A PERCEPTUAL EVALUATION OF DISTANCE MEASURES FOR CONCATENATIVE SPEECH SYNTHESIS

Johan Wouters *Michael W. Macon*

cslu.cse.ogi.edu/tts
Center for Spoken Language Understanding
Oregon Graduate Institute, PO Box 91000, Portland, OR 97291-1000, USA

ABSTRACT

In concatenative synthesis, new utterances are created by concatenating segments (units) of recorded speech. When the segments are extracted from a large speech corpus, a key issue is to select segments that will sound natural in a given phonetic context. Distance measures are often used for this task. However, little is known about the perceptual relevance of these measures. More insight into the relationship between computed distances and perceptual differences is needed to develop accurate unit selection algorithms, and to improve the quality of the resulting computer speech. In this paper, we develop a perceptual test to measure subtle phonetic differences between speech units. We use the perceptual data to evaluate several popular distance measures. The results show that distance measures that use frequency warping perform better than those that do not, and minimal extra advantage is gained by using weighted distances or delta features.

1 INTRODUCTION

To produce high quality concatenated speech, it is important to combine segments with appropriate coarticulation or phonetic "coloring". Many concatenative synthesizers store exactly one segment for each phonetic context (e.g. diphones). Recently, researchers have addressed the challenge of selecting segments from any database of naturally spoken text. Several unit selection algorithms that rely on objective distance measures have been proposed [2, 3, 8].

In speech recognition, distance measures have been used more widely than in speech synthesis. Recognition algorithms based on template matching, using Dynamic Time Warping (DTW), applied distance measures directly. Currently, more research is aimed at improving feature representations of speech, which are the basic building block of distance measures and are used as a front-end to Hidden Markov or Neural Network based recognizers. Distance measures are also important in speech coding, for use in vector quantization and as objective measures of speech quality [14].

Relatively few studies have attempted a large scale comparison of distance measures. Two reasons can be found for this. First, a distance measure is the result of many design choices, and to investigate all possible combinations is an enormous task. Second, often the only criterion to decide whether a certain distance measure is better than another, is its performance as part of a speech recognizer, coder, or synthesizer. Conclusions reached on performance of a distance measure within a certain algorithm or application, may not be valid in a different setting.

An early study comparing several distance measures was conducted by Gray and Markel [5]. They investigated measures based on spectral and cepstral coefficients, log area ratios, and the Itakura-Saito distance. They showed that the cepstral distance with 10 to 20 coefficients is an efficient estimation of the log spectral distance, and proved other relations between the measures both in theory and experimentally. Nocerino, Rabiner, and Klatt [13] studied the performance of several feature representations in a DTW recognizer. They concluded that warped frequency scales (such as mel scale and bark scale) did not improve performance. The opposite was found by Hermansky and Junqua [7], and Krishnan and Rao [11], in different recognizers. Krishnan and Rao also found promising results for features based on line spectral frequencies.

Such comparative studies give an insight into the range of distance measures that can be designed. They also provide evidence that certain feature representations capture more of the variability in speech that is relevant in recognition. However, for the purpose of unit selection in speech synthesis, we are interested in the relation between computed distance measures and human perception. We were able to find surprisingly little research on this topic, although many researchers [5, 16] have pointed out its need.

An exception is the work by Quackenbush, Barnwell and Clements [14]. In their book, "Objective Measures of Speech Quality," they study the perceptual effects of several speech coding distortions, and investigate the potential of a large number of distance measures to predict the perceptual data. Among their conclusions, they report a correlation of 0.7 between perceptual quality measures and some of the best automatic measures. However, coding distortions have a different effect on the speech signal than allophonic variations. New research is needed to study which measures best predict differences between allophones. Also, a perceptual test must be designed that is not aimed at judging overall quality of distorted speech, but at specific measurements of phonetic changes.

The paper is organized as follows. In Section 2, we describe a perceptual experiment. A database of speech samples and associated perceptual distances is constructed. In Section 3, we compare several frequently used distance measures with the perceptual data. In conclusion, we discuss their merits for concatenative speech synthesis.

2 A PERCEPTUAL TEST FOR ALLOPHONIC DIFFERENCES

2.1 Design

Each phoneme in a language can be realized as a continuum of phonetic variations in natural speech. Such variations are called allophones. For example, the /i/ in "tip" differs from the /i/ in "tick". Not only phonetic context, but also lexical stress, mood of the speaker, speech rate, etc. influence the realization of a phoneme.

We aim to measure the subtle perceptual differences between allophonic speech segments. A naive perceptual test would consist of extracting samples of a phoneme from different contexts, and playing the isolated sounds to a lis-

*This work was supported by a grant from Intel Corporation, and by the members of the CSLU Industrial Consortium.

tener. It would be a strenuous task for a listener to judge the differences between these samples. Not only are the phoneme durations very short, but also the differences become salient only when placed in a phonetic context.

In our test, listeners are instead presented with pairs of words that are identical except for one segment. Segments from different phonetic contexts are inserted, causing perceptible differences in the pronunciations of the words. A segment is set to be one half of a phoneme, which is the basic unit of our concatenative synthesizer [12]. Figure 1 illustrates this process.

The number of phonetic contexts in English is very large, and we can cover only a small fraction of the phonetic spectrum in the perceptual test. Hence, we limit the substituted segments to three specific cases of *vowels*. These cases are explained below. We expected that variations of vowel segments would result in a relatively wide range of perceptual differences, which would allow listeners to rate them on a five-point scale (distances from 0 to 4). This is similar to perceptual experiments with synthesized vowels, such as reported by Kewley-Port and Atal [9], and Klatt [10].

2.2 The Test Database

Every word pair in the perceptual test consists of a reference word and a modified version of this word. The reference word is realized by a diphone synthesizer, which is assumed to produce 'correct' allophonic variations. The modified word differs by one half of a diphone, extracted from a different phonetic context. The diphones are joined pitch-synchronously, but without spectral smoothing. Pitch contours and phoneme durations are identical for both words. Energy differences between the original segment and the substituted segment are not corrected, although care was taken during recording of the diphones to maintain equal vocal effort.

The test database contains 166 word pairs. The reference words are mono-syllabic English words, chosen from three categories. Category I consists of words that end in a nasal (/n/, /m/ or /ng/). Category II contains words beginning with a glide (/w/, /r/, /l/ or /y/). Category III consists of words ending with a voiceless stop consonant (/p/, /t/, or /k/). For each category, four reference words are chosen, with central vowels /aa/, /ae/, /iy/ and /uw/, respectively. These vowels correspond with distinct tongue positions, and allow study of a large part of the vowel space.

Modified words are generated for each category. In category I, the second half of the center vowel is substituted with vowel segments preceeding different nasals in the database. For example, in the reference word "lamb", the second half of /ae/ is replaced by the second half of /ae/ in "fan". Yet another version for "lamb" is created by inserting the second half of /ae/ from "sang". For words of category II, the first half of the vowel is replaced by instances of that vowel following different glides. For words of category III, segments are inserted that preceed any of the plosives /p/, /b/, /t/, /d/, /k/, or /g/.

2.3 Procedure

There were fifteen participants in the perceptual test. For each participant, the same word pairs were used, but the order was randomized. Also the order of words within a pair was decided at random. Care was taken to select only native speakers of American English.

Participants were trained by listening to 25 random word pairs before the test began. They were asked to rate the differences between the word pairs on a five-point scale, and could listen to the words an arbitrary number of times

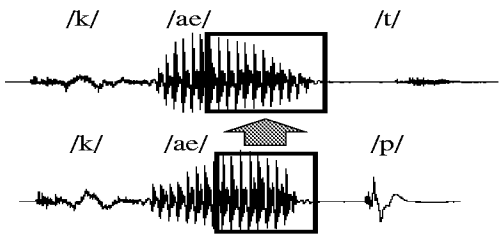

Figure 1. The pronunciation of a word is modified by substituting an allophonic speech segment.

before making a decision. The interface permitted the listeners to listen to the words individually, as well as in sequence. Study of the responses showed that all listeners had used the entire scale, except one subject, whose data were rejected. All listeners agreed that judging small perceptual differences was a difficult task, but felt that they had been able to make consistent decisions after the initial training.

As a validity check, we studied the responses for a group of 38 control pairs, in which no segment was altered (i.e. the words were identical). Almost all responses were 0 or 1 (on a scale from 0 to 4), 1.5 % of the responses were 2, and 3 was selected once.

3 EVALUATION OF DISTANCE MEASURES

The experiment described in Section 2, measures the perceptual effect of inserting certain allophones in a new context. Our goal is to investigate the potential of objective distance measures to predict these effects. Since the words in a pair are identical except for one segment, only the objective distance between the original and the substituted segment is used for the prediction. In general, the spectral discontinuity between a newly inserted segment and the remainder of the word also has a perceptual effect (i.e. concatenation cost). However, we chose the reference words carefully, so that most inserted segments represent the second half of a vowel, followed by a stop consonant (category III, see Section 2.2), or a nasal stop (category I). Concatenation effects can be minimized in such cases [1]. For the segments preceeded by a glide (category II), we neglect concatenation effects, and investigate how well the perceptual changes can be predicted only by objective distances between segments.

We define the *perceptual distance* between the words in one pair as the average of the listeners' responses for that pair. Since only listeners that had used the full answer scale were retained in the database, no further normalization of listeners was undertaken. The correlation between objective distances and perceptual distances is used to evaluate the objective measures. However, a simple correlation between perceptual and objective distances gives results close to 0. When computed per vowel category, more meaningful correlations are found. Note that for each category, four types of segments were studied (corresponding to the vowels /aa/, /ae/, /iy/ and /uw/, see Section 2.2). Hence, we divided the word pairs into twelve subgroups, corresponding to each category and each vowel. The correlation between objective and perceptual distances was computed per subgroup, yielding 12 coefficients. These coefficients were then combined in a "population correlation,"

	linear	PLP	mel
FFT cepstra	0.49	0.62	0.64
LPC cepstra	0.48	0.61	0.64
LSF	0.34	0.57	0.58
log area	0.28	0.55	0.52
Itakura	0.50	0.61	0.64

Table 1. Correlation between perceptual distances and objective measures based on different feature representations.

using Fischer's transformation. The population correlation is used as a measure of goodness for the objective distance measures.

4 RESULTS

We report the correlations of several objective distance measures with the perceptual distances, as defined above. The distance measures fit in the following framework. First, the raw speech signal, sampled at 16 kHz, is converted into a stream of feature vectors (frames), extracted at 5 millisecond intervals. Since segments may have different durations[1], resulting in an unequal number of frames, the time scale of the second segment is adjusted linearly, and new frames are calculated by interpolating the original frames at the new extraction points. Distances between corresponding frames can then be calculated. Finally, the frame distances are combined into a global distance between segments.

4.1 Choice of Features

We studied five feature representations: FFT-based cepstra, LPC-based cepstra, line spectral frequencies (LSF), log area ratios (LAR) and a symmetrized Itakura distance.

All but the FFT-based cepstra were computed via linear predictive coding (LPC) coefficients. Hermansky [6] proposed to compute LPC coefficients from a "perceptual spectrum," using the Bark scale and equal loudness preemphasis. The analysis was called perceptual linear prediction (PLP). On the other hand, current recognition systems often employ mel cepstral coefficients, obtained by taking the inverse FFT of a mel-warped spectrum. In our experiments, we decided to compute the feature representations in three different ways: (1) using the FFT amplitude spectrum, (2) using a perceptual spectrum as described in [6], (3) using a mel-warped spectrum.

From Table 1 it can be seen that the PLP and mel scales improve the correlation between objective measures and the perceptual data. Mel based distances slightly outperform the PLP distances. However, the confidence boundaries of the correlations are approximately ± 0.05, due to the limited amount of perceptual data. Hence the differences between PLP and mel distances are not statistically significant ($p > 0.84$, two-sided). When we inspect Table 1 columnwise, we see that the cepstral measures and the Itakura distance perform best.

4.2 Weighted Distances

Cepstral Liftering

A weighted Euclidean distance can be used to measure the distance between two feature vectors. Placing weights on cepstral coefficients is called "liftering".

Figure 2 summarizes experiments with an exponential cepstral lifter, as described by Hermansky and Junqua [7]. The idea is to weight each cepstral coefficient c_i with a factor i^s, where i is the index and s is a parameter. Hermansky and Junqua reported an optimal value of $s = 0.6$ for "regular" cepstral coefficients, which is confirmed by our perceptual experiments. However, for PLP and mel cepstra, only very small increases in the correlation coefficient can be noted.

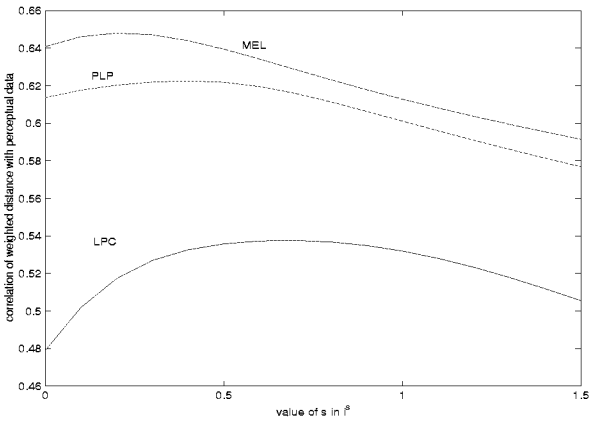

Figure 2. Effect of index weighting on correlation of distance measure with perceptual distances.

	linear		mel	
	Eucl	Mah	Eucl	Mah
cep	0.48	0.53	0.64	0.64
lsf	0.34	0.50	0.58	0.57

Table 2. Evaluation of cepstrum and LSF with Euclidean metric (left) and Mahalanobis metric(right). The value in the table is the correlation of the distance measure with the perceptual data.

Mahalanobis and Optimal Weighting

The Mahalanobis distance is based on weighting features with the inverse of their variance. Features with low variance are boosted, and have a better chance of influencing the total distance. For speech cepstra, index weighting is an approximation of the Mahalanobis distance [5]. In the general case, the Mahanalobis distance also involves estimation of feature covariances. Because the covariances cannot be estimated reliably from limited speech data, they are usually ignored. This corresponds with assuming that the features are uncorrelated.

In Table 2, we compare the Mahalanobis distances for cepstra and LSF measures. The variances of the feature vectors are calculated over the entire database. We find that the Mahalanobis distance gives an improvement for the linear frequency measures, but does not improve the correlation for the mel based measures. However, we believe that with more speech data, more reliable variances could be calculated for each vowel context, which could improve the performance of the Mahalanobis distance.

Ultimately, the weights could be optimized with an iterative search. We have not been very successful in this approach. For mel cepstra, the correlation can be maximized up to 0.68 starting from different initializations. This is not a big increase. Moreover, the weights do not form a pattern that can be easily interpreted. Our LSF feature representation consisted of the sums and differences of spectral pairs, which can be interpreted as spectral poles and bandwidths [15]. The optimized weights seemed to favor the middle poles of the LSF representation, and to

[1] Segments form part of vowels with equal duration, but the duration of a segment within a phoneme can be different.

	baseline	only Δ	combined
mel cepstra	0.64	0.64	0.66
mel LSF	0.58	0.53	0.59

Table 3. Correlations for mel cepstra and for mel LSF. The first column gives the baseline, the second column shows the correlation for delta features only, and the third column gives the result for a combined measure

attach less importance to the bandwidths.

Time Scale Weighting

Several frame distances need to be combined in order to obtain a distance between segments. In DTW algorithms, frame distances were sometimes weighted according to the amount of time warping needed. In a distance measure developed by Sondhi and Ghitza [4], frames are weighted more heavily towards phoneme boundaries.

In the results reported so far, we defined the distance between two segments as the average of their frame distances. Another choice is to take the maximum frame distance. This is motivated by the intuition that peak differences may perceptually be more important than global differences. However, from experiments with mel cepstra, mel LSF and mel Itakura distances, we concluded that taking the maximum frame distance did not lead to objective measures that correlated better with perceptual distances.

4.3 Delta Features

The features explored so far reflect the static frequency characteristics of speech at a certain point in time. *Delta features* are estimations of the time derivatives of static features, thus capturing more of the speech dynamics.

In Table 3, we summarize experiments with distances based on mel cepstra and mel LSF. Using delta features only, we find correlations that approximate results with regular features. When the delta features are combined with the original features, a small increase in correlation is found. Since the delta features are much smaller than the original coefficients, the average amplitude of both feature sets was normalized. We also trained an additional weight to optimize the contribution of delta features in the frame distance, but this improved the correlations only marginally.

It is disappointing that although the feature representation is doubled in size, the effect on the distance measure is quite small. Further research is needed to incorporate delta features more successfully into distance measures.

5 CONCLUSIONS

We have developed a test to measure perceptual differences between allophones. The collected data have allowed us to study the perceptual relevance of distance measures for speech. We found mel-based cepstral and Itakura distances to be the most powerful, with minimal added benefit found by utilizing weighted distances or delta features.

The results of our study give insights into the use of distance measures for unit selection. We have provided a perceptual validation of distance measures, showing that a reasonable (0.66) correlation with perceptual distances exists. On the other hand, this correlation is not high enough to consider distance measures as reliable predictors of perceptual differences. This was confirmed in a classification experiment, where we used the best objective measure to decide whether the words in a pair were either similar (perceptual score ≤ 1) or different (perceptual score ≥ 2). Classification errors ranged between 20% and 50%, for various decision points.

In this paper, we have evaluated some of the better known distance measures. It is a challenge for future work to study other measures, and to develop new ideas to improve the accuracy of objective distance measures.

REFERENCES

[1] P. Bhaskararao. Subphonemic segment inventories for concatenative speech synthesis. In *Fundamentals of Speech Synthesis and Speech Recognition*, pages 70–85. John Wiley & Sons Ltd, 1994.

[2] A. W. Black and N. Campbell. Optimising selection of units from speech databases for concatenative synthesis. In *ESCA Eurospeech'95*, pages 581–584, 1995.

[3] R. E. Donovan. *Trainable Speech Synthesis*. PhD thesis, Cambridge Univ. Eng. Dept., June 1996.

[4] O. Ghitza and M. M. Sondhi. On the perceptual distance between speech segments. *J. of the Acoustical Soc. of Am.*, 101(1), 1997.

[5] A. H. Gray and J. D. Markel. Distance measures for speech processing. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 24(5):380–391, October 1976.

[6] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *J. of the Acoustical Soc. of Am.*, 87(4):1738–1752, April 1990.

[7] H. Hermansky and J. C. Junqua. Optimization of perceptually-based ASR front-end. *Proc. IEEE ICASSP*, 9:219–222, 1988.

[8] A. J. Hunt and A. W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *ICASSP'96, Atlanta*, 1996.

[9] D. Kewley-Port and B. S. Atal. Perceptual differences between vowels located in a limited phonetic space. *J. of the Acoustical Soc. of Am.*, 85(4), 1989.

[10] D. Klatt. Prediction of perceived phonetic distance from critical-band spectra: A first step. In *Proc. IEEE ICASSP*, pages 1278–1281, 1982.

[11] S. Krishnan and P. Rao. A comparative study of explicit frequency and conventional signal representations for speech recognition. *Digital Signal Processing*, 6:249–284, 1996.

[12] M. Macon, A. Cronk, J. Wouters, and A. Kain. OGIresLPC: Diphone synthesizer using residual-excited linear prediction. Technical Report CSE-97-007, OGI, September 1997.

[13] N. Nocerino, F. K. Soong, L. Rabiner, and D. Klatt. Comparative study of several distortion measures for speech recognition. *Speech Communication*, 4:317–331, 1985.

[14] S. R. Quackenbush, T. P. Barnwell III, and M. A. Clements. *Objective Measures of Speech Quality*. Prentice Hall, Englewood Cliffs, New Jersey, 1988.

[15] F. K. Soong and B.-H. Juang. Line spectrum pairs (LSP) and speech data compression. In *Proc. IEEE ICASSP*, pages 1.10.1–1.10.4, 1984.

[16] J. van Santen. Prosodic modeling in text-to-speech synthesis. In *Proc. Eurospeech-97*, 1997.