PERCEPTUAL AND OBJECTIVE DETECTION OF DISCONTINUITIES IN CONCATENATIVE SPEECH SYNTHESIS

Yannis Stylianou and Ann K. Syrdal

AT&T Labs-Research, SIPS, Shannon Laboratories, 180 Park Avenue, Florham Park, NJ 07932, U.S.A.
email : {yannis, syrdal }@research.att.com

ABSTRACT

Concatenative speech synthesis systems attempt to minimize audible signal discontinuities between two successive concatenated units. An objective distance measure which is able to predict audible discontinuities is therefore very important, particularly in unit selection synthesis, for which units are selected from among a large inventory at run time. In this paper, we describe a perceptual test to measure the detection rate of concatenation discontinuity by humans, and then we evaluate 13 different objective distance measures based on their ability to predict the human results. Criteria used to classify these distances include the detection rate, the Bhattacharyya measure of separability of two distributions, and Receiver Operating Characteristic (ROC) curves. Results show that the Kullback-Leibler distance on power spectra has the higher detection rate followed by the Euclidean distance on Mel-Frequency Cepstral Coefficients (MFCC).

1. INTRODUCTION

Many speech synthesis systems today are based on non-uniform unit concatenation. In an effort to minimize audible signal discontinuities at the concatenation points, these systems try to select units from large speech databases in an optimum way [1] [2] [3]. For instance, in [1], a target cost and a concatenation cost are attributed to each candidate unit. Target cost is calculated as the weighted sum of the differences between prosodic and phonetic parameters and contexts of the target and candidate units. Concatenation cost is intended to achieve a smooth fit between two successive acoustic units. It is determined by the weighted sum of cepstral distance at the point of concatenation and the absolute differences in log power and f_0. The total cost for a sequence of units is the sum of the target and concatenation costs. Unit selection is then performed by a Viterbi search for the lowest cost path through the lattice of candidate units. Recent studies [4] [5] have attempted to determine which concatenation cost distance measures are best able to predict audible discontinuities. Units that are predicted to produce audible discontinuities in concatenation will be assigned higher concatenation costs, and thus they will be less likely selected.

Distance measures have many applications in other speech technologies. For speech coding, they are applied in the design of vector quantization algorithms and as objective measures of speech quality [6]. In speech and speaker recognition, measuring the spectral difference between two speech patterns is used to compare patterns and make similarity decisions [7]. A study comparing several distance measures, primarily for the above speech technologies, was conducted by Gray and Markel [8]. The best performance was obtained with the Root-Mean-Squared Log Spectral Distance. Nocerino et al. [9] found that perceptually motivated warped frequency scales (such as Mel and Bark scales) did not improve the performances of speech recognition systems, while Hermansky and Junqua [10] found the opposite. Currently, the most widely used distance in speech recognition is the Euclidean distance between MFCCs.

Motivated by speech recognition methods, some speech synthesis unit selection algorithms [1] and optimal join algorithms [11] use the Euclidean distance between MFCCs. However, because of the differing objectives of speech synthesis and speech recognition or coding, a perceptual evaluation of distance measures for concatenative speech synthesis and their ability to predict audible discontinuities have been investigated recently. Klabbers and Veldhuis [4] found that the Kullback-Leibler distance [12] on LPC power spectra was the best predictor of discontinuities. Wouters and Macon [5] found that the Euclidean distance on mel-scale LPC-based cepstral coefficients was a good predictor.

In this paper, we present a psychoacoustic experiment on the detection of signal discontinuities by humans in our speech database of a female voice and an evaluation, based on these perceptual data, of several measures of spectral distance. We conducted a two-part experiment. The first phase was a psychoacoustic experiment on listeners' detection of concatenation discontinuities in a large number of test words generated by concatenative synthesis. In the second phase, we compared concatenation cost estimates derived from various algorithms with the listeners' detection results. Since concatenation costs are intended to estimate the perceptual salience of concatenation discontinuities, our experiment constitutes a rigorous evaluation of the validity of various concatenation cost algorithms used for unit selection in text-to-speech synthesis.

The paper is organized as follows. Section 2 describes the perceptual experiment is described. In Section 3 we present the feature representations and the distances that were compared. Results from the perceptual experiment and an evaluation of various distance measures are presented in Section 4. A summary and a discussion of the obtained results and of future work concludes the paper.

2. PERCEPTUAL EXPERIMENT

2.1. Test Stimuli

A set of 2016 monosyllabic test words were generated by concatenative synthesis using an acoustic inventory of recordings from one adult female speaker. An experimental version of the AT&T Next-Generation text-to-speech (TTS) synthesizer [3] was used to synthesize the test stimuli. AT&T's TTS system is based on an extension of the unit selection algorithm of the CHATR synthesis system [1], and it is implemented within the framework of the Festival Speech Synthesis System [13].

The acoustic inventory used for synthesis consisted entirely of recordings of the 336 monosyllabic test words that constitute the Modified Rhyme Test (MRT)[14][15][16], a standard test of speech intelligibility [17]. The MRT is composed of 56 sets of 6 similar words. The 6 words within a set differ by either the initial consonant(s) (such as "book, took, shook, cook, hook, look") or the final consonant(s) (such as"dent, bent, went, tent, rent, sent"), and all words in a set contain the same vowel nucleus. A restricted domain voice was built with the MRT inventory for the AT&T TTS system.

The 2016 synthetic test words were synthesized by concatenation of selected portions of the 336 recorded words contained in the acoustic inventory. Each recorded word in the inventory was essentially divided into two parts, its initial and final halves. The initial half consisted of the word-initial consonant(s) and the first half of the vowel nucleus. The second half consisted of the second half of the vowel nucleus and the word-final consonant(s). For each of the 56 6-word sets, 36 test stimuli were synthesized. All possible combinations of the 6 initial halves and 6 final halves within a set were concatenated to generate 36 synthetic test words. Of the 36 test words synthesized from each 6-word set, 30 combined the first half of a word with the second half of a different word, and these 30 test words had the potential of containing detectable concatenation discontinuities. Six of the 36 test words synthesized per set were resynthesized versions of the first and second halves of the same word, and they would be expected to contain no detectable concatenation discontinuities.

An extremely simple concatenation method was used by the synthesizer to concatenate the first and second halves of words at approximately the mid-point of the vowel. Using the raw waveforms, the concatenation point was determined by a minimum in the cross-correlation function calculated over a narrow window around the vowel mid-points. In this way, concatenation discontinuities due simply to arbitrary abutment of the two halves was avoided, and pitch period continuity was maintained.

2.2. Perceptual Test Procedure

The listening test followed a simple single interval forced choice (Yes/No) signal detection paradigm [18] commonly used in psychoacoustic experiments. After hearing a test stimulus, a listener reported whether or not (s)he heard a concatenation discontinuity. Each stimulus was presented once per listener. The entire test battery was divided into a series of subtests; each subtest contained 72 test stimuli and normally took under 10 minutes to complete. Each listener received a different randomization of the stimuli in a subtest. Typically, a listener would participate in no more than one subtest a day. Written instructions to listeners and one example of a stimulus for each response type (a detectable concatenation discontinuity and no discontinuity) were provided at the beginning of a subtest. Listeners were automatically prompted if they did not complete any part of the subtest, and their complete response record was stored in a log file identifiable by listener and subtest.

Listening tests were web-based and interactive. Listeners normally took the tests from workstations or PCs in their quiet private walled offices using the relatively high quality audio equipment normally available there. Listeners initiated the presentation of each stimulus by clicking an icon. Concatenation detection responses were made by clicking one of two radio buttons (one indicating that a discontinuity was detected, and the other, that no discontinuity was detected). Listeners were encouraged to use headphones, and the large majority indicated that they did so. The volume was adjusted to suit their individual preferences. Stimuli were sampled at 16 kHz.

2.3. Listeners

Sixteen adult volunteer listeners participated in at least one listening subtest. The average number of subtests per listener was 10. All listeners were employees or contractors working at AT&T Labs Research. They represented diverse language backgrounds, since native language was not considered relevant for the auditory task of detecting concatenation discontinuities. The hit rate, false alarm rate, and *d'* [18] (an index of detectability) per subtest were monitored for each listener. Rarely (5% of the time), a listener's responses were rejected for a particular subtest if their *d'* score was substantially lower than the other listeners' *d'* scores for that subtest. There were at least five acceptable listeners for every stimulus word in the test set, and the average was 5.9 acceptable listeners per stimulus. There were 11,808 total acceptable observations in the entire listening test.

3. SPECTRAL DISTANCE MEASURES

The distance measures used in this paper were the following:

1. The Euclidean distance between Log Power Spectra computed from a) FFT (D1), b) LPC (D2) and c) Perceptual Linear Prediction, PLP [19], (D3).
2. The Euclidean distance between Line Spectrum Frequencies (LSFs) computed from a) LPC (D4) and b) PLP (D5).
3. The Weighted Euclidean distance between cepstral coefficients computed from a) LPC (D6) and b) PLP (D7).
4. The Euclidean distance between Mel-Frequency Cepstral Coefficients (MFCC) (D8).
5. The Kullback-Leibler distance between Power Spectra computed from a) FFT (D9), b) LPC (D10) and c) PLP (D11).
6. The Kullback-Leibler distance between LSFs computed from a) LPC (D12) and b) PLP (D13).

We tested various other features and distances with less interesting results than the ones listed in the paper. For example, we tested the sine parameters ($arcsin(k_i)$, where k_i are the reflection coefficients) and the Log Area parameters. Other distances were the Cosh distance, the Itakura-Saito 1 and 2, and the Itakura distance. The Kullback-Leibler distance (D_{KL}) is used to compute the distance (or divergence) between two probability distributions. Here we use it in a way similar to [4]; for instance, if $P(\omega)$ and $Q(\omega)$ are two power normalized spectra, then D_{KL} is defined as:

$$D_{KL} = \int (P(\omega) - Q(\omega)) \log \frac{P(\omega)}{Q(\omega)} d\omega \qquad (1)$$

The weights for the cepstral coefficients obtained from LPC (LPCC) were the warping parameters given in [7] (Table 4.3, p.189) which warp the linear frequency scale to Bark scale. The weights for cepstral coefficients obtained from PLP (PLPCC) were an exponential cepstral lifter, as described in [10]; this is to weight each cepstral coefficient, c_i, with i^s, where s was set to 2.0 [10] (otherwise refered to as group delay distortion measure). The MFCC were computed in the way described in [7] (pp. 186-189). The first cepstral coefficient from all the cepstral formats (i.e., MFCC, LPCC and PLPCC) was excluded from the distance calculation. The order of LPCC and MFCC was 20 while the order of PLPCC was 5. For each unit, one speech frame of $40ms.$ at the concatenation point was obtained and an FFT of size 1024 was computed. All speech frames were normalized before any transformation ((i.e., LPC, FFT, etc.) was applied. The distances (D1-D13) were only computed for test words in which the first and second halves were taken from different words, since in the case of concatenated halves of the same word, these distances are either zero or not defined (D_{KL}).

The evaluation of the distance measures was based on three criteria:

1. The detection rate, P_D, when the false alarm rate, P_{FA} was set to 5%.
2. The Bhattacharyya distance, B_d[20]:

$$B_d = \frac{1}{8}(\mu_2 - \mu_1)^2\,\frac{\sigma_1^2+\sigma_2^2}{2} + \frac{1}{2}\,\ln\frac{\frac{\sigma_1^2+\sigma_2^2}{2}}{\sqrt{\sigma_1^2\,\sigma_2^2}} \qquad (2)$$

   which is a measure of the separability of two distributions (not necessarily for normal only distributions).
3. The Receiver Operating Characteristic (ROC) curve

For each distance measure, Dx, two probability density functions, $p(Dx|0)$ and $p(Dx|1)$ were computed depending on the results from the perceptual test; if the synthetic sentence was perceived as continuous (0) or discontinuous (1) by the listeners. Then the detection rate for that distance, Dx, is computed as:

$$P_D(\gamma) = \int_{\gamma}^{\infty} p(Dx|1)dDx \qquad (3)$$

where γ is defined by:

$$P_{FA}(\gamma) = \int_{\gamma}^{\infty} p(Dx|0)dDx = 0.05 \qquad (4)$$

A plot of pairs $\{P_D(\gamma),\ P_{FA}(\gamma)\}$ for all values of γ constitutes an ROC curve.

4. RESULTS

4.1. Perceptual Test Results

Pooling all the acceptable listeners' responses, the group correct detection rate was 61.4% and the false alarm rate (the incorrect detection rate for test words concatenated from the first and second halves of the same word) was 6.1%. These results yield a *d'* score of 1.83, representing overall human perceptual performance. Note that the nature of the detection test was somewhat different for evaluating the distance algorithms than it was for humans. For the purposes of the distance measure evaluation, the human detection rate defined whether a test word did or did not contain an audible discontinuity. That is, the 61.4% detection rate by human listeners was equivalent to a 100% detection rate by the algorithms. The 38.6% of test words concatenated from different words but for which listeners could not detect discontinuities were used to determine false alarm statistics for algorithm evaluation.

4.2. Evaluation of Concatenation Cost Estimation Algorithms

In Table 1 the evaluation of the distance measures by the first two criteria (P_D and B_d) is reported. The distances are sorted in descending order by the detection rate. Note that the false alarm was set to 5%. As seen in Table 1, none of the detection rates are very high. The highest rate is obtained by the Kullback-Leibler distance on the FFT-based power spectra (D9), followed by the Euclidean distance between MFCC (D8). If we look only at the power spectra, we see that the non-parametric form (FFT-based) is the winner in both distances (Kullback-Leibler (D9) and Euclidean (D1)), followed by the Kullback-Leibler distance on PLP-based power spectra (D11) and as fourth is the Euclidean distance on LPC-based spectra (D2). The Kullback-Leibler distance on the normalized LPC-based spectra (D10) is among the poorer performers on the list. Therefore, although the Kullback-Leibler distance predicts better than the Euclidean distance for FFT-based power spectra, it is worse than the Euclidean distance for LPC-based power spectra. Among various parametric forms of power spectra, the Euclidean distance of MFCC (D8) yields the best score, and it's a close second to the one obtained by the Kullback-Leibler distance for FFT-based power spectra. The second best parametric form is the PLP-based cepstral coefficients (D7) while very close to this one is the LPC-based cepstral coefficients (D6). For these two parametric forms, the weighted Euclidean distance (D7) performs similarly to the Kullback-Leibler distance (D11). In addition to the distance measures reported here, we also tested the absolute difference of pitch around the concatenation points, which had a detection rate of 19.981%.

Distance	P_D %	B_d
D9	37.162	0.237
D8	35.811	0.187
D1	28.764	0.208
D7	25.579	0.154
D6	23.263	0.088
D12	23.166	0.077
D11	22.780	0.137
D2	21.429	0.070
D5	21.139	0.141
D3	20.946	0.137
D13	19.305	0.157
D10	18.243	0.105
D4	9.749	0.025

Table 1: Evaluation of Concatenation Cost Estimation Algorithms

The ROC curves for the first three best distances are depicted in Figure. 1. For comparison purposes, the worst distance (D4) is also shown. Although the results obtained in this paper are based on different speech data than those used in other experiments [5] [4], it is interesting to compare our results with previously published research. Our results (except for the Itakura distance) seem to be in accordance with those obtained by Wouters and Macon [5], in that the Euclidean distance for MFCC performs very well. On the other hand, our results are partially in agreement with those obtained by Klabbers and Veldhuis [4], since the Kullback-Leibler distance is a good predictor for audible signal discontinuities. However, in our study, it performs similarly to the Euclidean distance for MFCC, in contrast to what was observed in [4]. Furthermore, this only holds for the FFT-based power spectra; the Kullback-Leibler distance based on LPC-based spectra

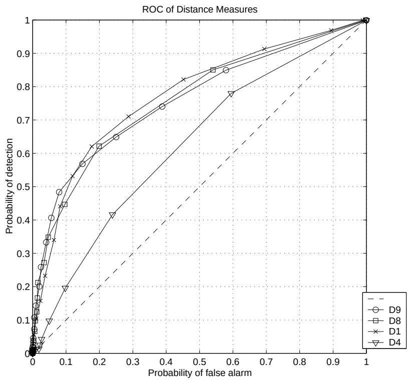

Figure 1: ROC curves for the first three disances D9 (circle), D8 (square) and D1 (x). For comparison, the worst distance, D4, is also depicted (triangle).

(D10) performs worse than the corresponding Euclidean distance (D2).

It is obvious from the above results that the evaluated spectral distance measures cannot predict very well the results from humans. The best distance measure predicts only 37% of the audible signal discontinuities. Therefore, there is a need for further investigations of new distances and new features. It will be interesting to also explore various combinations of distances.

5. SUMMARY AND CONCLUSIONS

This paper has presented a two part experiment. First, a psychoacoustic experiment was conducted on the detectability of signal discontinuities in concatenative speech synthesis by humans. Based on the perceptual results obtained, we have compared the ability of many distance measures to predict audible signal discontinuities. We have found that the Kullback-Leibler distance between FFT-based power spectra and the Euclidean distance between MFCC have the highest prediction rates. However, even the best obtained prediction score cannot be considered high. Further investigation (with additional data from more speakers) of new distances or combinations of distances and an exploration of new speech features better characterizing the phenomena during the concatenation of two units is of considerable importance for a high quality speech synthesis system. The study also has implications for extending our understanding of the human auditory perception of speech.

6. REFERENCES

[1] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using large speech database," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 373–376, 1996.

[2] W. N. Campbell and A. Black, "Prosody and the selection of source units for concatenative synthesis," in *Progress in Speech Synthesis*, R. Van Santen, R.Sproat, J.Hirschberg, and J.Olive, Eds. 1996, pp. 279–292, Springer Verlag.

[3] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T Next-Gen TTS System.," *137th meeting of the Acoustical Society of America*, 1999, http://www.research.att.com/projects/tts.

[4] E. Klabbers and R. Veldhuis, "On the reduction of concatenation artefacts in diphone synthesis," *International Conference on Spoken Language Processing ICSLP 98*, pp. 1983–1986, 1998.

[5] J. Wouters and M. Macon, "Perceptual evaluation of distance measures for concatenative speech synthesis," *International Conference on Spoken Language Processing ICSLP 98*, pp. 2747–2750, 1998.

[6] S. Quackenbush, T. P. Barnwell III, and M. A. Clements, *Objective Measures of Speech Quality*, Prentice-Hall, Englewood Cliffs, New Jersey 07632, 1988.

[7] L. R. Rabiner and B-H. Juang, *Fundamentals of speech recognition*, PTR Prentice-Hall, 1993.

[8] A. H. Gray and J. D. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, no. 5, pp. 380–391, Oct 1976.

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

[10] H. Hermansky and J. C. Junqua, "Optimization of perceptually-based asr front-end," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1988, pp. 219–222.

[11] A. Conkie and S. Isard, "Optimal coupling of diphones.," in *Progress in Speech Synthesis*, R. Van Santen, R.Sproat, J.Hirschberg, and J.Olive, Eds. 1996, pp. 293–304, Springer Verlag.

[12] S. Kullback and R. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 79–86, 1951.

[13] A. Black and P. Taylor, "The Festival Speech Synthesis System: system documentation," *Technical Report HCHC/TR-83*, 1997.

[14] A. S. House, C.E. Williams, M. H. L. Hecker, and K. D. Kryter, "Psychoacoustic speech tests: A Modified Rhyme Test," Tech. Doc. Rept. ESD-TDR-63-403, U. S. Air Force Systems Command, Hanscom Field, Electronics Systems Division, June 1963.

[15] A. S. House, C. E. Williams, M. H. L. Hecker, and K. D. Kryter, "Articulation testing methods: Consonantal differentiation with a closed-response set," *J. Acoust. Soc. Amer.*, vol. 37, pp. 158–166, 1965.

[16] E. J. Kreul, J. C. Nixon, K. D. Kryter, D. W. Bell, J. S. Lang, and E. D. Schubert, "A proposed clinical test of speech discrimination," *J. Speech and Hearing Research*, vol. 11, pp. 536–552, 1968.

[17] American National Standards Institute, "Method for measuring the intelligibility of speech over communication systems," Revised Standards Report ANSI S3.2-1989 - A revision of ANSI S3.2-1960, American Standards Association, New York, 1989.

[18] J.A.Swets, *Signal detection and recognition by human observers: Contemporary readings*, Peninsula Press, 1988.

[19] H. Hermansky, "Perceptual Linear Predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[20] Keinosuke Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 1990.