch-period-based frame size would improve the resulting speech quality. Despite occasional distortion, the interpolated region was generally acceptable. Thus, WI is an appropriate way to generate the residual for LP interpolation techniques.

3.3. LP techniques

LP interpolation techniques are often used to smooth LP-filter coefficients in LP coding (LPC) and sometimes also for speech synthesis (Shadle and Atal, 1979). The basic strategy is to model the speech signal as separate spectral and residual (filter and source) components and to adjust each component separately. Here, we perform LP spectral parameter interpolation in one of several domains, while the residual is interpolated using WI.

If the LP spectra are directly interpolated, formants will rise and fall in the interpolated frames of speech rather than move smoothly in frequency, amplitude, and bandwidth as is desired. Fig. 5 shows an example of improper results from interpolating LP spectra (two anchors with four interpolated frames); for comparison with a more sophisticated interpolation algorithm, the anchor-frame spectra are the same as for Fig. 6. Thus, rather than LP spectra interpolation, we strongly recommend interpolating LP parameters in a domain where the parameters are closely linked to formant location. To perform spectral smoothing, the LP parameters should be interpolated and recombined with a separately-interpolated residual.

LPC analysis yields less frame-to-frame variation and smoother evolution of the coefficients when analysis is performed on pitch-synchronous windows (Paliwal and Kleijn, 1995). Thus, it works well in PSOLA-based systems.

LP interpolation has a number of advantages but also some disadvantages. Most importantly, LP methods allow direct manipulation of the spectral envelope and thereby indirect manipulation of formants in a way desirable for smoothing. On the downside, it is difficult to determine which parameters control which formants and how to match parameters between frames. Also, in some domains, the inherent ordering of parameters does not give the best matching of parameters. Additional algorithm processing is required to translate the signal into the desired LP domain and back unless the data is already LP-encoded. Since LP analysis is based on an all-pole model, it does not adequately model nasal consonants and nasalized vowels. Despite these limitations, we have found that LP interpolation techniques can provide good results which exceed those of the other algorithms tested in this study.

LP interpolation was examined in two different domains. The following two subsections give details on these approaches and discuss their individual strengths and weaknesses.

3.3.1. LP pole shifting

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. Nonetheless, LP poles are a useful rep-

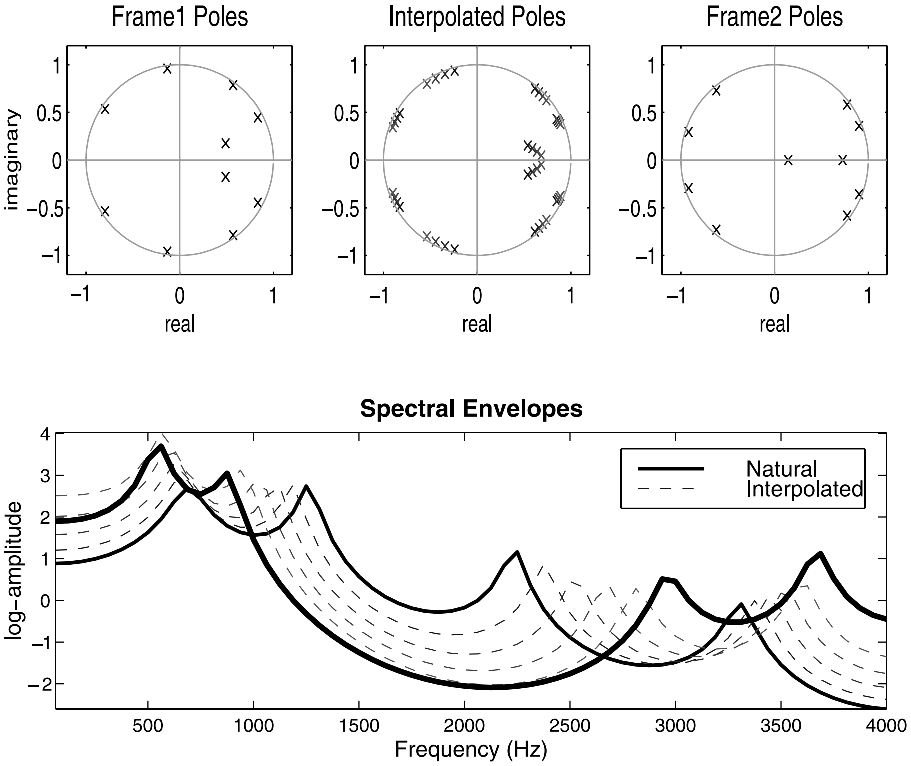

Fig. 6. Example of a successful LP pole shifting scenario.

resentation for interpolation. The two major problems involved with pole shifting are (1) matching the poles between the two anchor frames and (2) performing the interpolation.

Pole matching is not a simple problem, and it is related to the problem of finding formants based upon LP poles (Snell and Milinazzo, 1993). For coding natural speech, poles are unlikely to move very far between known values, and thus the matching problem is easier than for concatenative synthesis where poles may be at widely different positions in adjacent segments. The obvious solution of simply aligning poles by frequency order fails to consider several situations, such as the presence of real poles or other poles that do not directly correspond to formants but instead contribute to overall spectral shape. Moreover, there may be cases where formants split or merge, or arise in such a way that there is not a one-to-one correspondence between poles in different frames. A good approach is to search for the optimal pole match using an appropriate distance measure formula such as the following (Goncharoff and Kaine-Krolak, 1995):

$$D(p_0, p_1) = \begin{cases} \left|\ln\left(\frac{p_1}{p_0}\right)\right| \left\{\frac{\ln((1-r_0^2)/(1-r_1^2))}{\ln(r_1/r_0)}\right\}, & r_0 \neq r_1, \\ \left|\ln\left(\frac{p_1}{p_0}\right)\right| \{2r^2/1 - r^2\}, & r = r_0 = r_1, \end{cases} \tag{2}$$

where p_i are complex pole positions and r_i are pole radii. Our experiments have shown that this distance formula gives insufficient weight to the radius and thus may match a formant-connected pole with an overall spectral shaping pole nearby in frequency. This distance measure has this weakness because it was derived from equations for frequency and bandwidth which are based on a single pole rather than a multi-pole system. An improved distance measure could lead to better automatic pole matching and thereby better spectral smoothing.

In pole matching, a common problem arises when one frame of speech has more real poles than the adjoining speech segment frame. Fig. 6 illustrates this scenario, where four frames of speech are interpolated between Frames 1 and 2 pole plots. One solution to the pole-matching problem is to convert the pole constellation to a domain where each pole has a complex conjugate and then use a distance measure to match poles (Goncharoff and Kaine-Krolak, 1995). Another approach is to first match conjugate pairs that result in the minimum overall distance between matched pairs. For each remaining unmatched conjugate pair, the nearest single real pole is selected as a match.

Fig. 7 shows an important pole-matching scenario where improper matching yields poor results. Whether the poles are matched between Frames 1 and 2 by frequency ordering or by using Eq. (2), the first two poles become inappropriately criss-crossed over the four interpolated frames. With user assistance, the proper pole match could easily be made, but both automatic algorithms fail. As a result, the movement between Frames 1 and 2 is smooth for F_2 and F_3 and the overall spectral slope, but F_1 (ca 550 Hz) suddenly drops in amplitude for the interpolated frames and then suddenly rises in amplitude.

Once poles have been appropriately matched between anchor frames, their positions must be interpolated. This should not be performed directly in the complex plane, but instead the magnitude and phase of the poles should be interpolated separately. Separate interpolation of real and imaginary components in the z-plane can produce values which are not truly intermediate, but interpolating in the magnitude-phase domain produces more reasonable results. It is known that the magnitude of a pole relates to formant bandwidth, while the angle relates to formant frequency. While the pole radii can be interpolated

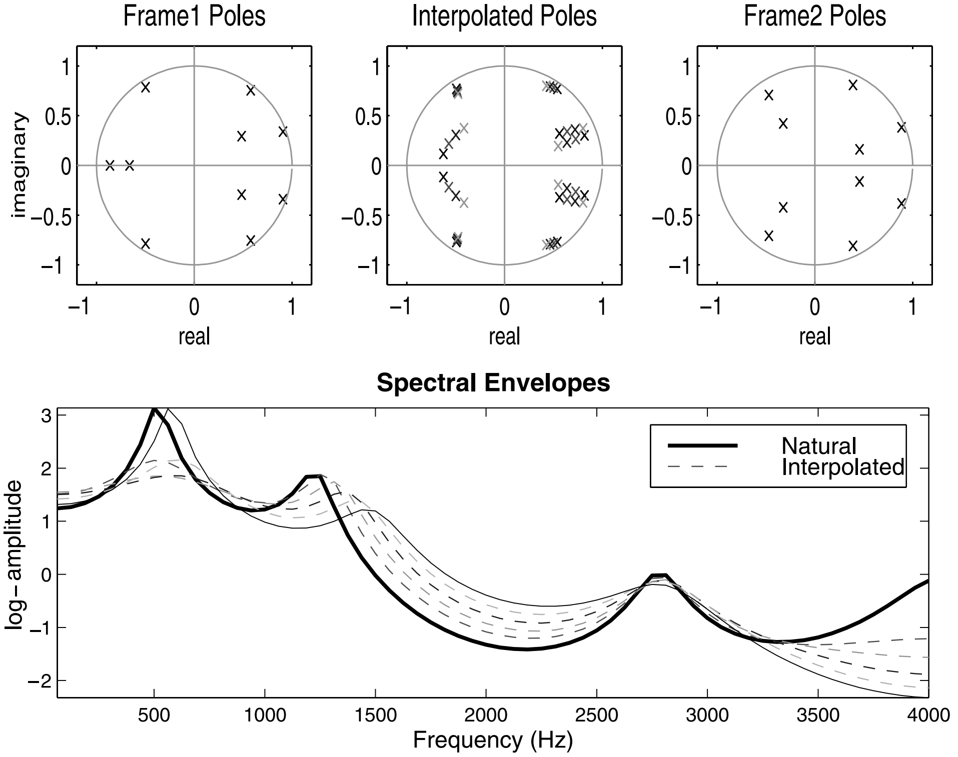

Fig. 7. Example of a failed LP pole shifting scenario from Frame 1 to Frame 2 with four interpolated frames.

directly, bandwidth can be directly interpolated by using the standard formulae that relate frequency F and bandwidth BW of a single pole p_i to its angle θ_i and radius r_i and the sampling period T_s,

$$F_i = \frac{\theta_i}{2\pi T_s}, \quad BW_i = \frac{-\ln(r_i)}{\pi T_s}. \tag{3}$$

Thus, to linearly interpolate with the frequency and the bandwidth, the angle should be interpolated linearly while the radius is interpolated geometrically. Specifically, if the new pole $p_i = r_i \angle \theta_i$ is a fraction k_1 from a known pole $p_1 = r_1 \angle \theta_1$, and a fraction k_2 from pole $p_2 = r_2 \angle \theta_2$, where $k_1 + k_2 = 1$, then the components of p_i should be generated by the equations

$$\theta_i = k_1\theta_1 + k_2\theta_2 \quad \text{and} \quad r_i = r_1^{k_1} + r_2^{k_2}. \tag{4}$$

Ideally, each LP pole pair would correspond to a single formant, but in practice multiple poles will affect the location and bandwidth of each formant and some poles will contribute to overall spectral shape. Thus, although pole shifting does modify formants, it can have undesired effects such as formant bandwidth spreading. Quite often, the LP model order is selected with the notion that smoothing will be applied (i.e., for $f_s = 8$ kHz, studies will select an order of $P = 9$, corresponding to approximately four formant pole-pairs and one real overall shaping pole). Other research on waveform synthesis has been successful in separating individual formants based on poles for formant frequency modification (Mizuno et al., 1993). Key points of that work are that (1) formants are connected with only those poles with a low value of BW_i divided by F_i and (2) pole extraction is checked by comparing the target and calculated spectral intentions in an iterative procedure.

Fig. 18(e) shows an example spectrogram of a segment concatenated phrase with smoothing performed via LP pole shifting.

In summary, LP pole manipulation has excellent spectral smoothing potential, yet several weaknesses are present. Shifting pole location gives the ability to shape the desired interpolated speech spectral structure. When each pole corresponds to a formant and the poles move little between anchor frames, then the interpolation is simple and of high quality. In more complex situations, the relationship between pole location and spectral envelope must be considered to ensure that pole matching and interpolation gives the desired results. The results can be quite good, but even more recent techniques are not sufficient to be applied in a completely unsupervised manner. In a minority of cases, pole interpolation can yield results which are worse than no smoothing. Future efforts should consider ways to automatically assess the success of the interpolated pole shifted frames.

3.3.2. LSF interpolation

The line spectral frequency (LSF) representation, also known as line spectrum pair (LSP), is often used for speech coding (Papamichalis, 1987). Interpolation between LSFs has been used not only for coding but also for synthesis and even spectral smoothing. LSFs are calculated from the LP poles in a technique that yields two sets of interleaved zeros on the unit circle. Representing the LPC filter in the LSF domain ensures its stability and is thus appropriate for coding and interpolation.

For coding, the LSF representation is generally accepted as giving the best performance in terms of spectral distortion, and it always yields stable filters after interpolation (Paliwal and Kleijn, 1995; Paliwal, 1995). Some comparative studies have shown that LSF interpolation gives better results than other representations when used for interpolation in coding as measured by spectral distortion (Paliwal and Kleijn, 1995; Paliwal, 1995) or prediction error (Erkelens and Broersen, 1994). Other comparison studies showed no inherent advantage for LSFs (Atal et al., 1989).

LSFs can also be interpolated for speech synthesis. For waveform synthesis, however, there is limited benefit from the compression and quantization advantages which LSFs offer for coding.

The two major problems of pole shifting are trivial for LSF interpolation. Unlike pole shifting, LSF interpolation provides an inherent order for parameter matching. When LSF pairs are matched in the obvious sequence of increasing frequency,

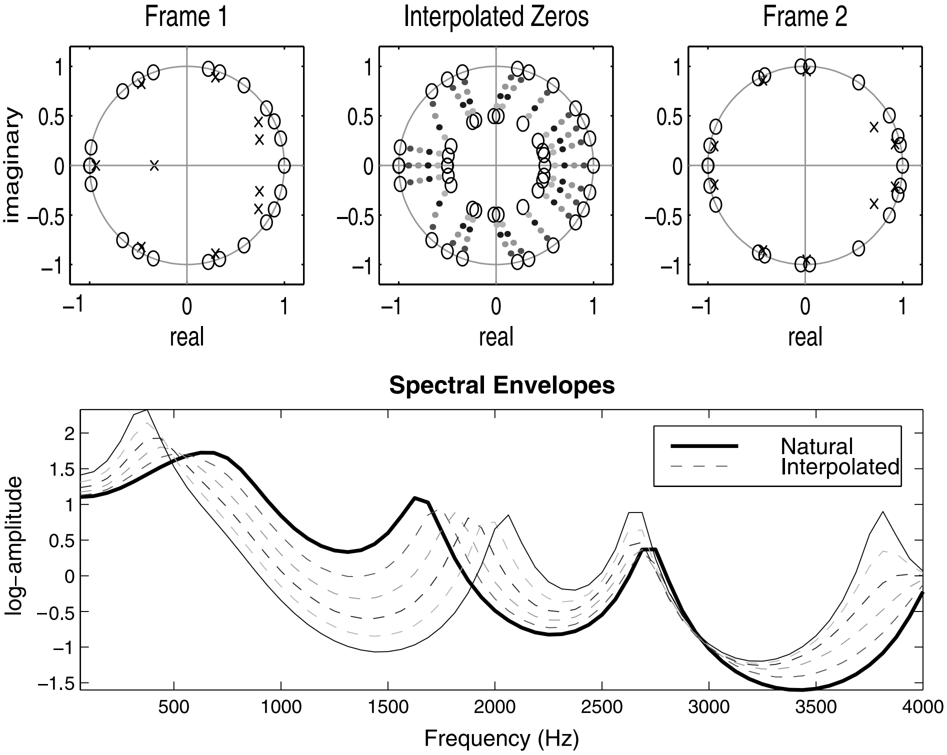

Fig. 8. Example successful LSF interpolation scenario. Note that the two circles for interpolated zeros are used only for imaging purposes; all zeros for all frames are actually on the unit circle.

however, this sequence is not always in the order which yields the best results. As with pole matching, there are cases where a parameter that corresponds to a formant will be matched with a parameter that corresponds to general spectral slope. The interpolation process is also straightforward since there is only one dimension (frequency) involved. The two major interpolation methods are to either interpolate each P and Q zero separately or to first interpolate the P (position) zeros and then interpolate the difference parameters. Since the position parameters correspond to formant position while difference parameters roughly correspond to bandwidth, the latter approach is more intuitive. Specifically, if a new zero pair P_i, Q_i is a fraction k_1 from known zero pair P_1, Q_1 and fraction k_2 from zero pair P_2, Q_2, where $k_1 + k_2 = 1$, then this form of interpolation yields $P_i = k_1 P_1 + k_2 P_2$ and $Q_i = P_i + k_1(Q_1 - P_1) + k_2(Q_2 - P_2)$.

Fig. 8 shows a scenario where LSF interpolation succeeds, while Fig. 9 shows an example where it performs poorly. In each figure, the z-plane plots for anchor Frames 1 and 2 show both the LP poles and the LSF zeros, while the plot of interpolated zeros shows the transition from the first frame zeros (outermost ring) to the last frame zeros (innermost ring). In Fig. 8, all formant peaks have moved smoothly across frequency as desired on the spectral envelope plot. In Fig. 9, the formants located near 800, 1400, and 1800 Hz do not move in frequency as desired but instead shift only in amplitude. As another example of performance, Fig. 18(f) shows a spectrogram of a phrase with smoothing performed via LSF interpolation.

Despite some obvious strengths, the use of LSFs for interpolation can also display some inherent drawbacks. The interpolation technique itself is simple: the zeros have an inherent order for matching, and the interpolation is in a single di-

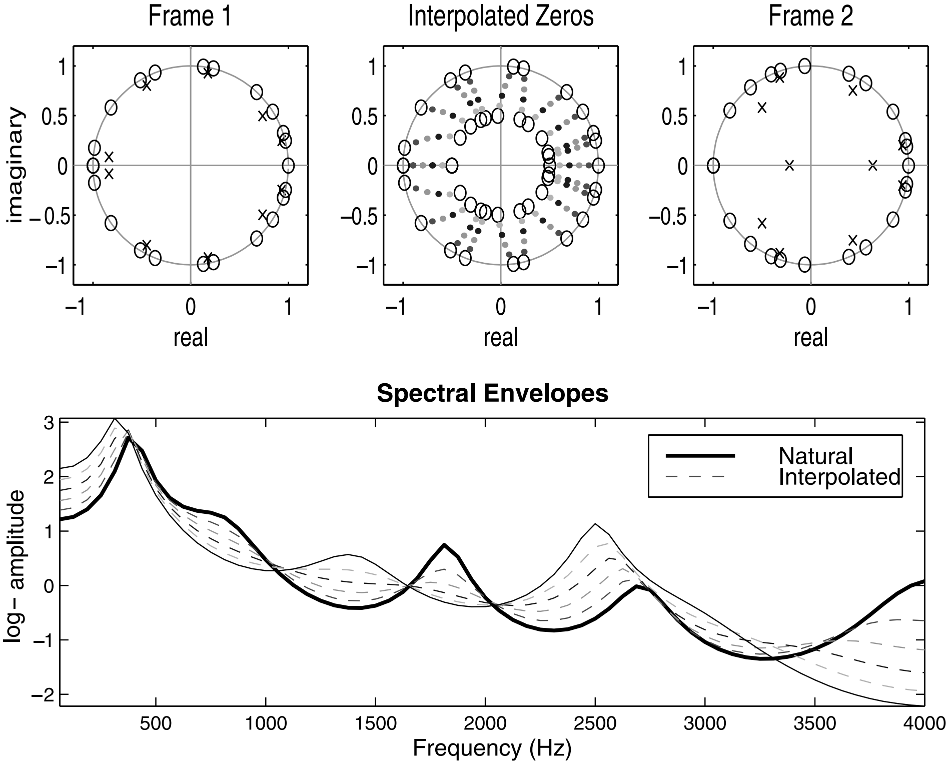

Fig. 9. Example failed LSF interpolation scenario. Note that the two circles for interpolated zeros are used only for imaging purposes; all zeros for all frames are actually on the unit circle.

mension. The zeros' inherent order does not always match between frames in the manner that could yield the best results after interpolation. More importantly, little can be done to improve the basic algorithm. As with LP pole manipulation, there are a small but noticeable number of scenarios in which LSF interpolation yields results which are worse than no smoothing. Although LSF interpolation is simple and can give good results, it does not hold the potential to be universally successful for direct LP pole interpolation.

3.4. Continuity effect

The fourth smoothing approach does not perform audio signal interpolation but instead masks discontinuities. The continuity effect is a psychoacoustic phenomenon that is suggested here as a possible method for spectral smoothing. 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" (Bregman, 1990; Moore, 1997). Fig. 10 illustrates the phenomenon.

Perceptual closure occurs when a missing sound gap 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 either that the masking sound contain the frequency content of the missing, theoretically masked sound or that the masking sound be near enough in frequency or time to the missing sound for simultaneous masking to occur

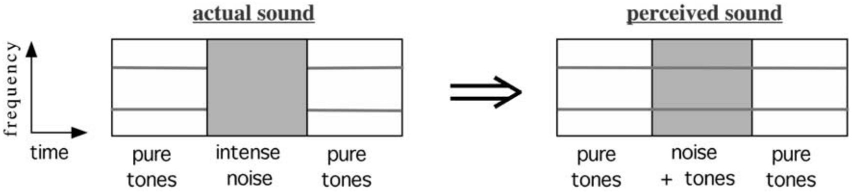

Fig. 10. Illustration of the continuity effect.

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 irregularly spaced bursts of noise interrupting speech at the rate used in phone or diphone concatenation (about 6 per second) is near a minimum in the effect of noise on speech comprehension. Moreover, with this interruption frequency and the desired fraction of time spent on speech versus noise (~91%), listener tests revealed a very high word articulation rate. In some circumstances, interrupting noise has been shown to actually increase intelligibility (Bregman, 1990; Moore, 1997). Similar perceptual studies have found that replacing a phone with an extraneous sound results in listeners reporting the presence of the phone, while replacing the phone with silence results in correct detection of the gap (Warren, 1970).

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 extend the concept 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.

Fig. 11 shows an example of how a frequency-domain filter is obtained for inserted noise. The spectral envelopes of the two original speech frames are compared, and the filter is constructed to meet the maximum of the two envelopes at all points and to interpolate between any peaks (presumably formants) between the two spectra. To generate this spectral envelope for the noise, all peaks are found and sorted by frequency for the spectral envelopes for both of the original frames. For each frequency range between two spectral peaks, the anchor-frame envelopes are compared as follows. If the amplitude of one of the original envelopes is larger than the other at all frequencies in the range between two peaks, then that portion is directly used in the new envelope. Otherwise, that frequency range will have amplitude values constructed by interpolating between the two

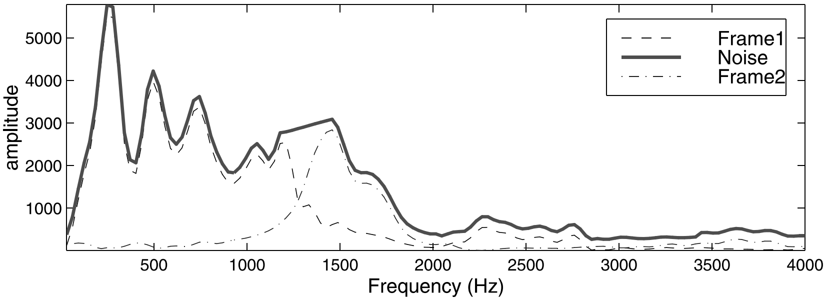

Fig. 11. Example noise envelopes for continuity effect.

peaks. Once the new spectral envelope is constructed, Gaussian white noise is passed through the filter to create shaped noise that will mask any hypothetical speech between the two natural frames without introducing more noise than necessary for auditory masking.

Although the proposed use of the continuity effect does not provide great spectral smoothing in all situations, we have found that it performs well in some cases and has distinct potential for improvement. From our experiments, a section of inserted shaped noise about 45–125 ms long generally yields the best results. A section of noise that is too long gives the impression of either inserting a stop or simply playing noise. In most cases, it is clear that some noise is present in the signal, but it is also feasible that the speech continues smoothly in the background. In other cases, the shaped noise sounds worse than nothing, but in rare cases it sounds very good and natural. Our experiments revealed an optimal amplitude of the noise at about 1/4 the amplitude of the rms of the mean of the adjacent natural frames. A lower noise amplitude sounds more natural and acceptable, but it can also sound more often like an inappropriately inserted stop. Fig. 18(g) shows a sample spectrogram.

Inserting shaped noise is noticeably better than white noise; it sounds more like the noisy signal of the correct spectrum. When smoothing is performed between concatenated speech segments, the use of closure may fail when the formants of the two segments are naturally too far apart. When noise is played intermittently with natural speech, the formants will typically be close enough together for the continuity effect to apply, but such situations do not always occur with concatenation. Still, many concatenative databases have segments selected such that their formants are nearly aligned. In summary, using shaped noise can provide perceptive spectral smoothing in some cases, but in other cases it can be very annoying. We have shown that application of the method is promising, especially for phonemes with frication where LP and WI techniques fail. Further research is warranted in determining the perceptually optimal spectral envelope for shaping the inserted noise.

4. Determining smoothing required

Rather than blindly applying the same spectral smoothing algorithm in the same manner at all concatenation points, we suggest that several methods may be necessary since it is important to determine the proper type and amount of spectral smoothing required. Not all segment joints benefit from spectral smoothing, and no single smoothing algorithm performs best in all situations. Relevant factors – including phonemes, articulator positioning, and spectral perception – help determine the type, amount, and duration of modification required for smoothing. It is noted that we have not emphasized F_0 issues since standard methods such as PSOLA can modify the pitch to the desired frequency.

In this study we used knowledge from both speech science and signal processing to determine the smoothing required in different scenarios. We have compiled this data into a table (see Section 4.2 and Table 3) for use as an aid in smoothing. In addition, we provide further details on smoothing for several example phone pairs (see Section 4.3). Although we performed this analysis on phone segments, the same concepts are applicable for diphone concatenation.

One important issue of spectral smoothing is determining the circumstances under which the smoothing should be performed. If two segments have a sufficiently close spectral match, then the distortion introduced by smoothing techniques may sometimes outweigh the performance gain. On the other hand, spectral smoothing generally performs better on segments with similar spectral characteristics, and attempting to smooth very different spectral envelopes can yield poor results. Moreover, many smoothing techniques are inappropriate for use with unvoiced speech. The two pieces of data used in automatically determining whether smoothing is appropriate for a joint are (1) knowledge of the phonemes involved and (2) the ANBM score (see Section 2.3) for the joint.

Certain smoothing algorithms are better for certain phonemes. For example, LP techniques are not as successful for nasals and nasalizations because they employ all-pole models and thus do not reflect the anti-resonances.

Another issue is determining the best time span over which to interpolate. The pitch will remain continuous if data is inserted in blocks equal to an integer number of pitch periods. Many of our experiments have used a pitch-synchronous synthesizer, and we have seen that inserting three to five interpolated periods generally works well. While we have considered experiments which range between 0 and 10 pitch periods for insertion, future studies should be done to determine the optimal number of frames (i.e., pitch periods) of smoothing for specific phone-to-phone circumstances.

4.1. Method

To determine the desired spectral smoothing, we consider aspects of both speech science and signal processing. The articulation between phonemes gives an indication of the expected formant transitions. Analysis of the ANBM scores for natural and concatenated phones indicates the approximate scores for perceptively smooth speech. Knowledge of the phonemes and smoothing algorithms leads to recommendations as to which algorithms should be used for various circumstances.

We propose using the ANBM to evaluate the perceived spectral smoothness between segments. In order to test the ANBM across phone boundaries in natural speech, we applied the measure to phone transitions from the TIMIT database [1] resampled at 8 kHz. Using the phoneme labels supplied with TIMIT, the ANBM score was calculated for each phone-to-phone transition in the database. We recorded the measure scores and calculated relevant statistics such as the sample mean and the unbiased sample standard deviation. The resulting ANBM scores are used to assess phone-to-phone spectral transition information.

Table 2 shows the net ANBM results and statistics across all phone-to-phone transitions. It reports the sample mean and sample standard deviation for the entire data set. Measure analysis was performed on only male speakers in the training portion of the database. Fig. 12 shows a histogram of the resulting ANBM scores at phone transitions for the entire database.

Table 2
Net ANBM results for TIMIT

Number of speakers	326
Number of phone transitions	124,193
Sample mean	222.24
Sample standard deviation	100.27
Maximum score	732
Minimum score	6

Analyses with the ANBM have shown that measure scores vary according to several factors: naturalness, the specific phonemes concatenated, the speaker, and the dialect. Subjectively, different speakers yield different amounts of smoothness when their phonemes are concatenated, and statistical analysis of ANBM scores confirms this observation. Calculating the probability that the population mean difference d is at least 0 (PdAL0) (Berry, 1996) indicates that several of the dialect regions have scores which are statistically significantly different from each other even at the 99% probability interval. Although there is no single threshold which is equally suited to all speakers, we have included data from speakers of different American English dialect regions in determining approximate thresholds for general use.

Fig. 13 shows the sample mean and sample standard deviation for those phoneme-pair junctions with 400 or more examples in the database. The first entry shows the sample mean and sample standard deviation for the overall dataset, and individual transition results follow. This figure shows that different phoneme-pairs produce shifted expected measure score ranges.

In order to use ANBM scores to determine whether smoothing is necessary, we have used a probe listener test to connect ANBM scores with subjective evaluations. For each possible phoneme class junction, we generated approximately 100 concatenated examples from TIMIT. We then used a nine-value subjective opinion score to measure the perceived spectral smoothness of both these concatenated examples and a small set of natural examples. Based on the subjective scores, we suggest using an ANBM threshold to deter-

[1] Available from the Linguistic Data Consortium at http://www.ldc.upenn.edu/.

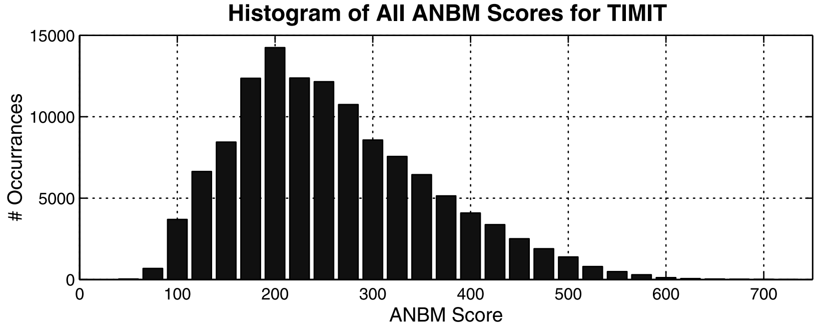

Fig. 12. Histogram of ANBM Scores for TIMIT.

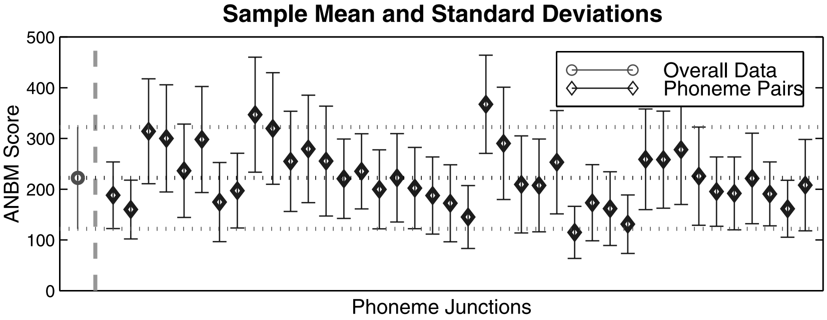

Fig. 13. Sample mean and standard deviation of ANBM score for overall data set and for different phoneme pair junctions.

mine when a concatenated joint is acceptably smooth. Joints with ANBM scores below the threshold will be considered acceptable, while scores above the threshold will indicate a need for spectral smoothing. We compared ANBM scores and opinion scores to establish relative correlation measures. The correlation coefficients for different phoneme class pairs varied with better values for more strongly voiced data: the correlation coefficient for combinations of vowels (VL) is 0.21; for combinations of VL and/or diphthongs (DT) it is 0.17; and for VL, DT, and semi-vowels (LG; liquids and glides) it is 0.09.

Based on our evaluations and observations, we recommend a threshold of the sample mean of the natural ANBM scores for a given phoneme class pair. This is not a clear-cut threshold, but raising or lowering it can change the fraction of joints that are smoothed. Fig. 14 shows ROC curves that illustrate the trade-off on the probabilities of detection (P_d) and false alarm (P_f) for joints that need smoothing for different sets of phoneme classes. Curves are shown for junctions involving three sets of phoneme classes: vowels and diphthongs; vowels, diphthongs, and semi-vowels; and all phoneme classes. The detection in this figure means that using the ANBM to find joints that are subjectively marked in the "poor" to "good" range, with "very good" and "excellent" joints being accepted as already sufficiently smooth.

In some practical situations, time constraints may prevent the use of spectral smoothing on all segment joints. It is suggested that the ANBM score should be used to rank-order the perceived smoothness of joints. The ANBM score will list the joints in the same general order that a human listener would rank the perceived smoothness. The spectral smoothing system can then devote more processing resources to smoothing only the top X% of the joints, where X is smaller for larger-database synthesizers. There is a natural balance between resulting speech quality and segment

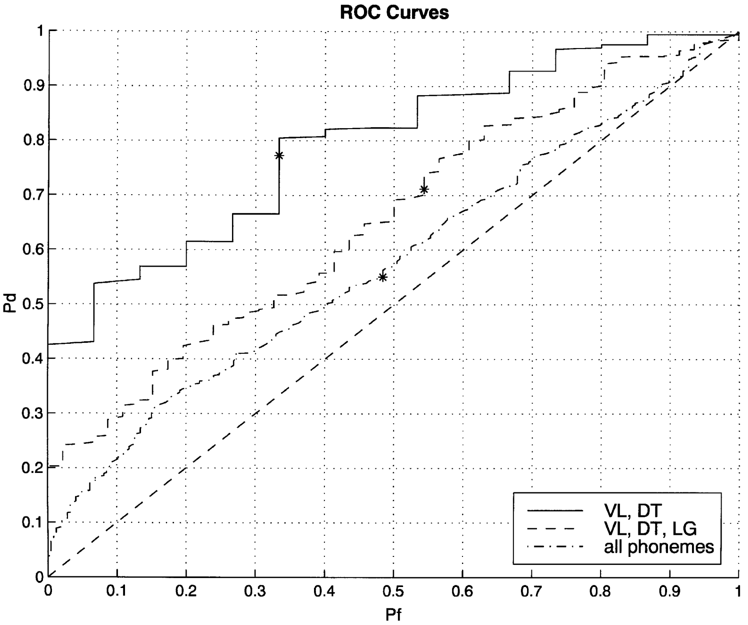

Fig. 14. ROC curves for detection of "very good" and "excellent" joints with the ANBM score for different phoneme class sets. P_d = probability of detection; P_f = probability of false alarm; VL = vowels; DT = diphthongs; LG = semi-vowels; $*$ = results from a threshold of the mean ANBM score.

synthesis database size. It is therefore fair to say that for small database size systems, larger amounts of segment smoothing will be necessary. It also stands to reason that very large database size systems may require little or no spectral smoothing. As the database size increases, so too decreases the fraction of segment joints that require spectral smoothing.

As demonstrated by the measurements we have reported, the ANBM score should not be taken as an absolute scale for determining smoothness. Instead, the score should be considered in the context of typical scores for the given phoneme transition and the given speaker. The relative score gives a better measurement of the perceived continuity.

The just-noticeable difference (JND) or difference limen (DL) of formant frequency, bandwidth, and intensity should also be considered in spectral smoothing. The JND for formant frequencies is 3–14% of the formant frequency value. The formant bandwidth JND is about 20–40%. For formant amplitudes, typical JNDs are approximately 1.5 dB for $F1$ and 3 dB for $F2$ (Flanagan, 1972; O'Shaughnessy, 1990; Rabiner and Juang, 1993). The precise JND depends upon whether the speech is natural or steady-state and whether one or more formants are changing simultaneously.

4.2. Recommendations table

Our experience with concatenating and smoothing various phonemes allows us to make recommendations as to the spectral smoothing algorithms which perform best for each case according to the classes of phonemes joined. The natural smoothness and expected amount of smoothing also follow from given phoneme classes. Although each specific phoneme pair varies slightly, practical space limitations force us to list results by phoneme classes.

Table 3 shows our recommendations for spectral smoothing according to phoneme class. We show results only for those phoneme class pairs that have at least 100 natural examples within

Table 3
Recommendations on spectral smoothing by phoneme class[a]

Phoneme	Nat. ANBM score				Smoothing	
Pair	Min.	Mean	Max.	S.D.	Alg.	Amount
Stop → stop	46	308	662	108	Closure	Large
Stop → nasal	43	253	610	107	Closure	Large
Stop → fricative	38	211	604	98	Closure	Large
Stop → semi-vowel	39	220	673	93	Closure	Large
Stop → whisper	48	181	395	72	Closure	Small
Stop → vowel	24	191	609	78	Closure	Large
Stop → diphthong	46	200	529	88	Closure	Large
Stop → affricate	54	249	565	98	Closure	Small
Nasal → stop	41	258	642	109	Closure	Small
Nasal → nasal	23	181	383	87	LP	Large
Nasal → fricative	36	228	527	90	Closure	Small
Nasal → semi-vowel	18	193	514	95	LP	Large
Nasal → whisper	40	220	510	100	Closure	Small
Nasal → vowel	16	215	604	90	LP	Large
Nasal → diphthong	37	233	531	90	LP	Large
Fricative → stop	52	233	560	81	Closure	Small
Fricative → nasal	64	221	503	80	Closure	Large
Fricative → fricative	48	179	510	73	Closure	Large
Fricative → semi-vowel	56	204	453	77	Closure	Small
Fricative → whisper	63	166	389	63	Closure	Small
Fricative → vowel	30	200	546	76	Closure	Large
Fricative → diphthong	51	209	537	78	Closure	Large
Semi-vowel → stop	51	283	662	109	Closure	Large
Semi-vowel → nasal	44	210	515	98	LP	Small
Semi-vowel → fricative	51	224	567	84	Closure	Small
Semi-vowel → semi-vowel	32	181	546	81	LP	Large
Semi-vowel → vowel	18	178	600	80	LP	Large
Semi-vowel → diphthong	27	194	515	69	LP	Large
Whisper → vowel	45	177	534	83	Closure	Small
Whisper → diphthong	58	179	441	76	Closure	Small
Affricate → stop	88	246	431	75	Closure	Large
Affricate → vowel	63	175	533	55	Closure	Small
Vowel → stop	35	260	661	103	Closure	Small
Vowel → nasal	22	216	647	92	LP	Large
Vowel → fricative	33	198	565	75	Closure	Small
Vowel → semi-vowel	9	172	596	81	LP	Large
Vowel → whisper	40	164	449	67	Closure	Small
Vowel → vowel	6	145	538	66	LP	Large
Vowel → diphthong	32	143	368	68	LP	Large
Diphthong → stop	44	243	608	101	Closure	Small
Diphthong → nasal	37	211	562	90	LP	Large
Diphthong → fricative	47	181	477	71	Closure	Small
Diphthong → semi-vowel	35	182	414	71	LP	Large
Diphthong → vowel	20	128	427	60	LP	Large

[a] ANBM scores are shown for natural joints, and suggestions are given for the algorithm and amount of smoothing to use.

TIMIT. Although some phonemes within each class (and certainly individual phones) may have different results, it is reasonable to generalize to phoneme classes. These recommendations are derived not only from our objective experiments (see Section 5) but also from our subjective experience as reflected in the specific examples shown in Section 4.3.

For each phoneme class pair in the table, we show relevant statistics of the ANBM score calculated for naturally-occurring examples of the joint from TIMIT. We show the range (minimum and maximum), sample mean, and sample standard deviation. These values can be useful in determining the relative smoothness of concatenated joints as well as establishing phone-to-phone class thresholds for directing an appropriate smoothing method.

In addition, we make recommendations as to the type of smoothing appropriate for each phoneme class pair. Of the various algorithms considered in this study, both LP techniques (see Section 3.3) and closure (the continuity effect; see Section 3.4) give results that are broadly useful. Optimal coupling can be used to supplement either algorithm if desired.

Although it is difficult to give a quantitative representation of the extent of smoothing necessary for all situations in a category, we do give an indication of how much smoothing is typically needed for each class joint (i.e., large versus small levels of smoothing). The amount of smoothing specified indicates not only the typical perceived spectral distance between phone segments but also the relative amount of speech data that should be inserted in an interpolation region. When formants lie near each other between segments, then the amount of smoothing is typically small, whereas large discontinuities require a larger amount of smoothing. The amount and duration of smoothing are related since a longer duration is typically needed to smooth a larger discontinuity, but they are not necessarily the same. For example, despite the large discontinuity in formant position between a nasal and a vowel, a short smoothing duration is appropriate since the spectral change results from the rapid motion of the velum.

4.3. Specific examples

This section considers several specific examples of smoothing based on the previously-described table and smoothing algorithms. We examine three phoneme-to-phoneme combinations and consider each smoothing algorithm to conclude which approach is most effective for each scenario. These three scenarios are considered to be examples rather than general categories.

For each example phoneme pair, we examine four or five sample cases from different speakers – three male and one female – selected arbitrarily from TIMIT. We extracted one single sample phone for each phoneme for each speaker, and we concatenated pairs to make continuous speech. We then applied the previously described spectral smoothing algorithms and examined the results. We note that in the following examples, a frame consists of one pitch period of speech data.

Table 4 shows the compiled data for these specific examples. Fig. 15 shows a histogram of the resulting ANBM scores from TIMIT for each of the three phone-transition examples described below.

4.3.1. Vowel-to-vowel joint: /iy/–/aa/

As an example of a vowel-to-vowel transition, we considered /iy/ to /aa/. The phoneme /iy/ is a high front vowel, while /aa/ is a low back vowel. In this transition, articulation is due primarily to the tongue. In /iy/, formants are expected to lie at 270, 2290, 3010 Hz, and in /aa/, they should lie near 730, 1090, 2440 Hz. Thus, the first three formants

Table 4
Specific examples smoothing table

Phoneme	Nat. ANBM Score			Articulation		Smoothing		
Pair	Min.	Mean	Max.	Movement	Formants	Alg.	Amount	Duration
/m/ → /iy/	42	232	463	Lips	F_2	LP	Large	23 ms
/d/ → /ah/	110	214	448	Tongue	F_1, F_2	Closure	Large	38 ms
/iy/ → /aa/	50	144	251	Tongue	F_1, F_2, F_3	LP	Small	30 ms

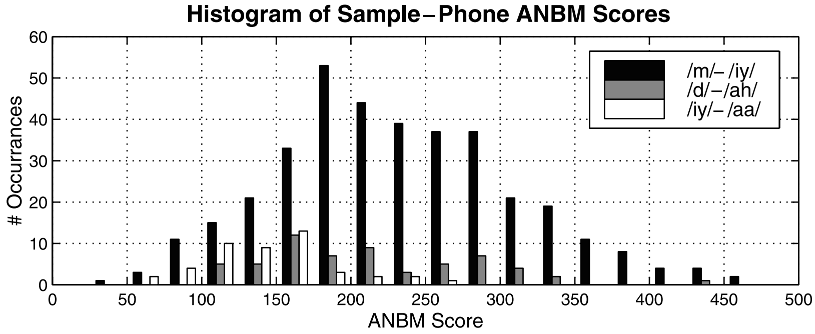

Fig. 15. Histogram of ANBM scores for examples.

are expected to be significantly offset between segments.

Of the smoothing algorithms under consideration, optimal coupling has the least noticeable effect on this phoneme pair. While the measure score improves with coupling, there is minimal perceived difference after coupling and minimal visual difference in the spectrograms.

Waveform interpolation provides some small amount of smoothing. Although formant locations do not actually shift in the interpolated region with WI, the transition sounds smoother for three of the four speakers. With a large number of WI frames, the result sounds noisy, and the best results are when around three or four frames are used.

In comparison, LP pole shifting does yield actual formant movement as desired and thus gives better-sounding smoothing. Results vary according to the order of LP analysis with no one order working best in all cases. The pole-matching problem arises as previously mentioned, and poor matching can yield poor results. Four or five frames of smoothing typically works best, though only one or two frames were appropriate in the one sample where the formants were nearly aligned naturally.

LSF interpolation also moved the formants properly (in three of the four samples) and yielded audibly acceptable results. The optimal interpolation duration varies for each sample from two to seven frames. In two cases, long interpolation regions yielded noisy speech.

The continuity effect can yield feasible results for /iy/–/aa/, but only one of the tested cases gave good results. Generally four to six frames of shaped noise worked best, though the best case – in which the formants were naturally nearly aligned – had the best results for one to two frames. In two cases, the noisy region gave the false illusion of the presence of an extra phoneme.

Thus, we recommend using one of the LP interpolation methods for spectral smoothing of the /iy/–/aa/ transition. A typical good interpolation period is equal to about four pitch periods, or around 30 ms for a typical male speaker. While up to three formants may have to move a fair distance, the LP algorithms can provide appropriate smoothing.

4.3.2. Stop-to-vowel joint: /d/–/ah/

The /d/ to /ah/ transition is an example of a stop-to-vowel phoneme pair. The phoneme /d/ is a voiced alveolar oral stop, while /ah/ is a mid vowel. In this transition, the articulation is primarily with the tongue. Based on the phonemes involved, we expect $F1$ to rise in frequency in /d/ since constriction of front of the oral cavity lowers $F1$. $F2$ and $F3$ should have a slight fall in frequency in /d/. In /ah/, formants are expected to lie at 640, 1190, 2390 Hz.

Optimal coupling gave very little benefit to the smoothness of this joint. In four of the five cases we examined, there was no change at all from coupling, and the fifth case gave only minimal change with no perceived improvement in quality.

Applying waveform interpolation to /d/–/ah/ tends to overly smooth the stop consonant. The /d/ sound can become indistinct or change to be perceived as /b/. Thus, WI should not be applied to this phoneme pair.

The LP interpolation methods can produce appropriate formant transitions as visible on spectrograms. The audible quality, however, is arguably worse than without interpolation. After interpolation, the /d/ phone can lose some of its stop qualities or sound more like /b/.

The continuity effect gives good results for this phoneme pair. In two cases, the results are reasonable, while for two others they produced sufficient improvement so as to make a previously poor concatenation sound like a perfect match.

Therefore, of all the methods examined, the results using the continuity effect clearly outshine the others. A typical good interpolation period is equal to about five pitch periods, or around 38 ms for a typical male speaker. While the stop consonant loses its character under most forms of spectral smoothing, the presence of shaped noise can noticeably increase perceived smoothness and quality.

4.3.3. *Nasal-to-vowel joint: /m/–/iy/*

As a sample nasal-to-vowel pair, we examined /m/ to /iy/. The phoneme /m/ is a voiced bilabial nasal stop, while /iy/ is a high front vowel. In this transition, the articulation is primarily with the lips, though the tongue also plays some role. In /m/, we expect to find F_1 near 250 Hz and F_3 near 2200 Hz; F_2 should be weak, and an antiresonance should be around 750–1250 Hz. In /iy/, formants are generally located around 270, 2290, and 3010 Hz. Thus, formants are not likely to be too far apart between segments.

With several of the sample cases we examined, the formant intensity was so much stronger in the /iy/ phone than in the /m/ that it is difficult to determine spectral smoothness via visual inspection of spectrograms. Still, the most important judgement is auditory perception of smoothness.

Applying optimal coupling to this phoneme pair gives only slight improvement. Though the coupling algorithm yields larger shifts in phone boundary positions for this pair than for other phoneme pairs examined in detail, the results were only marginally better in formant matching and perceived quality. In comparison, waveform interpolation yields only slight improvement for this phoneme pair. None of the test cases gave lower performance with WI, but results ranged from no noticeable change to slight perceived improvement. The transition was sometimes smoother as formants faded in and out.

Both LP interpolation algorithms gave some smoothness improvement. An interpolation period which is too long yields a buzzy, scratchy quality, but the results otherwise generally sound smoother and more natural. The change in formant location and bandwidth is also noticeably smoother. LSF interpolation yielded slightly better performance than pole interpolation.

Using noise with the continuity effect does not give much smoothing for this junction. In three of the four cases, using shaped noise resulted in an unacceptable "hollow" sound. Using white noise was feasible in some cases but at times was worse than raw concatenation. Inserting too much noise could result in the /m/ being perceived as a frication. In general, the continuity effect results were poor.

Therefore, the LP algorithms – LSF interpolation in particular – give the best spectral smoothing performance for /m/–/iy/. We found that a typical good interpolation period is around three pitch periods, or 23 ms for a typical male speaker. Incorporating optimal coupling as well can yield further improvement, but LSF interpolation provides the most useful impact.

5. Results and evaluations

The net results of the examined algorithms showed improvement over standard techniques applied to small databases. The final speech is more natural-sounding than direct concatenation of selected units with no spectral processing. Still, even the best smoothing algorithms sometimes yield poor results at certain joints. Thus, blind use of a smoothing algorithm to all segment joints can result in speech that is of worse net quality than

Table 5
Summary of interpolation algorithms

Optimal coupling	
Summary	Adjust segment boundaries to improve spectral match
Advantages	Does not modify actual signal
Disadvantages	Limited benefit gained
Results	Most consistent improvement in quality
Evaluation	Better than nothing but not sufficient
Recommendation	Useful as is
Waveform interpolation	
Summary	Interpolate between two pitch periods
Advantages	Simple
Disadvantages	Does not consider formant locations
Results	Occasionally yields good results
Evaluation	Not good by itself but useful on LP residual
Recommendation	Useful on LP residual
LP interpolation	
Summary	Interpolate between linear predictive parameters
Advantages	Averages formants when parameters match formants
Disadvantages	Poor matching of parameters will give poor results
Results	Performance varies from good in many cases to poor
Evaluation	Often quite good; warrants more work
Recommendation	Useful as is; warrants more research
Closure (*continuity effect*)	
Summary	Insert noise shaped to match desired spectral envelope
Advantages	Spectrally matches surrounding signal
Disadvantages	Still possesses noisy quality
Results	Offers improvement primarily for transitions with noise-like sounds
Evaluation	Holds potential, but not good enough yet
Recommendation	Warrants more research

direct concatenation, but proper use of smoothing can noticeably increase the quality.

Table 5 summarizes the four major approaches which have been considered in this study. To evaluate these spectral smoothing algorithms, we performed comparisons based on a perceptual listener test and an objective quality measure. In addition, we show and describe a set of sample spectrograms that compare the algorithms' performance. We combine these results with our own subjective observations from this study to draw conclusions about the algorithms' effectiveness.

5.1. Data tested

For these evaluations, we chose to use two different speech databases. Our informal tests and evaluations were primarily based on the TIMIT database, where each speaker provides only ten phonetically-balanced sentences with approximately 400 phones. The phoneme labels for TIMIT include 60 distinct allophones.

In comparison, for our formal tests we used data collected specifically for this research study. For each speaker, the new database includes 35 phonetically balanced sentences and 114 words for a total of approximately 2300 phones. The corpus includes 34 continuous read sentences, of which 18 are adapted from the TIMIT database and 16 are adapted from the Brown text corpus.[1] There are also a large number of mono-syllabic, isolated words: 19 words are read three times each, and 57 words are read once each. Finally, there is 10 s of continuous, spontaneous speech. This database was phoneme-labeled with the RSPL speech time-aligner (Pellom and Hansen, 1998; Pellom, 1998), which models 46 different phoneme units.

For the comparison of smoothing algorithms, 14 sample words were created by constructing new words from the collected database. The first half of one word was matched with the second half of another word. Pitch periods were manually aligned between segments, and the segments were chosen so that the pitch varied by no more than 2.5 Hz at each joint. The original prosody of each segment was left unaltered. Spectrally smoothed frames were inserted via overlap-add.

The concatenated words and their source contexts are listed in Table 6. We chose to test segment joints that are common in English. Our work also places an emphasis on voiced data, and therefore all combinations included transitions to or from vowels. The following segment joints under test fall within seven of the nine most frequent phoneme-class transitions as measured from TIMIT (Pellom and Hansen, 1998): vowel → stop, semi-vowel (liquid/glide) → vowel, stop → vowel, fricative → vowel, vowel → nasal, vowel → semi-vowel, nasal → vowel.

5.2. Listener test

A number of informal listener tests were performed – both subjective and objective – in the various stages of examining the spectral smoothing algorithms. The results of several of these tests were mentioned or incorporated into the preceding text. Here we report only the final, formal listener test results.

In order to reach a large number of potential listeners, the listener test was implemented across the World Wide Web (WWW). Both the instructions and the test itself were placed on the WWW with speech available in multiple audio file formats to accommodate listeners on a variety of computers. Although direct control of the exact listening environment was not possible, we did make recommendations and ask listeners to report on the equipment used in performing the evaluation.

The test included several variations on 14 different words concatenated from segments in the database (see Section 5.1). Algorithm variations included the natural speech, raw concatenation without smoothing, optimal coupling, waveform interpolation, LP pole shifting, LSF interpolation, and shaped noise (continuity effect). In a brief training phase, listeners were presented with anchor speech signals to exemplify the high and low judgement categories. Listeners were asked to make a category judgement and give an opinion score for each word under test with nine ratings on a 1.0–5.0 scale (Deller et al., 2000; Quackenbush et al., 1988).

A total of 33 listeners with no history of hearing problems performed this mean option score (MOS) evaluation. Using a five-point MOS scale, with half-step ratings allowed, each algorithm

Table 6
Words used for listener test[a]

Word	Phonemes	Source	Context
Bear	/b//eh/r/	/b/oa/t/	/hh/eh/r/
Dog	/d//ao/g/	/d/eh/r/	/w/ao/g/
Fear	/f//iy/r/	/f/ay/r/	/w/iy/r/
Hair	/hh/eh//r/	/hh/eh/d/	/h/ay/r/
Here	/hh/ /ih/r/	/hh/aa/d/	/m/ih/r/
Hide	/h/ay//d/	/h/ay/r/	/hh/aa/r/d/
Make	/m/ey//k/	/m/ey/n/	/w/ow/k/
Mat	/m//ae/t/	/m/ao/n/	/b/ae/t/
Moon	/m/uw//n/	/m/uw/n/	/m/aa/n/
Nut	/n//ah/t/	/n/aw/	/b/ah/t/
Wait	/w//ey/t/	/w/eh/g/	/b/ey/t/
Wine	/w/ay//n/	/w/ay/k/	/m/ow/n/
Wire	/w/ay//r/	/w/ay/k/	/hh/ih/r/
Wood	/w//uh/d/	/w/iy/k/	/hh/uh/d/

[a] Phones were extracted from two source words and combined at a single concatenation point to synthesize a new word. Phoneme labels are from the TIMIT set. In the given phonemic spellings, "//" indicates the point of concatenation between phones.

Table 7
MOS results from listener test[a]

Algorithm	MOS	Better	Worse
Natural speech	4.13	N/A	N/A
Raw concatenation	3.53	N/A	N/A
Optimal coupling	3.82	77.0%	23.0%
Waveform interpolation	2.80	40.1%	59.9%
Pole shifting	2.69	38.4%	61.6%
LSF interpolation	3.14	39.4%	60.6%
Shaped noise (closure)	2.43	20.1%	79.9%

[a] Included are percent of tokens rated better than raw concatenation and percent worse than raw concatenation.

received at least one vote across all nine possible grades. This result points to the wide range of listener preference for all methods considered (e.g., from raw concatenation to each of the smoothing methods). Table 7 shows the mean opinion scores for each of the algorithms tested.

Fig. 16 plots these same MOS results. For each spectral smoothing technique, the mean opinion score is shown both for the entire test set and for each phoneme class. Each transition tested includes a vowel and a consonant (NA = nasal, LG = semi-vowel, ST = stop, or FR = fricative). The vertical bar indicates the sample standard deviation (SD) of the overall opinion scores with tick marks at 0.5 and 1.0 SD.

Note that the optimal coupling scores given here are only for those concatenated words for which coupling did change the final speech signal. In 8 of the 14 words of the test, the point of concatenation was the same for the raw concatenation and the optimally coupled forms. Although including the unmodified data for these eight words in with the coupled data decreases the MOS results for coupling, it still leaves coupling with a higher rating than raw concatenation without any smoothing.

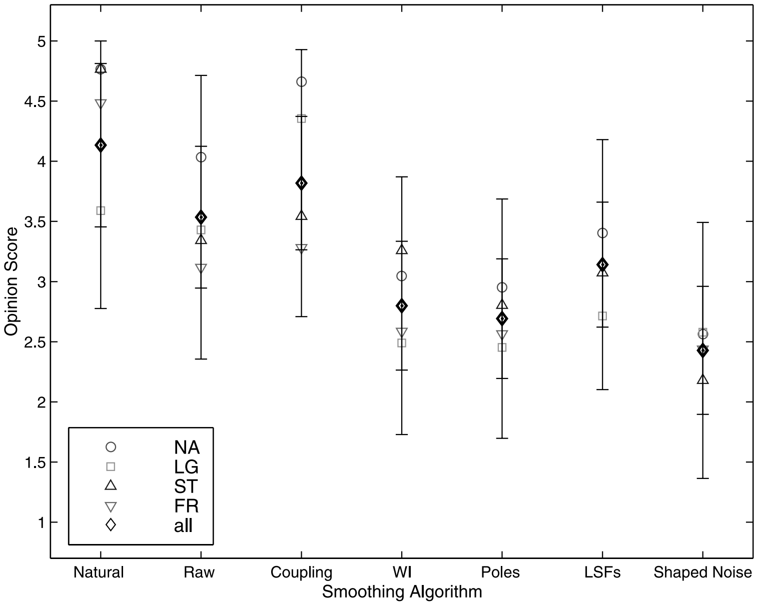

Fig. 16. Mean opinion scores from listener test. Evaluations are of transitions between vowels and specified phone classes.

The mean scores for each category indicate the general performance, but analysis was also performed on the relative scoring of each algorithm for each of the evaluated words. Table 7 also reports the fraction of cases in which each smoothing algorithm was rated better or worse than the same word with no smoothing. With the exception of optimal coupling, the opinion scores were generally lower after smoothing was performed. While the focus here was on the relative performance across the different smoothing methods, it is possible that the absolute MOS scores could change for a larger number of tested segment joints. Still, these results are in line with observations that smoothing can at times be useful or harmful depending upon the phoneme circumstances. Optimal coupling was the only algorithm which consistently performed well enough to receive the recommendation of general use without consideration of the situation.

5.3. Objective scoring

The ANBM (see Section 2.3) was used to provide an objective measure for assessing segment discontinuity. ANBM scores were calculated for 40 concatenated words which were smoothed with each of the techniques under evaluation. This word set includes all the examples in the listener test (see Section 5.2) in addition to other concatenated words. For each concatenated word, the ANBM score was obtained at multiple points around each joint with a comparison made for the maximum ANBM scores.

Fig. 17 reports the ANBM measure scores in the same format as the MOS scores in Fig. 16 (see Section 5.2). The large standard deviations for several of the algorithms reflect how the results vary widely from measureably improved to mildly degraded over pure concatenation. Note that the ANBM scores do vary by phoneme junction, and the standard deviation marks help indicate how each phoneme class performs for different algorithms compared with the overall test set. The white noise results are reported for comparison with shaped noise.

5.4. Spectrogram comparison

Fig. 18 shows one example spectrogram from each smoothing algorithm. The phrase "carry an oily rag" from TIMIT is used for these spectro-

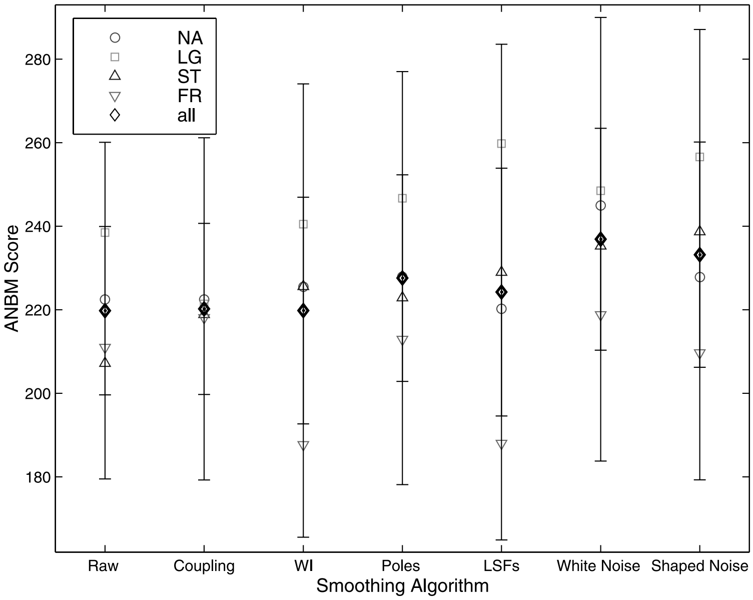

Fig. 17. ANBM scores for smoothed joints. Examples are from transitions between vowels and specified phone classes.

grams. The naturally-produced speech in Fig. 18(a) demonstrates that most formants have smooth transitions between phones yet some segment joints have rougher formant transitions.

Each subsequent spectrogram in the figure shows synthesized speech produced by concatenating phones with the same phone segments used in each example. The ANBM was used as the concatenation cost in selecting phone segments to attempt to find good spectral alignment in segments from the database. Note that the formants tend to be smoother and more continuous with several of the smoothing techniques, especially with LP pole shifting.

Fig. 18(b) does not include any spectral smoothing; although several joints have only small formant movement, others (e.g., at 0.62 s) have large and unnatural jumps in formant position. In Fig. 18(c) the /oy/ phone (0.4–0.6 s) and the /l/ phone (0.6–0.7 s) are clearly different than for the other examples shown; one result is a smoother formant transition at 0.6 s. In Fig. 18(d), several segments are noticeably different at the joints due to waveform interpolation; in the example at 0.6 s, the widely separated formants move towards each other compared with (b) but still show a rough transition. In the LP pole shifting example shown in Fig. 18(e), good spectral transitions are present for segments in the time region (0.1–0.4 s); however, poor smoothing is present near 0.6 s. In Fig. 18(f), LSF interpolation not only provides good smoothing for time region (0.1–0.4 s) but also distinctly improves the formant transitions near 0.6 s. It is important to note that processing via the continuity effect (as shown in Fig. 18(g)) is perceptually motivated, and as such the standard spectrogram will not display the perceived level of spectral smoothness.

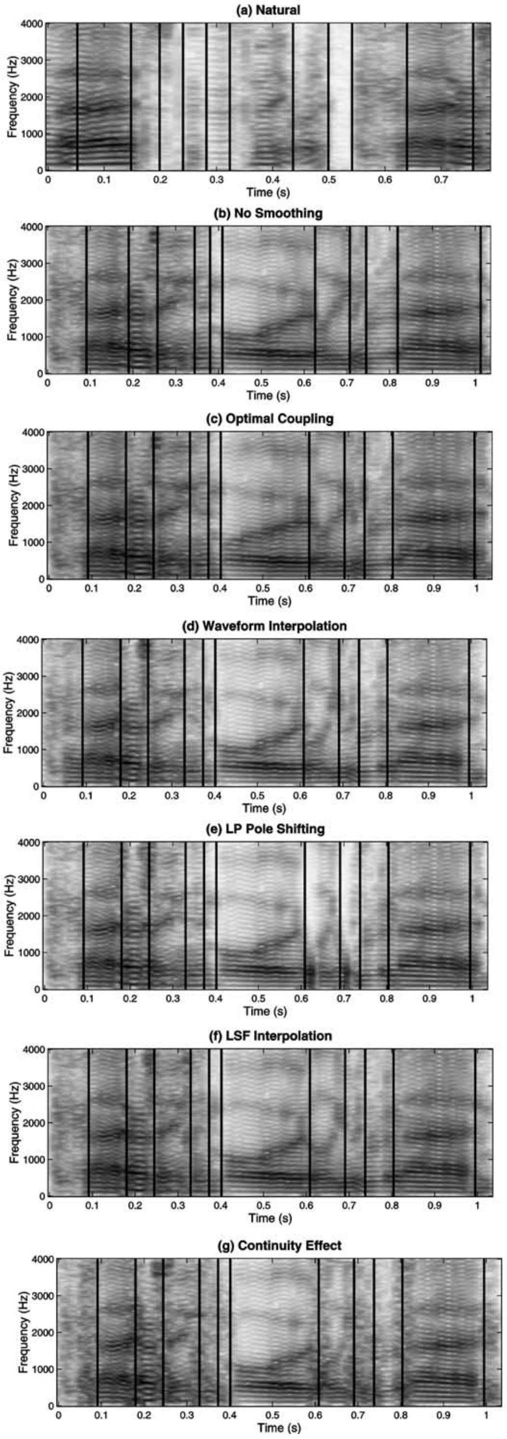

Fig. 18. Spectrograms of the phrase "carry an oily rag": (a) naturally produced and (b)–(g) concatenated speech. Solid vertical lines mark actual phone segment boundaries.

5.5. Discussion

The evaluations presented here show that spectral smoothing can at times improve the quality of speech yet at times can degrade it further. When smoothing succeeds it can noticeably increase the continuity of speech. In some scenarios it is better to perform no processing at all.

Moreover, which algorithm (if any) is best to use depends upon the circumstances.

Successful spectral smoothing can reduce the disfluency of speech. Smoothing seems to affect naturalness more than it affects intelligibility. The range of results from these evaluations does not imply that smoothing as a whole is not good, but instead it indicates that there is no single solution that can properly smooth all spectral discontinuities.

Thus, indiscriminate use of spectral smoothing is a poor choice because the results can produce further discontinuities. With current techniques, it generally would be best to manually inspect each joint after smoothing, but such user-assisted labor is typically impractical. The results from MOS listener evaluations and ANBM scores show that no smoothing method is clearly superior and that more effective methods are necessary. We recommend use of an existing automated quality-checking procedure such as rating with the ANBM.

The evaluations presented here have emphasized applications in concatenative synthesis with a limited data set of phonemes. Many of the concepts described herein also apply to synthesis with diphone sets and large corpora, but the emphasis shifts in such situations. When spectral smoothing is used in speech and audio coding, the situation differs because the speech was originally spectrally continuous. For coding applications, smoothing is typically simpler and does not encounter some of the previously mentioned problems that accompany concatenative synthesis. For example, LP parameter matching has a higher success rate with coding applications, and the duration of smoothing is sometimes shorter. Moreover, the quality of the resulting smoothed speech is generally higher for coding.

6. Conclusions

In this study, we have focused on a comparison of four algorithms for spectral smoothing of concatenated speech. The algorithms considered include three major existing techniques for smoothing – optimal coupling, waveform interpolation, and LP interpolation – and one technique (application of the continuity effect) which was considered for spectral smoothing for the first time. In addition to performing extensive informal comparisons of the algorithms, we have reported results from a formal listening test and scoring with an auditory-based objective measure. These evaluations have been performed in the context of a phoneme concatenation synthesizer with a small data set. The net results of the discussed algorithms are that no method is clearly superiod (see Table 5) and no single algorithm performs best in all phone joint circumstances. Application of smoothing methods to many of the smoothed segment joints demonstrate noticeable improvements over direct concatenation, while other joints are of noticeably worse quality after applying a smoothing algorithm. This study has shown that most segment based smoothing methods are not universally successful for all segment joints and that the use of an objective measure of segment joint quality is necessary to direct more effecting smoothing.

Although synthesis systems with smoothing typically apply a single algorithm indiscriminately, we recommend using a smart system. Knowledge of the phonemes involved in each joint enables selection of an appropriate smoothing algorithm. Scoring with an objective measure such as the ANBM enables automated evaluation of whether the smoothing has improved or degraded the perceived quality of each transition.

While using such a smart system enables improvements over current techniques, there is still room for enhancement in spectral smoothing. For example, LP pole shifting could greatly benefit from the derivation of a better distance measure for matching poles in the z-plane across frames. In addition, the proposed use of the continuity effect is novel and promising enough that there may be ways to improve it that we have not yet considered.

When spectral smoothing is appropriately applied, the final speech has smoother, more continuous formants and is often more natural-sounding than direct concatenation of segments without processing. These improvements benefit speech coding by enabling appropriate generation of intermediate data between transmitted frames. These smoothing algorithms also improve the results of concatenative speech synthesis with a limited database.

Acknowledgements

This study was supported in part by SPAWAR under grant no. N66001-92-D-0092. The authors thank M. Ostendorf (now at Washington University) for providing C code for the Carney auditory-nerve model during the 1993 DoD Workshop on Robust Speech Recognition, Rutgers University. The authors also thank Bryan Pellom and Robert Ross, also with the Robust Speech Processing Laboratory, for their assistance in collecting and labeling some of the speech corpora used in this study.

References

Atal, B.S., Cox, R.V., Kroon, P., 1989. Spectral quantization and interpolation for CELP coders. In: Proc. 1989 IEEE ICASSP, Glasgow, Scotland, Vol. 1, pp. 69–72.

Berry, D.A., 1996. Statistics: A Bayesian Perspective. Duxbry, Belmont, CA.

Breen, A.P., Jackson, P., 1998. A phonologically motivated method of selecting non-uniform units. In: Proc. 1998 ICSLP, Sydney, Australia.

Bregman, A.S., 1990. Auditory Scene Analysis: The Perceptual Organization of Sound. MIT Press, Cambridge, MA.

Carney, L.H., 1992. A model for the responses of low-frequency auditory-nerve fibers in cat. Journal of the Acoustical Society of America 93, 401–417.

Chappell, D.T., Hansen, J.H.L., 1997. An auditory-based measure for improved phone segment concatenation. In: Proc. 1997 IEEE ICASSP, Munich, Germany, Vol. III, pp. 1639–1642.

Coker, C.H., 1976. A model of articulatory dynamics and control. Proc. IEEE 64, 452–460.

Conkie, A.D., Isard, S., 1997. Optimal coupling of diphones. In: van Santen, J.P.H., Sproat, R.W., Olive, J.P., Hirschberg, J. (Eds.), Progress in Speech Synthesis, Springer, New York, Chapter 23, pp. 293–304.

Deller Jr., J.R., Hansen, J.H.L., Proakis, J.G., 2000. Discrete-Time Processing of Speech Signals. IEEE Press, New York.

Donovan, R.E., 1996. Trainable speech synthesis. Ph.D. thesis, Department of Engineering, Cambridge University.

Dutoit, T., 1994. High quality text-to-speech synthesis: a comparison of four candidate algorithms. In: Proc. 1994 IEEE ICASSP, Adelaide, South Australia, Vol. 1, pp. 565–568.

Dutoit, T., Leich, H., 1993. MBR-PSOLA: Text-To-Speech synthesis based on an MBE re-synthesis of the segments database. Speech Communication 13, 435–440.

Erkelens, J.S., Broersen, P.M.T., 1994. Analysis of spectra interpolation with weighting dependent on frame energy. In: Proc. 1994 IEEE ICASSP, Adelaide, South Australia, Vol. 1, pp. 481–484.

Fant, G., 1960. Acoustic Theory of Speech Reproduction. Mouton, The Hague.

Flanagan, J.L., 1972. Speech Analysis, Synthesis and Perception, second ed. Springer, New York.

Goncharoff, V., Kaine-Krolak, M., 1995. Interpolation of LPC spectra via pole shifting. In: Proc. 1995 IEEE ICASSP, Detroit, MI, Vol. 1, pp. 780–783.

Hansen, J.H.L., Chappell, D.T., 1998. An auditory-based distortion measure with application to concatenative speech synthesis. IEEE Transactions on Speech and Audio Processing 6, 489–495.

Hirokawa, T., Hakoda, K., 1990. Segment selection and pitch modification for high quality speech synthesis using waveform segments. In: Proc. 1990 ICSLP, Kobe, Japan, Vol. 1, pp. 337–340.

Huang, X., Acero, A., Hon, H., Ju, Y., Liu, J., Meredith, S., Plumpe, M., 1997. Recent improvements on Microsoft's trainable text-to-speech system–Whistler. In: Proc. 1997 IEEE ICASSP, Munich, Germany, Vol. II, pp. 959–962.

Hunt, A.J., Black, A.W., 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In: Proc. 1996 IEEE ICASSP, Atlanta, GA, pp. 373–376.

Klabbers, E., Veldhuis, R., 1998. On the reduction of concatenation artifacts in diphone synthesis. In: Proc. 1998 ICSLP, Sydney, Australia, Vol. 5, pp. 1983–1986.

Kleijn, W.B., Haagen, J., 1995. Waveform interpolation for coding and synthesis. In: Kleijn, W.B., Paliwal, K.K., (Eds.), Speech Coding and Synthesis, Elsevier, Amsterdam, Chapter 5, pp. 175–207.

Kleijn, W.B., Shoham, Y., Sen, D., Hagen, R., 1996. A low-complexity waveform interpolation coder. In: Proc. 1996 IEEE ICASSP, Atlanta, Georgia, Vol. 1, pp. 212–215.

Ladefoged, P., 1975. A Course in Phonetics, third ed. Harcourt Brace, New York.

Ladefoged, P., 1981. Preliminaries to Linguistic Phonetics. University of Chicago, Chicago.

Liberman, M.C., 1982. The cochlear frequency map for the cat: labeling auditory-nerve fibers of known characteristic frequency. Journal of the Acoustical Society of America 72, 1441–1449.

Mizuno, H., Abe, M., 1995. Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt. Speech Communication 16, 153–164.

Mizuno, H., Abe, M., Hirokawa, T., 1993. Waveform-based speech synthesis approach with a formant frequency modification. In: Proc. 1993 ICASSP, Vol. 2, pp. 195–198.

Moore, B.C.J., 1997. An Introduction to the Psychology of Hearing, fourth ed. Academic Press, New York.

Moulines, E., Charpentier, F., 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Communication 9, 453–467.

Moulines, E., Laroche, J., 1995. Non-parametric techniques for pitch-scale and time-scale modification of speech. Speech Communication 16, 175–205.

O'Shaughnessy, D., 1990. Speech Communication: Human and Machine. Addison-Wesley, New York.

Paliwal, K.K., 1995. Interpolation properties of linear prediction parametric representations. In: Proc. EuroSpeech'95, Madrid, Vol. 2, pp. 1029–1032.

Paliwal, K.K., Kleijn, W.B., 1995. Quantization of LPC parameters. In: Kleijn, W.B., Paliwal, K.K. (Eds.), Speech Coding and Synthesis. Elsevier, Amsterdam, pp. 433–466.

Papamichalis, P.E., Cliffs, N.J, 1987. Practical Approaches to Speech Coding. Prentice-Hall, Englewood.

Parthasarathy, S., Coker, C.H., 1992. On automatic estimation of articulatory parameters in a text-to-speech system. Computer Speech and Language 6, 37–75.

Pellom, B.L., 1998. Enhancement, segmentation, and synthesis of speech with application to robust speaker recognition. Ph.D. thesis, Robust Speech Processing Laboratory, Dept. of Electrical Engineering, Duke University.

Pellom, B.L., Hansen, J.H.L., 1998. Automatic segmentation of speech recorded in unknown noisy channel characteristics. Speech Communication 25, 97–116.

Pickett, J.M., 1980. The Sounds of Speech Communication: A Primer of Acoustic Phonetics and Speech Perception. University Park Press, Baltimore.

Plumpe, M., Acero, A., Hon, H., Huang, X., 1998. HMM-based smoothing for concatenative speech synthesis. In: Proc. 1998 ICSLP, Sydney, Australia, Vol. 6, pp. 2751–2754.

Quackenbush, S.R., Barnwell, T.P., Clements, M.A., 1988. Objective Measures of Speech Quality. Prentice-Hall, Englewood Cliffs.

Rabiner, L., Juang, B.-H., 1993. Fundamentals of Speech Recognition. Prentice-Hall, Englewood Cliffs, NJ.

Savic, M., Nam, I.-H., 1991. Voice personality transformation. Digital Signal Processing 1, 107–110.

Shadle, C.H., Atal, B.S., 1979. Speech synthesis by linear interpolation of spectral parameters between dyad boundaries. The Journal of the Acoustical Society of America 66, 1325–1332.

Shiga, Y., Matsuura, H., Nitta, T., 1998. Segmental duration control based on an articulatory model. In: Proc. 1998 ICSLP, Sydney, Australia, Vol. 5, pp. 2035–2038.

Slaney, M., Covell, M., Lassiter, B., 1996. Automatic audio morphing. In: Proc. 1996 IEEE ICASSP, Atlanta, Georgia, pp. 1001–1004.

Slifka, J., Anderson, T.R., 1995. Speaker modification with LPC pole analysis. In: ICASSP-95, Vol. 1, pp. 644–647.

Snell, R.C., Milinazzo, F., 1993. Formant location from LPC analysis data. IEEE Transactions on Speech and Audio Processing 1, 129–134.

Stevens, K.N., House, A.S., 1955. Development of a quantitative description of vowel articulation. Journal of the Acoustical Society of America 27, 484–493.

Syrdal, A., Stylianou, Y., Garrison, L., Conkie, A., Schroeter, J., 1998. TD-PSOLA versus harmonic plus noise model in diphone based speech synthesis. In: ICASSP-98, Seattle, Vol. 1, pp. 273–276.

Warren, R.M., 1970. Perceptual restoration of missing speech sounds. Science 167, 392–393.

Witten, I.H., 1982. Principles of Computer Speech. Academic Press, New York.

Zemlin, W.R., 1968. Speech and Hearing Science: Anatomy and Physiology. Prentice-Hall, Englewood Cliffs, NJ.