AN AUDITORY-BASED MEASURE FOR IMPROVED PHONE SEGMENT CONCATENATION

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

ABSTRACT

This paper describes a new auditory-based distance measure intended for use in a concatenated synthesis technique wherein time- and frequency-domain characteristics are used to perform natural-sounding speaker synthesis. Whereas most concatenation systems use large databases (often +100,000 units), we begin from a small, limited database (approx. 400 units) and use a new spectral distortion measure to aid in the selection of phones for optimal concatenation. At the transition between speech segments, the new auditory-based distance metric assesses perceived discontinuities in the frequency domain. The distortion measure, which employs the Carney auditory model, is used to select phones which minimize the perceived distortion between concatenated segments. Moreover, time- and frequency-domain methods can shape the prosodic and spectral characteristics of each speech segment. The final results demonstrate improved performance over standard concatenation methods applied to small databases.

1. INTRODUCTION

For several applications, it is desirable to synthesize short passages of a speaker's voice from limited training data. Such applications include response systems in voice dialog applications and voice e-mail simulation via synthesis.

Common text-to-speech systems based on concatenation produce continuous speech by selecting waveform units from speech databases. Most of these systems employ databases with a large number (e.g., over 100,000) of available segments with varied characteristics. Although such work can yield high-quality speech [1, 2], these algorithms rely upon their large databases and are oriented primarily for text-to-speech usage. Creation and labeling of such databases require large commitments of time and resources. These large-database synthesis schemes generally concentrate on segment choice and reducing unit selection search times since they have enough sample units to be able to find a suitably close match for each desired phoneme.

In contrast, we propose an algorithm for limited speech synthesis that uses a small database of only about four hundred phone unit waveforms per speaker. The time required to create and label such a limited data set is several orders of magnitude smaller than that for larger schemes. In particular, we chose the TIMIT database, which provides only ten phonetically-balanced sentences **for each speaker**. For each TIMIT speaker, the ten-sentence training corpus typically includes five to eight examples of each phoneme with only three or four strings of usable adjacent matching phones for each synthesized sentence. With such limitations, the process of unit selection does not take priority since even the best segment selection is not likely to suffice. Instead, each sample segment must be modified to fit the desired properties. In addition to matching each phone's internal parameters, the spectral characteristics of the beginning and ending of each phone must be molded to smooth the transition between adjoining units.

In this paper, we propose a new auditory-based measure which aids in the production of natural-sounding speech via segment concatenation from a limited database. The basic waveform unit used is the phone, with diphones and multiple adjacent phones used when appropriate matches reside in the database of ten source sentences. While standard schemes choose phones for concatenation based upon their values for pitch, duration, and power, we use a newly-developed metric to measure the perceived spectral match of adjacent units. Each selected phone is shaped with a time-domain analysis-synthesis normalization scheme to provide the desired pitch and duration [2]. The spectral features may also be shaped to improve the aforementioned new spectral matching metric. By combining these approaches, the concatenated speech can not only closely approximate the desired perceptual speech characteristics but also be continuous in both the time domain and spectral structure.

2. TIME-DOMAIN PARAMETER MATCHING

With the limited number of phones in a small database environment, it is rarely possible to find an existing phone with the desired prosodic characteristics. A search routine can scan all possibilities to locate the closest match, but direct concatenation yields a jumble of mismatched units, none of which have quite the desired parameters. If the characteristics can be modified, however, then even the units available in the limited data set can yield acceptable phones. Although the goal is to have a given speaker produce the speech naturally, for purposes of comparison, we have a second speaker produce the desired sentence.

Thus, we use the Pitch-Synchronous Overlap and Add (PSOLA) algorithm to adjust the pitch and duration of each phone. By manipulating pitch-synchronous analysis windows, the method provides a simple mechanism for prosodic adjustment [3]. Once each speech segment has been selected from the database, the PSOLA scheme adjusts the pitch and duration of each voiced phone to match the goal. While it would be preferable to have a perfectly matched phone, modifying the available data is a practical method of achieving quite similar results.

In addition to the phone pitch and duration manipulation via PSOLA, the power of each speech segment is also adjusted. Standard time-domain techniques are used to change the RMS amplitude of each phone to the desired value. Modifying these three characteristics—pitch, duration, and power—allows the limited database to produce synthesis results similar to those of much larger data sets.

In contrast, methods with huge databases scan for near-exact matches of prosodic characteristics. They find the closest match and minimize the introduction of artifacts which might appear due to prosodic modification [1]. With our small database, though, the best prosodic matches still differ enough from the target that use of PSOLA is necessary despite the introduction of any artifacts.

3. SPECTRAL DISTORTION MEASURE

To aid in both the selection and transition smoothing of phones, we propose a new auditory-based distortion measure. This metric bears similarities to common objective speech quality measures such as those based on LP properties [4, 5], critical-band structure, the Bark-scale, or line spectral pairs [6]. While these measures have been used historically to assess performance of vocoder distortion, the present distortion metric is intended to identify how much a spectral distortion is noticeable from a perceptual basis.

3.1 Motivation

In one study, Schuyler Quackenbush considered over 2000 objective speech quality measures and found that the best predictors of subjective quality are those derived based on auditory criteria; of these, Klatt's weighted-spectral slope measure (WSSM) gives the best results [4]. Thus, to best use such auditory criteria, our proposed metric measures the "distance" between the perceived spectral characteristics of adjacent time-slices in the speech signal.

Several existing measures are comparable to the proposed metric, but we emphasize measurement of the human auditory perception of mismatches between segments. Simply taking the difference between LP spectral envelopes yields an expression of spectral continuity [2, 4]. Our metric is similar in concept to spectral distance measures, but whereas traditional spectral measures are based upon smoothed spectra, we investigate a more perceptual-based spectral representation via the Carney model of auditory-nerve (AN) fibers.

3.2 Auditory Model

The perceptual basis for this new measure arises from the use of Laurel Carney's computational model for generating firing rates of synapses of auditory nerves [7]. This

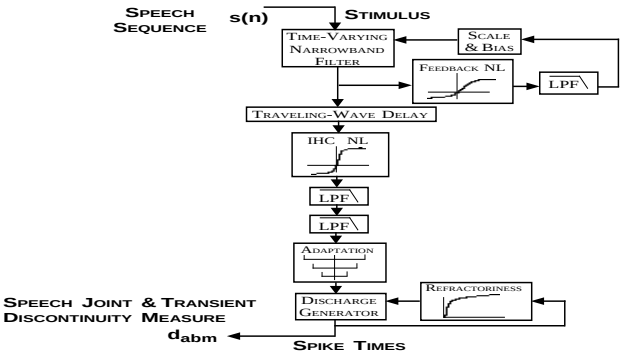

Figure 1: Carney Model Flow Diagram

model (see Figure 1) is based upon and closely approximates measurements of AN fibers in cats. The parameters of the narrow-band filter vary continuously as a function of stimulus level via a feedback mechanism which simulates the compressive nonlinearity associated with the mechanics of the basilar membrane. A nonlinearity and pair of lowpass filters are used to simulate membrane transduction properties of the inner hair cell (IHC). Along with Carney's basic algorithm , we used Mark Liberman's placement map of IHC characteristic frequencies [8].

One form of data output by the Carney model is the spike rate of the IHC-AN synapse. This data can be used to generate a cochleagram (a.k.a. neurogram), which is a picture of the firing rates versus time for an array of AN fibers. Figure 2 illustrates a cochleagram of the phone /AO/ from the phrase "all year." It is this data which our measure employs on a per-AN basis.

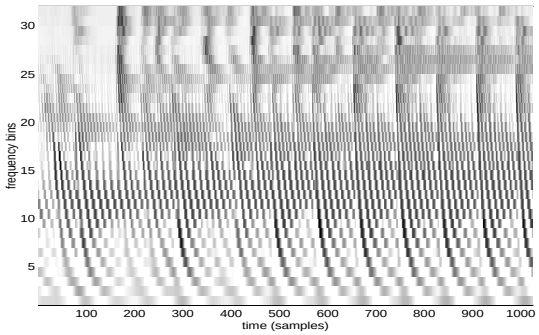

Figure 2: Sample Cochleagram

This measure takes advantage of a computational model which produces realistic temporal response properties and average discharge rates in response to simple and complex stimuli. The parameters of the narrow-band filter vary continuously as a function of stimulus level via a feedback mechanism which simulates the compressive nonlinearity associated with the mechanics of the basilar membrane. A nonlinearity and pair of lowpass filters are used to simulate membrane transduction properties of the inner hair cell.

Despite its advantages, the Carney algorithm is only a model, and it is not accurate in all aspects. In particular, it was designed for a single low-frequency AN and does not account for crosstalk, lateral neural inhibition, and other

effects that arise in practice with a full array of hair cells and nerve fibers. Moreover, use of the Carney model is not a requirement for this style of measure, and one could use another auditory model which generates cochleagram data.

3.3 Measure Algorithm

Malcolm Slaney and Richard Lyon have studied correlograms, which are series of short-time autocorrelations of AN firing rate data. They propose that the correlogram is biologically feasible and may be a representation by which the brain organizes sound information [9]. Figure 3 shows an example of one time-slice of the correlogram of the phone /AO/ from the phrase "all year." The distinct periodicity of the response of each AN corresponds to the frequency data extracted and used in our measure. Other research also indicates that the perception of sound and speech depends upon the spectra of the firing rates of the auditory nerves.

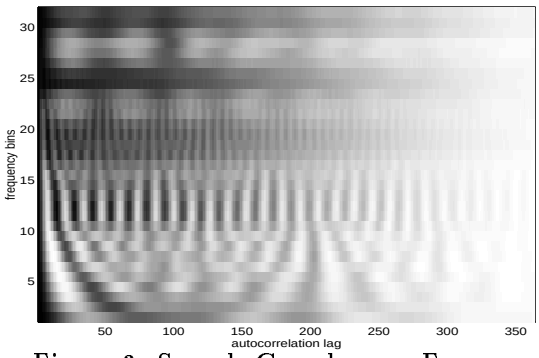

Figure 3: Sample Correlogram Frame

Figure 4 illustrates the flow chart for generating feature vectors for this distance measure. The input data is the rate of AN synapse firing from the Carney model. The analysis stage of the algorithm calculates the primary frequency at which each AN fires. For each AN frequency bin, we 1) apply a Hamming window, 2) take the linear autocorrelation, 3) find the highest-amplitude frequency from the autocorrelation, and 4) store the frequency value within the feature data vector. This vector is calculated for a single window frame at the beginning or end of a speech segment. For the beginning of a speech file, about 25 ms of AN data should be skipped to allow transient effects to fade. Once the feature vectors have been generated, the final scalar measure value is found by summing the difference of values between two vectors using the city block metric. When applying this measure to selecting phones for concatenation synthesis, we use the vectors from the end of one segment and the beginning of the next segment which is a candidate for concatenation. Saving the feature vectors reduces the computational requirements of calculating AN data for each speech sample.

To find the primary AN firing frequency, one could use the power spectrum of the synapse output, but in practice linear autocorrelation performs better. To find the primary frequency of AN firing, one can find the maximum peak of the FFT of the autocorrelation or directly measure the distance between peaks in the linear autocorrelation. Note that we use linear correlation because tests show that it

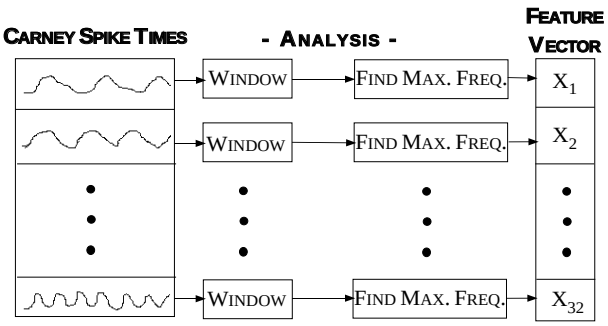

Figure 4: Auditory Measure Flow Diagram

performs better in practice than either circular correlation or a mere FFT. Our correlograms similarly use the linear rather than the circular autocorrelation function.

Since it is used to measure how well segments match spectrally at their edges, this metric forms one basis for choosing potential speech units from the database. Moreover, this measure yields information on where in the frequency domain there is perceptual mismatch or distortion so that additional signal processing can smooth any discontinuities. As such, it can provide valuable and immediate feedback as to the level of spectral mismatch in the segment concatenation string under consideration.

This distance function meets the criteria for being a metric of feature vectors and also qualifies as being invariant; thus, it has a high degree of mathematical tractability.

4. SPECTRAL SMOOTHING

Even though time-domain adjustments can shape the prosodic characteristics of each phone, the resulting waveforms do not immediately fit together to produce natural-sounding speech. Whereas a large database is more likely to contain sample segments with closely-matching spectral characteristics, we need to use additional means to ensure that the selected phones are spectrally aligned so as not to be discordant at phoneme transitions.

Within the time domain, segments can be truncated to begin and end only at pitch peaks, but synchronizing the pitch is only a first step. The fundamental frequency, or pitch, can be adjusted using standard pitch-synchronous analysis-synthesis methods. Other frequency-domain methods are needed, however, to provide a smooth spectral transition between formants. Frequency-domain FD-PSOLA and other spectral techniques can make the available segments fit together smoothly and thereby improve the quality over that of direct concatenation [5].

Our proposed measure can yield some limited information on the location of discontinuities in the frequency domain. Analysis of the positions of the differences between measure vectors yields an indication of where perceptual distortion lies in the frequency domain. However, a sufficiently strong sound-pressure level will cause AN fibers to respond to frequencies far from their characteristic frequencies with each AN responding according to its tuning curve. Thus, changes in actual frequencies present in the waveform may affect distant IHCs; in particular, at slight spectral shifts the change appears primarily at the IHCs which are just on the edge of

being affected by the stimulating frequency. For example, simulations with sine-wave pseudo-formants reveal that a change in one formant can affect IHCs with a characteristic frequency of up to 800 Hz away. Instead of relying solely on our measure to correct spectral discontinuity, one could use this auditory measure to decide whether a perceived mismatch is present and then use the Klatt spectral slope measure to determine locations for spectral smoothing.

5. EVALUATIONS

To evaluate this proposed algorithm, we set up two styles of tests. We first demonstrated the measure's performance on a set of summed sinusoids and then used the measure to select phones for concatenation synthesis.

We compared our auditory-neural distance measure (ANDM) to the Itakura-Saito weighted log-likelihood distortion measure (ISM) and WSSM to evaluate their responses to changes in a sine-wave pattern of artificial formants. To provide a set of quantitative evaluations, we generated patterns of sine waves to simulate three formants with the position of a single formant shifting to a higher frequency. The frequency and amplitude of these sine waves are based upon mean measured data for the vowel /AE/ [10]. In one set of trials, the pseudo-formants are changed to different frequencies for comparison; a second set of trials has a consistent frequency shift but has different amplitudes of the formant which shifts. The intention of these simulations is to allow detailed measurements that are applicable to real concatenation situations. Figure 5 illustrates the results for each of the three measures for these shifts for multiple frequencies and amplitudes. Although our measure does not show ideal monotonic increase with greater amplitude and frequency shift, it does reflect both the amount of spectral change and a dependence upon signal amplitude.

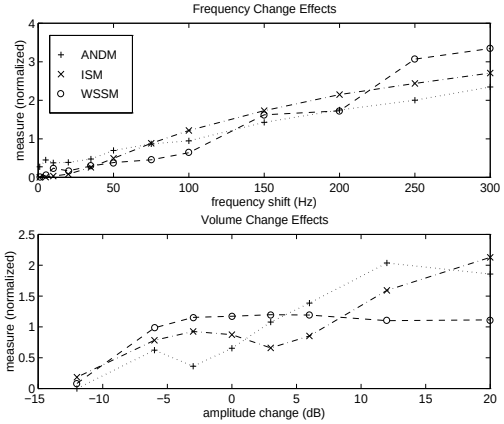

Figure 5: Distance Measure Comparison

To demonstrate the capabilities of this new measure, we employed it to select phones from a TIMIT speaker for concatenative synthesis. Based upon the measure, we constructed synthesized phrases from a 10-sentence phone corpus from the TIMIT database. In Figure 6, spectrograms of the synthesized phrase "all year" illustrate how the formants match better when our distance measure is used to select phones rather than more traditional prosodic matching.

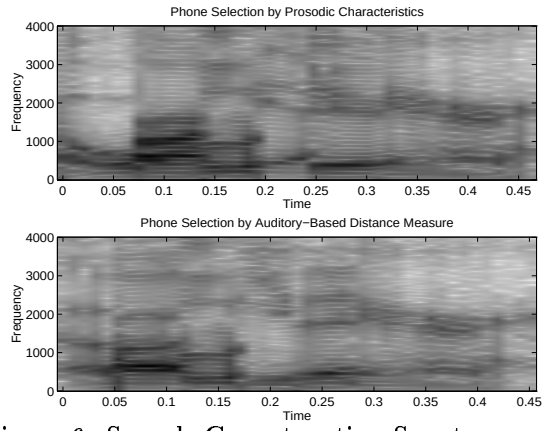

Figure 6: Sample Concatenation Spectrograms

6. CONCLUSIONS

In this study, a new auditory-based distance metric is developed and shown to provide a measure of dissimilarity of the perception of speech segments. When used as a basis for phone selection, the metric provides smoother spectral transitions for concatenative synthesis. When combined with direct processing of the time- and frequency-domain characteristics, our measure allows a small database to yield concatenation synthesis results similar to those previously achieved only with very large databases. Thus, based upon experiments conducted using the Carney model of auditory-nerve fibers, we feel that the proposed measure would be an effective approach to assess segment joint mismatch for low-segment-size codebooks for speech synthesis.

ACKNOWLEDGEMENT

The authors thank Mari Ostendorf of Boston University for providing C code for the Carney auditory-nerve model at the 1993 DoD Workshop on Robust Speech Recognition.

References

[1] A. J. Hunt, A. W. Black, "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database," *Proc. 1996 IEEE ICASSP,* pp. 373-376.

[2] T. Hirokawa, K. Hakoda, "Segment Selection and Pitch Modification for High Quality Speech Synthesis using Waveform Segments," *Proc. 1994 ICSLP,* pp. 337-40, 1994.

[3] E. Moulines, J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Communication,* vol. 16, pp. 175-205, 1995.

[4] S. R. Quackenbush, T. P. Barnwell, M. A. Clements, *Objective Measures of Speech Quality,* Prentice-Hall, 1988.

[5] D. Rainton, S. J. Young, "Time-Frequency Spectral Analysis of Speech," *Proc. 1994 ICSLP,* pp. 349-52.

[6] H. J. Coetzee, T. P. Barnwell III, "An LSP Based Speech Quality Measure," *Proc. 1989 ICASSP,* pp. 596-99.

[7] L. H. Carney, "A model for the responses of low-frequency auditory-nerve fibers in cat," *J. Acoust. Soc. Am.,* vol. 93, pp. 401-17, 1993.

[8] M. C. Liberman, "The cochlear frequency map for the cat," *J. Acoust. Soc. Am.,* vol. 72, pp. 1441-49, 1982.

[9] M. Slaney, R. F. Lyon, "On the Importance of Time–A Temporal Representation of Sound," *Visual Representations of Speech Signals,* M. Cooke, S. Beet, M. Crawford, eds., pp. 95-116, 1993.

[10] G. E. Peterson, H. L. Barney, "Control methods used in a study of the vowels," *J. Acoust. Soc. Am.,* vol. 24, pp. 175-84, 1952.