OBJECTIVE DISTANCE MEASURES FOR SPECTRAL DISCONTINUITIES IN CONCATENATIVE SPEECH SYNTHESIS

Jithendra Vepa[1,2]*, Simon King*[1]

[1]Centre for Speech Technology Research
University of Edinburgh
Edinburgh, UK

Paul Taylor[2]

[2]Rhetorical Systems
Edinburgh, UK
www.rhetorical.com

ABSTRACT

In unit selection based concatenative speech systems, *join cost*, which measures how well two units can be joined together, is one of the main criteria for selecting appropriate units from the inventory. The ideal join cost will measure *perceived* discontinuity, based on easily measurable spectral properties of the units being joined, in order to ensure smooth and natural-sounding synthetic speech. In this paper we report a perceptual experiment conducted to measure the correlation between *subjective* human perception and various *objective* spectrally-based measures proposed in the literature. Our experiments used a state-of-the art unit-selection text-to-speech system: *rVoice* from Rhetorical Systems Ltd.

1. INTRODUCTION

Unit-selection based speech synthesis systems have become popular recently because of their highly natural-sounding synthetic speech. These systems have large speech databases containing many instances of each speech unit (e.g. diphone), with varied and natural distribution of prosodic and spectral characteristics. When synthesising an utterance, the selection of the best unit sequence from the database is based on a combination of two costs: target cost (how closely candidate units in the inventory match the required targets) and join cost (how well neighbouring units can be joined) [1]. The target cost is calculated as the weighted sum of the differences between the various prosodic and phonetic features of target and candidate units. The concatenation cost is also determined as the weighted sum of sub-costs, such as absolute differences in F0 and amplitude, mismatch in various spectral (acoustic) features, MFCCs, LSFs, etc. The optimal unit sequence is then found by a Viterbi search for the lowest cast path through the lattice of the target and concatenation costs.

The ideal join cost is one that, although based solely on measurable properties of the candidate units, such as spectral parameters, amplitude and F0, correlates highly with human perception of discontinuity at unit concatenation points. In other words: the join cost should predict the degree of perceived discontinuity. We report a perceptual experiment to measure this correlation for various join cost formulations.

A few recent studies have been conducted in this context. Klabbers and Veldhius [2] examined various distance measures on five Dutch vowels to reduce the concatenation discontinuities in diphone synthesis and found that a Kullback-Leibler measure on LPC power-normalised spectra was the best predictor. A similar study by Wouters and Macon [3] for unit selection, showed that the Euclidean distance on Mel-scale LPC-based cepstral parameters was a good predictor, and utilising weighted distances or delta coefficients could improve the prediction. Stylianou and Syrdal [4] found that the Kullback-Leibler distance between FFT-based power spectra had the highest detection rate. Donovan [5] proposed a new distance measure which uses a decision tree based context dependent Mahalanobis distance between perceptual cepstral parameters.

All these previous studies focused on human detection of audible discontinuities in **isolated words** generated by concatenative synthesisers. We extend this work to the case of **polysyllabic words in natural sentences** and new spectral features, Multiple Centroid Analysis (MCA) coefficients.

2. PERCEPTUAL LISTENING TESTS

A listening test was designed to measure the degree of **perceived** concatenation discontinuity in natural sentences generated by the state-of-the art speech synthesis system, using an adult North-American male voice.

2.1. Test Design & Stimuli

A preliminary assessment indicated that spectral discontinuities are particularly prominent for joins in the middle of diphthongs, presumably because this is a point of spectral change (due to moving formant values). This study therefore focuses on such joins. Previous studies have also shown that

Thanks to Rhetorical Systems Ltd. for funding this work

diphthongs have higher discontinuity detection rates than long or short vowels [6].

We selected two natural sentences for each of five American English diphthongs (ey, ow, ay, aw and oy) [7]. One word in the sentence contained the diphthong in a stressed syllable. The sentences are listed in Table 1.

diphthong	*sentences*
ey	More **plac**es are in the pipeline. The government sought autho**riz**ation of his citizenship.
ow	European shares resist **glob**al fallout. The speech sym**pos**ium might begin on Monday.
ay	This is **high**ly significant. Primitive **trib**es have an upbeat attitude.
aw	A large **hous**ehold needs lots of appliances. Every picture is worth a **thous**and words.
oy	The **boy** went to play Tennis. Never exp**loit** the lives of the needy.

Table 1. The stimuli used in the experiment. The syllable in bold contains the diphthong join.

These sentences were then synthesised using the experimental version of *rVoice* speech synthesis system. For each sentence we made various synthetic versions, by varying the two diphone candidates which make the diphthong and keeping all the other units the same. We removed the synthetic versions which were worse at the joins of neighbouring phones of the diphthong. The remaining versions were further pruned based on target features of the diphones making the diphthong, to ensure similar prosody among synthetic versions. This process resulted in around 30 versions with variation in concatenation discontinuities at the diphthong join. We manually selected the best and worst synthetic versions by listening to these 30 versions based on authors perception of the join. This process was repeated for each sentence in Table 1.

2.2. Test Procedure

There were around 17 participants in our perceptual listening test, most of them are PhD or MSc students with some experience of speech synthesis. Most of them are native speakers of British English.

Subjects were first shown the written sentence, with an indication of which word contains the join. At the start of the test they were first presented with a pair of reference stimuli: one containing the best and the other the worst joins (as selected by the authors) in order to set the endpoints of a 1-to-5 scale. Subjects could listen to the reference stimuli as many times as they liked and they could also review them at regular intervals (for every 10 test stimuli) throughout the test.

They were then played each test stimulus in turn and were asked to rate the quality of that join on a scale of 1 (worst) to 5 (best). They could listen to each test stimulus up to three times. Each test stimulus consisted of first the entire sentence, then only the word containing the join (extracted from the full sentence, not synthesised as an isolated word).

The test was carried out in blocks of around 35 test stimuli, with one block for each sentence in Table 1. Subjects could take as long as they pleased over each block, and take rests between blocks. Each test block contained a few duplications of some test stimuli to validate the subjects scores, explained in Section 4.

3. OBJECTIVE DISTANCE MEASURES

A distance measure operates on a parameterisation of the speech signal, such as Mel Frequency Cepstral Coefficients (MFCCs), Line Spectral Frequencies (LSFs) and Multiple Centroid Analysis (MCA) coefficients. A distance measure between two vectors of such parameters can use various metrics: Euclidean, Absolute, Kullback-Leibler or Mahalanobis. We describe these briefly in Section 3.2.

3.1. Parameterisations

We used three parameterisations, MFCCs [8], LSFs [9], MCA coefficients. The third parameterisation – MCA – is less well known, so we briefly describe it below.

Multiple Centroid Analysis was introduced by Crowe & Jack [10] as an alternative to traditional formant estimation techniques, which employs a global optimisation based on a generalisation of the centroid. To compute centroids, we consider a multi-modal distribution such as a speech power spectrum, then split it into appropriate number of partitions say 4 or 5, as shown in Fig.1. The centroid of a specific partition of the distribution $P(n)$ bounded by $n = c_1$ and $n = c_2$ is estimated as the value that gives minimum squared error, as shown in the equation below:

$$e(c1, c2, k) = \sum_{c_1}^{c_2} (n - k)^2 P(n) \qquad (1)$$

This will be computed for every possible combination of partitions and a minimum error condition is used to determine the optimal partition boundary positions. If the spectral distribution within a single partition contains a single formant then the centroid and associated variance represents the formant frequency and bandwidth [11]. This is more robust than peak picking, so is an attractive alternative to linear prediction based formant trackers.

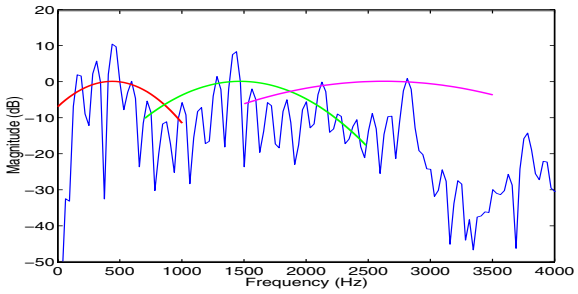

Fig. 1. Speech power spectrum and MCA (three centroids).

3.2. Distance metrics

Standard distance measures, such as Euclidean, Absolute, Kullback-Leibler, Mahalanobis distances were computed for all the above speech parameterisations, MFCCs, LSFs and MCA coefficients respectively.

The Euclidean distance between two feature vectors is:

$$Eu(X,Y) = \left(\sum_{i=1}^{n}(X_i - Y_i)^2\right)^{1/2} \quad (2)$$

The Absolute distance is computed as the absolute magnitude difference between individual features of the two feature vectors.

The Kullback-Leibler (K-L) distance [12] is used to compute the distance between two probability distributions $f(x)$ and $g(x)$:

$$KL(f,g) = \int (f(x) - g(x))\ log\left(\frac{f(x)}{g(x)}\right)dx \quad (3)$$

Mahalanobis distance [5] is a generalisation of standardised distance:

$$R(X,Y)^2 = \sum_{i=1}^{n}\left[\frac{X_i - Y_i}{\sigma_i}\right]^2 \quad (4)$$

where,σ_i is standard deviation of the i^{th} feature of the feature vectors.

4. RESULTS AND DISCUSSION

In Table 2, we present the number of subjects for each sentence and the number of subjects with more than 50% consistency in rating the joins. The consistency of subjects was measured on a validation set, which we included in the test stimuli for each sentence. Mean listener scores were computed only for the subjects with more than 50% consistency in rating the joins. Also, we manually checked all listeners' ratings, and removed the listener scores with all same rating (e.g all '1's) during mean listener computation.

	no. of subjects	consistent subjects
ey	13, 14	11, 8
ow	11, 13	6, 7
ay	17, 11	9, 6
aw	11, 13	11, 10
oy	13, 14	6, 6

Table 2. Consistency of subjects in listening tests, each number in a pair corresponds to the sentences listed in Table 1.

Correlation coefficients of various spectral distance measures with mean listener preference ratings are reported in Tables 3, 4 and 5. The correlation coefficients above the 1% significant level have been highlighted. It is clear that no distance measure performs well in all cases. The distance measures computed on MCA coefficients have a higher number of 1% significant correlations compared to those obtained from MFCCs and LSFs. Unfortunately, none of these measures yield 1% significant level correlation for four of the sentences. Using delta coefficients did not improve correlations; they are sometimes worse rather than better. Also, simple absolute distance is as good as any other distance measure.

	Euclidean		Absolute		Mahalanobis	
	mfcc	mfcc+Δ	mfcc	mfcc+Δ	mfcc	mfcc+Δ
ey	0.27	0.34	0.28	0.38	0.21	0.35
	0.60	**0.55**	**0.64**	**0.55**	**0.66**	**0.50**
ow	0.31	0.33	0.32	0.33	0.31	0.24
	0.53	**0.49**	**0.51**	**0.44**	**0.56**	0.42
ay	0.32	0.24	0.34	0.20	0.39	0.11
	0.63	**0.67**	**0.65**	**0.71**	**0.66**	**0.61**
aw	0.40	0.32	0.42	0.26	0.34	0.06
	0.74	**0.75**	**0.72**	**0.74**	**0.77**	**0.75**
oy	-0.01	-0.03	0.02	-0.01	0.17	0.15
	-0.01	0.06	-0.02	0.09	-0.01	0.15

Table 3. Correlation between perceptual scores and various objective distance measures based on MFCCs.

In Table 3 it can be seen that all three Euclidean, Absolute, Mahalanobis distance metrics based on MFCCs have good correlations with perceptual scores in many cases. The objective distance measures based on LSFs also have better correlations in some cases, as observed in Table 4. From Table 5, it is clear that all objective distance measures on MCA coefficients correlate well with perceptual scores in most of the cases compared to those of MFCCs and LSFs. As an additional advantage, the size of the MCA vector is only 12 (including deltas), whereas MFCCs are 26 and LSFs are 24. Considering the computational complexity and size,

the absolute distance measure based on MCA coefficients outperforms the other metrics, which has five 1% significant correlations out of ten cases.

	Euclidean		Absolute		Mahalanobis		K-L
	lsf	lsf+Δ	lsf	lsf+Δ	lsf	lsf+Δ	lsf
ey	0.05	0.06	0.14	0.20	0.29	0.37	0.30
	0.63	**0.63**	**0.64**	**0.64**	**0.64**	**0.58**	**0.68**
ow	0.42	0.40	0.37	0.29	0.35	0.21	0.37
	0.41	0.42	0.34	0.36	0.34	0.40	0.29
ay	0.15	0.13	0.12	0.05	0.21	0.01	0.35
	0.58	**0.65**	**0.59**	**0.69**	**0.64**	**0.61**	**0.68**
aw	0.33	0.39	0.22	0.38	0.31	**0.66**	0.29
	0.77	**0.78**	**0.76**	**0.77**	**0.78**	**0.78**	**0.78**
oy	0.16	0.18	0.13	0.18	0.12	0.28	0.12
	0.01	0.03	0.04	0.09	-0.01	0.17	0.18

Table 4. Correlation between perceptual scores and various objective distance measures based on LSFs.

	Euclidean		Absolute		Mahalanobis		K-L
	mca	mca+Δ	mca	mca+Δ	mca	mca+Δ	mca
ey	0.31	0.32	0.29	0.36	0.32	0.36	0.41
	0.59	**0.46**	**0.58**	**0.46**	**0.55**	**0.62**	**0.62**
ow	0.07	0.13	0.12	0.19	0.17	0.10	0.17
	0.37	0.43	0.39	**0.46**	**0.46**	0.39	0.32
ay	-0.04	0.11	-0.05	0.03	-0.02	0.01	0.07
	0.55	0.43	**0.50**	**0.45**	**0.53**	**0.50**	**0.57**
aw	**0.48**	0.27	0.37	0.35	0.39	0.34	0.37
	0.74	**0.58**	**0.73**	**0.57**	**0.77**	**0.69**	**0.81**
oy	0.32	**0.53**	0.28	**0.53**	0.21	0.22	0.21
	0.01	0.19	0.03	0.30	0.06	0.14	0.16

Table 5. Correlation between perceptual scores and various objective distance measures based on MCA coefficients.

5. FUTURE WORK

Our test stimuli was confined to five American English diphthongs, also we only used two sentences for each diphthong from a single speaker. It would be worthwhile to perform experiments using more sentences for each case, to get more insight into the various distance metrics. Also, it would be interesting to know how these distance measures detect discontinuities in liquids, which have been shown [2, 7] to be very susceptible to the spectral characteristics of the surrounding phones. Further research is needed to develop new distance measures, also to incorporate delta features into them, to improve their performance in all cases.

6. ACKNOWLEDGEMENTS

Thanks to all the experimental subjects: the members of CSTR, staff at Rhetorical Systems Ltd. and students on the M.Sc. in Speech and Language processing, University of Edinburgh. The authors also acknowledge the assistance of Dr. Alice Turk of the Dept. of Theoretical and Applied Linguistics in designing the listening tests.

7. REFERENCES

[1] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *Proc. ICASSP*, pp. 373–376, 1996.

[2] E. Klabbers and R. Veldhuis, "On the reduction of concatenation artefacts in diphone synthesis," *Proc. ICSLP98*, pp. 1983–1986, 1998.

[3] J. Wouters and M. Macon, "Perceptual evaluation of distance measures for concatenative speech synthesis," *Proc. ICSLP98*, pp. 2747–2750, 1998.

[4] Y. Stylianou and Ann K. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," *Proc. ICASSP*, 2001.

[5] Robert E. Donovan, "A new distance measure for costing spectral discontinuities in concatenative speech synthesisers," *The 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, 2001.

[6] Ann K. Syrdal, "Phonetic effects on listener detection of vowel concatenation," *Proc. Eurospeech*, 2001.

[7] J. Olive, A. Greenwood, and J. Coleman, *Acoustics of American English Speech: A Dynamic Approach*, Springer, 1993.

[8] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[9] F.K. Soong and B.H.. Juang, "Line spectrum pairs (LSP) and speech data compression," *Proc. ICASSP*, pp. 1.10.1–1.10.4, 1984.

[10] A. Crowe and M.A. Jack, "Globally optimising formant tracker using generalised centroids," *Electronic Letters*, vol. 23, no. 19, pp. 1019–1020, 1987.

[11] A.A. Wrench, "Analysis of fricatives using multiple centres of gravity," *Proc. International Congress of Phonetic Sciences*, , no. 4, pp. 460–463, 1995.

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