NEW 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

The quality of unit selection based concatenative speech synthesis mainly depends on how well two successive units can be joined together to minimise the audible discontinuities. The objective measure of discontinuity used when selecting units is known as the *join cost*. 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 describe a perceptual experiment conducted to measure the correlation between *subjective* human perception and various *objective* spectrally-based measures proposed in the literature. Also we report new objective distance measures derived from various distance metrics based on these spectral features, which have good correlation with human perception to concatenation discontinuities. Our experiments used a state-of-the art unit-selection text-to-speech system: *rVoice* from Rhetorical Systems Ltd.

1. INTRODUCTION

In unit-selection based speech synthesis systems, synthesised speech is produced by concatenating speech units selected from large database, containing many instances of each speech unit, with varied prosodic and spectral characteristics. Hence it is possible to synthesise more natural-sounding speech. 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 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. A few recent studies have attempted to determine which objective distance measures are best able to predict audible discontinuities. Klabbers and Veldhius [2] examined various distance measures on five Dutch vowels to reduce the concatenation discontinuities in diphone synthesis and found that the 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 vectors.

All these previous studies focused on human detection of audible discontinuities in **isolated words** generated by concatenative synthesisers. We extended this work to the case of **polysyllabic words in natural sentences** and new spectral features, Multiple Centroid Analysis (MCA) coefficients [6]. But we found that no single distance measure perform well for all cases. In this paper, we report new distance measures which correlate well with human perception to the concatenation discontinuities. These are weighted sums of the distance metrics of various spectral features.

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

Thanks to Rhetorical Systems Ltd. for funding this work

shown that diphthongs have higher discontinuity detection rates than long or short vowels [7].

We selected two natural sentences for each of five American English diphthongs (ey, ow, ay, aw and oy) [8]. 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 author**iza**tion of his citizenship.
ow	European shares resist **glob**al fallout. The speech sym**po**sium might begin on Monday.
ay	This is **high**ly significant. Primitive **trib**es have an upbeat attitude.
aw	A large **house**hold 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. The authors manually selected what they judged to be the best and worst synthetic versions by listening to these 30 versions. 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 were PhD or MSc students with some experience of speech synthesis. Most of them were 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, as explained in Section 4.

3. WEIGHTED DISTANCE MEASURES

We used three parameterisation of the speech signal, Mel Frequency Cepstral Coefficients (MFCCs), Line Spectral Frequencies (LSFs) and Multiple Centroid Analysis (MCA) coefficients [9] respectively. A distance measure between two vectors of such parameters can use various metrics: Euclidean, Absolute, Kullback-Leibler or Mahalanobis.

From our previous study [6], it is clear that no single distance measure computed on various spectral features performs well for all cases. One solution to this would be to use phone-specific objective distance measures. But, to decide which distance measure to use for each phoneme would require large amounts of perceptual data. However, our preliminary studies showed that a weighted sum of these various distances results in better correlations compared to those obtained from individual distances. Also, we observed that using delta coefficients in distance metrics does not improve correlations. Hence, we used other spectral parameters instead of considering delta features to maintain almost same feature vector size.

Consider, the weighted distance as shown in the equation below,

$$\sum_{i=1}^{N} w_i * D_i(e_j, s_j) = L_j \qquad \forall j \tag{1}$$

where, w_i is weight on distance (D_i) between two feature vectors, e_j and s_j, L_j is the mean listener rating. The set of equations is over-determined (more equations than unknowns) if $j > i$, j is the index ($1 \leq j \leq M$) of a join in the perceptual experiment results. Then, weights can be computed by solving this matrix shown below,

$$\mathbf{A} \cdot \mathbf{w} = \mathbf{l} \tag{2}$$

where $\mathbf{A}$ is M-by-N distance matrix and $\mathbf{l}$ is the column vector $\{L_1, \cdots, L_M\}$. We used standard least-squares method to find $\mathbf{w}$.

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. Then mean listener scores were computed only for the subjects with more than 50% consistency in rating the joins. Also, we manually checked all the 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.

4.1. Weighted distances of all three features

Table 3 summarises correlations of absolute distances based on MFCCs, LSFs and MCA coefficients and their weighted sum with mean listener ratings. The correlation coefficients above the 1% significant level are highlighted. The weights used for MFCCs, LSFs and MCA coefficients are 0.13, 0.03, 0.84 respectively. The correlation coefficients of Euclidean

	mfcc	lsf	mca	weighted dist.
ey	0.28	0.14	0.29	0.31
	0.64	**0.64**	**0.58**	**0.65**
ow	0.32	0.37	0.12	0.29
	0.51	0.34	0.39	**0.49**
ay	0.34	0.12	-0.05	0.26
	0.65	**0.59**	**0.50**	**0.65**
aw	0.42	0.22	0.37	**0.48**
	0.72	**0.76**	**0.73**	**0.74**
oy	0.02	0.13	0.28	0.09
	-0.02	0.04	0.03	-0.01

Table 3. Correlation between perceptual scores and absolute distances of MFCCs, LSFs, MCA coefficients and weighted sum of above three measures.

distance measures of all three spectral features and their weighted sum with mean listener ratings are reported in Table 4. Weights used are 0.15 (MFCCs), 0.35(LSFs) and 0.5 (MCAs).

	mfcc	lsf	mca	weighted dist.
ey	0.27	0.05	0.31	0.27
	0.60	**0.63**	**0.59**	**0.64**
ow	0.31	0.42	0.07	0.29
	0.53	0.41	0.37	**0.51**
ay	0.32	0.15	-0.04	0.25
	0.63	**0.58**	**0.55**	**0.65**
aw	0.40	0.33	**0.48**	**0.47**
	0.74	**0.77**	**0.74**	**0.76**
oy	-0.01	0.16	0.32	0.08
	-0.01	0.01	0.01	0.00

Table 4. Correlation between perceptual scores and Euclidean distances of MFCCs, LSFs, MCA coefficients and weighted sum of above three measures.

	mfcc	lsf	mca	weighted dist.
ey	0.21	0.29	0.32	0.31
	0.66	**0.64**	**0.55**	**0.64**
ow	0.31	0.35	0.17	0.25
	0.56	0.34	**0.46**	**0.53**
ay	0.39	0.21	-0.02	0.18
	0.66	**0.64**	**0.53**	**0.63**
aw	0.34	0.31	0.39	**0.44**
	0.77	**0.78**	**0.77**	**0.79**
oy	0.17	0.12	0.21	0.23
	-0.01	-0.01	0.06	0.03

Table 5. Correlation between perceptual scores and Mahalanobis distances of MFCCs, LSFs, MCA coefficients and weighted sum of above three measures.

From Table 4 it is evident that we can improve correlations by setting weights on individual distances, e.g *ow* has good correlation for MFCCs, similarly MCA coefficients yield better correlations for *aw*. However, weighted measure achieves good correlations for both the cases. In Table 5 we present correlations between perceptual scores and Mahalanobis distances of MFCCs, LSFs and MCA coefficients and their weighted sum. The weights used are 0.39(MFCCs), 0.0(LSFs) and 0.61(LSFs).

4.2. Weights on MCA parameters

We found that MCA coefficients have higher weights compared to MFCCs. Also the size of the MCA feature vector is only 12 (including deltas), whereas MFCCs are 26 and LSFs are 24. Hence, we carried out a further experiment in which the individual MCA coefficients were weighted. The least-squares method did not yield good solutions in this case. So, we randomly generated the weights and checked the corre-

lations and chose the ones which produce more 1% significant correlations (i.e those highlighted in the tables). Table 6 shows three different sets of weights on MCA parameters, and corresponding correlations obtained are shown in Table 7. Set2 and 3 produces seven 1% significant correlations out of ten cases, also achieved good correlations for *oy* diphthong, which has very poor correlations with other distance measures (see Tables 3,4,5).

MCA parameter	set1	set2	set3
F1	0.682	0.342	0.699
F2	0.168	0.528	0.181
F3	0.419	0.026	0.547
B1	0.782	0.211	0.982
B2	0.109	0.520	0.589
B3	0.623	0.887	0.237
E1	0.251	0.242	0.223
E2	0.271	0.367	0.141
E3	0.028	0.019	0.081
DF1	0.150	0.198	0.838
DF2	0.211	0.211	0.536
DF3	0.437	0.924	0.778

Table 6. Various weights used on MCA parameters, Formant frequency(F), Bandwidth(B), Energy(E), Delta-Formant frequency(DF).

	set1	set2	set3
ey	0.43	**0.44**	**0.45**
	0.47	**0.60**	**0.58**
ow	0.09	0.19	0.11
	0.45	**0.52**	**0.49**
ay	0.04	-0.02	0.07
	0.48	**0.49**	0.41
aw	**0.46**	**0.49**	**0.46**
	0.66	**0.62**	**0.67**
oy	**0.55**	**0.55**	**0.50**
	0.34	0.39	**0.44**

Table 7. Correlation between perceptual scores and absolute distances based on weighted MCA coefficients.

5. FUTURE WORK

Further work is need to tune these weights to achieve high correlation for all cases. Also, more perceptual experiments need to be carried out to determine phoneme specific distance measures.

The computation of join cost and spectral smoothing are closely related. Suppose, if we had a large database and a perfect measure of join cost then no smoothing would be required. Conversely, if we could smooth joins better, then the method of computing join would be less critical. Hence it would be optimal if we combine these two operations in some optimal way. Presently, we are investigating a single representation, which can be used for join cost computation as well as smoothing.

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] J. Vepa, S. King, and P. Taylor, "Objective distance measures for spectral discontinuities in concatenative speech synthesis," *ICSLP*, 2002.

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

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

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