

A New Distance Measure for Costing Spectral Discontinuities in Concatenative Speech Synthesisers

Robert E. Donovan

IBM T.J.Watson Research Center, PO Box 218
Yorktown Heights, New York, 10598, USA

red@watson.ibm.com

Abstract

In many modern concatenative speech synthesisers the unit sequence used to synthesise each sentence is determined at runtime by a search algorithm seeking to optimise a multi-dimensional cost function. One of these costs is usually some form of spectral continuity cost, computed between the end of one segment and the start of the following segment, intended to ensure that the synthetic speech does not contain any unpleasant spectral discontinuities. This paper presents the results of listening tests conducted to evaluate the performance of several possible continuity measures. It also describes a new continuity measure developed at IBM which substantially out-performs all other measures tested.

1. Introduction

Data-driven “unit-selection” concatenative approaches to speech synthesis have become increasingly popular in recent years, [1], [2], [3], [4], [5], [6], [7]. In these systems speech is synthesised by concatenating units selected from a database typically containing many thousands of units. The selection is usually made using a dynamic programming search to optimise a cost function given some target specification of the sentence to be synthesised. Typically two types of cost are employed, being target costs and concatenation costs. The target costs reflect how well a segment’s phonetic and wider context and intrinsic prosody match those required in the synthetic sentence. The concatenation costs are applied between potentially adjacent segments, and usually include some form of spectral continuity cost, and sometimes a pitch continuity cost.

This paper is concerned with the spectral continuity cost used to determine how smoothly two segments will concatenate during synthesis. Various measures were used in the systems mentioned above, often based on distances between cepstra or mel-cepstra, but in most cases little work appears to have been done to determine the usefulness of the measure selected. In [8] however, Klabbers & Veldhuis evaluated a number of likely measures by comparing the numerical distances computed with them to listener ratings of discontinuities in a large number of concatenated stimuli. The measures included the Euclidean distance between F_1 and F_2 (formant frequency) pairs, the Kullback-Leibler distance, the Euclidean distance between mel-frequency cepstral coefficients (MFCCs), the likelihood ratio, the mean-squared log-spectral distance, the loudness difference and the excitation difference. The results indicated that the Kullback-Leibler distance was the best measure tested by a fairly large margin, and that the Euclidean distance between MFCCs performed only a little better than chance.

The results obtained by Klabbers & Veldhuis prompted a

similar study to be undertaken at IBM, and consequently to the development of a new measure designed to give better correlation with human perception of join quality. This paper describes the new algorithm in detail, and reports on the listening tests conducted to determine its effectiveness.

2. New Distance Measure

Preliminary listening tests indicated that neither the Kullback-Leibler measure recommended by Klabbers & Veldhuis, nor the cepstral Mahalanobis distance (which they did not test) provided very impressive correlations with human perception of join quality, and research was therefore undertaken to try and find something better. Analysis of a large number of novel distance measures led to the observation that different changes in the underlying spectral vectors appeared to be allowable in different contexts. This prompted an analysis of the natural variation of log power spectra across segment boundaries in different contexts in the speech database used in the IBM synthesis system, and hence to the conclusion that the new distance measure should be context dependent. The new distance measure can be described as a decision-tree-based context-dependent Mahalanobis distance between perceptual cepstral vectors. The rest of this section describes the new measure in more detail.

A major problem that needs to be overcome with any context dependent system is that of data-scarcity, that is, a lack of training data in some contexts. A common solution to this problem, and the one which was adopted here, is to use top-down decision trees to cluster the data according to phonetic context. The data to be clustered was prepared from the speech database used to build the synthesis system. This database is segmented into hidden Markov model state-sized segments as part of the synthesis system construction process, [5], and these segments are the basic concatenation units in synthesis. For each state boundary in the database the context vector consists of the identity of the current phone and the boundary location for within-phone boundaries, or the identities of the preceding and following phones for cross-phone boundaries. For each boundary the change in the cepstra from the frame at the end of one segment to the frame at the start of the next segment is also noted. The frames used were 25ms pitch synchronous frames through regions of voiced speech, and 6ms frames at a 3ms frame rate through unvoiced speech, using the scaling method described in [1] to ensure that cepstra from the different sized frames were comparable. These frames were used because they were used to perform the state alignment when constructing the synthesis system, [5]. The boundary data was clustered using a single decision tree, by asking broad class questions about the preceding and following phonetic identity and the location of the

boundary within the phone. The splitting criterion was the standard increase in log-likelihood criterion often used in speech recognition systems, [9], and the stopping criterion was to ensure at least 60 data points in each leaf. A mean vector and diagonal covariance matrix was estimated from the data in each leaf of the decision tree. The stopping criterion was established by examining the performance, in terms of correlation with human ratings of discontinuities, of distances (see Equation 1) calculated using unclustered data in place of the decision tree for contexts in which it was available, and noting that performance began to degrade when less than approximately 60 data points were available.

Given the decision trees, the distance measure between the vector e at the end of one segment and the vector s at the start of the next segment can be computed as

$$D^2 = \sum_{i=1}^n \left[\frac{e_i - s_i - \mu_i^l}{\sigma_i^l} \right]^2 \quad (1)$$

where n is the dimensionality of the data, μ_i^l is the i th element of the mean vector in leaf l , $(\sigma_i^l)^2$ is the i th diagonal element of the covariance matrix for leaf l , and leaf l is the leaf reached by descending the decision tree for the context that the join is in.

The cepstra referred to above are actually 12 dimensional perceptually-modified MFCCs. The perceptual modification to the standard MFCCs was introduced because otherwise a significant distance could be computed between two cepstral vectors which both represented inaudible frames. More generally it seems likely that discontinuities in loud sonorants should generate larger distances than discontinuities occurring in quieter phones. The solution to the problem which was adopted was to adjust the mel-binned log FFT vectors used during cepstra computation by subtracting, in the log domain, a value in each mel bin corresponding to the threshold of hearing at that frequency. The log FFT vectors were then thresholded at zero (ie. not allowed to go negative) since zero now corresponded to the threshold of hearing, and the cepstra computed in the usual way. Since the cepstral coefficients used do not include C0 the perceptual cepstra are identical to standard cepstra for speech which exceeds the threshold of hearing at all frequencies. For quieter speech the cepstra change, generally diminishing in magnitude, dropping to an all-zero cepstral vector for inaudible speech.

The computation of the perceptual cepstral vectors just described requires knowledge of the threshold of hearing as a function of frequency in terms of appropriate values to subtract from each mel-binned log FFT vector. Standard plots of human hearing threshold vs frequency are usually given in terms of dB, and therefore cannot be used for this purpose without being calibrated to the IBM log FFT generation software. In practice the required numbers were obtained directly at each frequency by playing sinusoids, with a logarithmic fade in and fade out, over loudspeakers in a soundproofed room to the author. The audio level was set such that speech from the synthesis training database was at a comfortable listening level, and the same audio level used to measure the thresholds of hearing. The sinusoidal waveforms so obtained were then processed to obtain the threshold of hearing equivalents in each log FFT mel bin. While using a single subject to determine these thresholds is not ideal, the author is confident that he has excellent hearing.

The use of perceptual cepstra, which occasionally contain all-zeros, can lead to difficulties with maximum likelihood tree building, and even more importantly with Equation 1, because leaf standard deviations can go to zero. In order to prevent these

problems it is necessary to enforce a standard deviation floor. A suitable floor was estimated as the amount of variation seen in a cepstral vector caused by perturbing the corresponding log FFT vector by a few dB in each dimension. The rationale for using this as a standard deviation floor vector is that the ear is insensitive to variations in amplitude below about 1dB, and thus the computed floor vector represents approximately the minimum perceivable change in cepstra. Since a leaf standard deviation represents the spread in the shift in cepstra seen across a boundary in that context, it is pointless to consider standard deviations below the point where that spread can be perceived.

3. Listening Tests

Informal listening indicated that the listening test stimuli needed to be short in order to avoid unnecessary distraction, and it was therefore decided to base the stimuli on consonant-vowel (CV) pairs. Approximately 400 CV pairs are possible in English, and a subset of 112 of these was selected containing an equal balance of different phone classes (eg. A-vowels, I-vowels, nasals, voiced-fricatives, etc). The stimuli were synthesised in a male voice using a modified form of the IBM Trainable Speech Synthesis System [5]. The acoustic decision trees had only a single leaf in order that the segments concatenated came from potentially inappropriate contexts to encourage the occurrence of spectral discontinuities in the stimuli. The continuity cost function was altered so that each stimulus was synthesised from exactly two contiguous chunks of speech. The join was usually placed at the phone transition location, except for stimuli beginning with voiced plosives in which the join was one third of the way into the vowel. Synthesis duration and energy were obtained using decision trees, and the synthesis pitch was constant. Listening to the stimuli it seemed that the full CV waveform was perhaps still adding too much distraction to the task, since the vowel was often much longer than the consonant. The stimuli were therefore edited to retain only the 120ms (or less if necessary) of speech centered on the join.

The seven listening test subjects were members of the IBM speech group, but most had no significant experience listening to synthetic speech. They were told that the stimuli they were about to hear contained one join and that their task was to rate the quality of that join for naturalness on a scale of 1 (good) to 5 (bad). In order to give the listeners an idea of the range of quality they were about to hear they were told that the first two stimuli presented were, in the author's opinion, a 1 and a 5 respectively. The ratings for these two stimuli were not used to compute any results. The tests were conducted over loudspeakers in a soundproofed room. The listeners could play each stimulus as many times as they liked.

Unfortunately the ratings given by the listeners did not correlate well with each other. Even after removing one listener who correlated particularly poorly with everyone else, the mean listener to listener correlation was only 0.34. This number is disappointing, but perhaps reflects the difficulty of evaluating the quality of extremely short segments of speech, especially for inexperienced listeners; many listeners commented that they found the test very difficult. Mean ratings averaged over several listeners improve in reliability, however. The average correlation of a listener with the mean ratings computed from the other five listeners is 0.49. A group rating was therefore computed as the mean ratings of the six listeners.

In addition to the tests described above, the author also conducted a listening test on himself. The author's ratings correlated with the group rating above at 0.51. Unlike the subjects

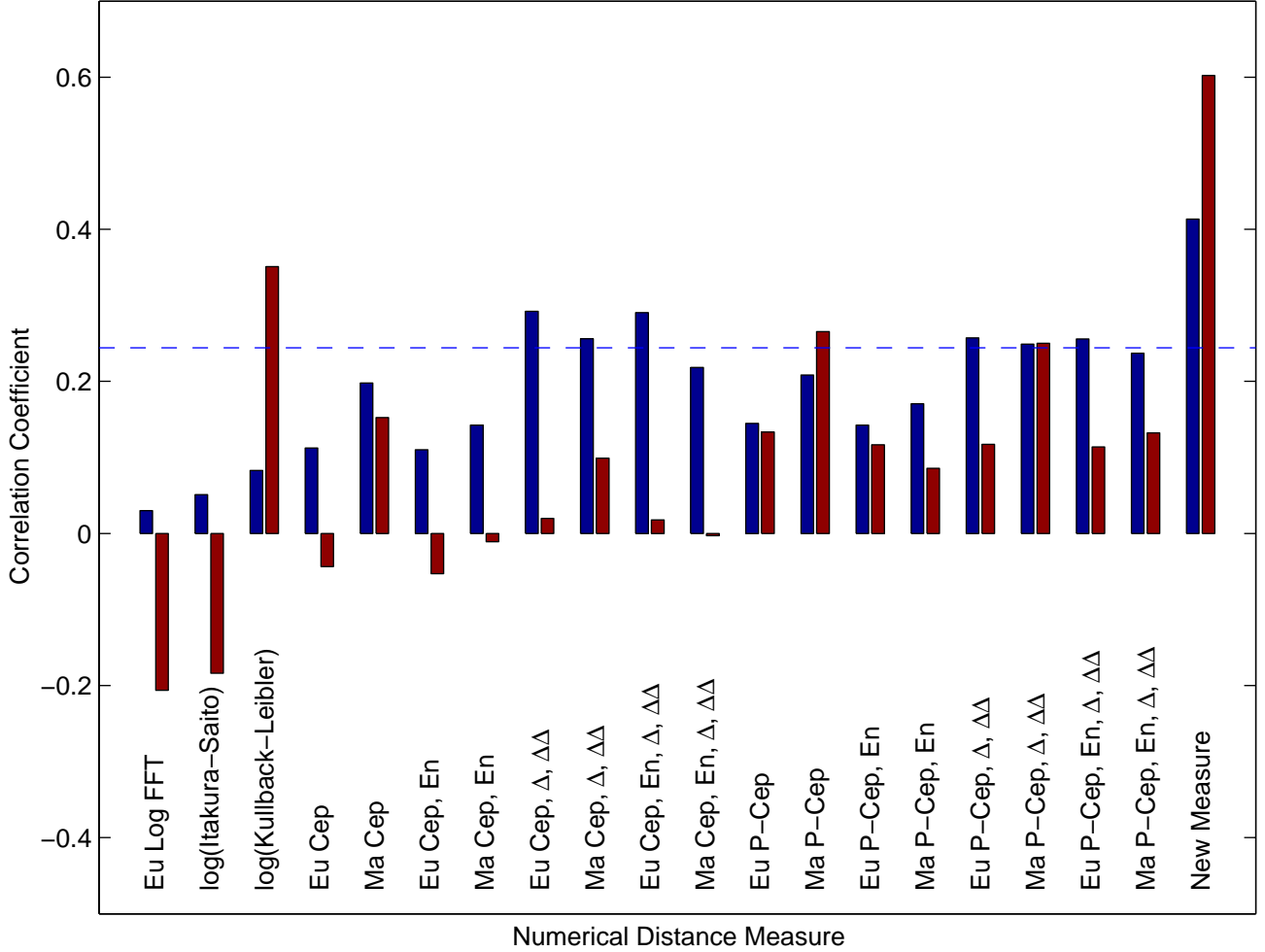


Figure 1: Correlation coefficients of various spectral distance measures with the group’s (left bar) and author’s (right bar) preference ratings for the 110 test stimuli described in Section 3. Eu = Euclidean, Ma = Mahalanobis, Cep = mel-frequency cepstral coefficients, En = log energy, P-Cep = perceptually modified mel-frequency cepstral coefficients (see Section 2), $\Delta, \Delta\Delta$ = 1st and 2nd time differentials of said coefficients. The dashed line shows the correlation required to be significant at the 1% level.

used in the group test, the author has 9 years of experience of listening to segmental defects in concatenative speech synthesis systems, and his ratings might therefore reflect more accurately the presence of spectral discontinuities.

4. Results

Figure 1 shows the correlation coefficients of various spectral distance measures with both the group’s mean ratings (the left bar of each pair) and the author’s ratings (the right bar of each pair) for the test stimuli described in Section 3. All measures were computed from 11kHz speech. During synthesis the segments composing each stimulus were scaled to the energy value predicted by the energy decision trees. Database derived log energy values and log FFT vectors were therefore adjusted to reflect the synthesis energy before computing distances dependent on these quantities. The perceptually modified cepstra however were computed from the database speech directly with no allowance made for any synthesis-time energy shift when determining the threshold of hearing.

The Euclidean log FFT distance was computed using 21 dimensional mel-binned log FFT vectors with equal weighting

in each dimension. The Itakura-Saito distance is given in [10] as

$$d_{IS}(S, S') = \int_{-\pi}^{\pi} \left[e^{V(w)} - V(w) - 1 \right] \frac{dw}{2\pi} \quad (2)$$

where $V(w) = \log S(w) - \log S'(w)$, and $S(w)$ and $S'(w)$ are power spectra. Since this distance is asymmetric, a symmetric version was used here, computed as

$$\begin{aligned} d_{IS_{sym}}(S, S') &= \frac{1}{2} [d_{IS}(S, S') + d_{IS}(S', S)] \quad (3) \\ &= \frac{1}{2} \int_{-\pi}^{\pi} \left[\frac{S(w)}{S'(w)} + \frac{S'(w)}{S(w)} - 2 \right] \frac{dw}{2\pi} \quad (4) \end{aligned}$$

where the integral was actually performed as a summation using mel-binned power spectra. The Kullback-Leibler distance is also asymmetric

$$d_{KL}(S, S') = \int_{-\pi}^{\pi} S(w) \log \left(\frac{S(w)}{S'(w)} \right) \frac{dw}{2\pi} \quad (5)$$

and again a symmetric version was used here, computed as

$$d_{KL_{sym}}(S, S') = \frac{1}{2} [d_{KL}(S, S') + d_{KL}(S', S)] \quad (6)$$

$$= \frac{1}{2} \int_{-\pi}^{\pi} \log \frac{S(w)}{S'(w)} [S(w) - S'(w)] \frac{dw}{2\pi} \quad (7)$$

with the integral again performed as a summation using mel-binned power spectra. The distance measures used were actually the log of the Kullback-Leibler and Itakura-Saito distances, in order to bring down the dynamic range which otherwise spanned many orders of magnitude.

The Euclidean cepstral distance was computed using 12 dimensional MFCC vectors with equal weightings in each dimension. The Mahalanobis cepstral distance was computed using the same vectors and Equation 1, with μ_i^l and $(\sigma_i^l)^2$ now the i th elements of the global cepstral mean and diagonal covariance matrix. The distances between perceptual cepstra were computed similarly, using the perceptual cepstral vectors described in Section 2.

The dashed line in Figure 1 shows the correlation required to be significant at the 1% level; a 5% level cannot be used because with 20 measures being evaluated there would be a high probability of one being judged significant just by chance.

5. Discussion

As can be seen from Figure 1, the author's ratings correlate significantly with the Kullback-Leibler distance (consistent with the result in [8]), the Mahalanobis distance between perceptual cepstra, the Mahalanobis distance between perceptual cepstra with deltas and delta-deltas, and most significantly of all with the new distance measure described in Section 2. The high correlation of the author's ratings with Mahalanobis perceptual cepstra during preliminary testing led to their use as the basis of the new measure. The group's ratings correlate significantly mostly with different measures, with the presence of delta and delta-delta coefficients being particularly helpful. Despite the differences however, the correlation with Mahalanobis perceptual cepstra is still relatively high, and the highest scoring measure is again the new measure described in Section 2.

The new distance measure was incorporated into the IBM Trainable Speech Synthesis System, [5], using a monotonically increasing mapping curve to convert the values it produced into costs suitable for use in the system. The mapping curve tried to ensure that the worst joins received costs comparable to the worst costs seen elsewhere in the system, and that the best joins were almost free. Analysis of the speech synthesised using the new measure revealed problems with phone transitions across word boundaries, as if the new cost was encouraging cross-word spectral transitions to sound too much like word-internal spectral transitions. This could happen because most of the data in the training database was word-internal, so the decision tree leaf means and variances reflected mainly word-internal transition statistics. The solution adopted was to build separate decision trees for word-internal and cross-word transitions, and descend the appropriate tree for each join at synthesis time.

Listening tests were conducted to determine whether using the new distance measure is worthwhile from a speech quality perspective compared to using a simple, yet well motivated, computationally-cheap alternative. Two synthesis systems, each with 4767 acoustic leaves [5], were compared. In both systems the continuity cost was zero for segments which were adjacent in the training database. In one system the new

word-boundary sensitive cost described above was used to compute all other continuity costs, while in the other system these costs were uniform. The uniform cost was set equal to the prosodic modification cost corresponding to the approximate limit of perceptually acceptable signal processing modification. This ensured that the use of non-contiguous speech was strongly discouraged but that the system would not introduce prosodic artifacts solely to avoid a join. Eight listeners rated fifty sentences generated randomly from the two systems using a 5-point opinion score. Collapsing across sentences, a t-test conducted on the listeners' system difference scores showed a (borderline) significant result ($p = 0.055$, 1 tailed test) that the system using the new continuity measure was superior.

6. Conclusions

A substantially improved distance measure for the costing of spectral discontinuities in concatenative speech synthesisers has been developed. Listening tests have justified the use of the new measure in the IBM synthesis system.

7. Acknowledgments

Thanks to Mike Monkowski for recording the speech data used in this work. Thanks also to the members of the IBM speech group who performed the listening tests.

8. References

- [1] Donovan, R.E. (1996) *Trainable Speech Synthesis*, PhD. Thesis, Cambridge University Engineering Department.¹
- [2] Hunt, A., and Black, A. (1996) Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database, *Proc. ICASSP'96, Atlanta*.
- [3] Huang, X., Acero, A., Adcock, J., Hon, H-W., Goldsmith, J., Liu, J., and Plumpe, M. (1996) Whistler: A Trainable Text-to-Speech System, *Proc. ICSLP'96, Philadelphia*.
- [4] Black, A., and Taylor, P. (1997) Automatically Clustering Similar Units for Unit Selection in Speech Synthesis, *Proc. Eurospeech'97, Rhodes*.
- [5] Donovan, R.E., and Eide, E.M. (1998) The IBM Trainable Speech Synthesis System, *Proc. ICSLP'98, Sydney*.
- [6] Beutnagel, M., Conkie, A., Schroeter, J., Stylianou, Y., and Syrdal, A. (1999) The AT&T Next-Gen TTS System, *Joint meeting of ASA, EAA, and DAGA, Berlin, 15-19 March 1999*.
- [7] Coorman, G., Fackrell, J., Rutten, P., and Van-Coile, B. (2000) Segment Selection in the L&H Realspeak Laboratory TTS System, *Proc. ICSLP 2000, Beijing*.
- [8] Klabbbers, E., and Veldhuis, R. (1998) On the Reduction of Concatenation Artefacts in Diphone Synthesis, *Proc. ICSLP'98, Sydney*.
- [9] Bahl, L.R., deSouza, P.V., Gopalakrishnan, P.S., and Picheny, M.A. (1993) Context Dependent Vector Quantization for Continuous Speech Recognition, *Proc. ICASSP'93, Minneapolis*.
- [10] Rabiner, L., and Juang, B-H. (1993) *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, New Jersey.

¹Available by anonymous ftp to svr-ftp.eng.cam.ac.uk