Feature Extraction for Spectral Continuity Measures in Concatenative Speech Synthesis

Barry Kirkpatrick, Darragh O'Brien and Ronán Scaife

Faculty of Engineering and Computing
Dublin City University, Dublin 9, Ireland
{bkirkpatrick, dobrien}@computing.dcu.ie, scaifer@eeng.dcu.ie

Abstract

The quality of concatenative speech synthesis depends on the cost function employed for unit selection. Effective cost functions for spectral continuity are difficult to define and standard measures often do not accurately reflect human perception of discontinuity across a concatenated join. In this study the performance of a number of standard distance measures are compared for the task of detecting audible discontinuities in concatenated speech. Feature sets derived from the phase spectrum are also investigated. Feature extraction based on wavelet analysis is proposed to overcome some of the limitations of the standard measures tested. Receiver Operating Characteristic (ROC) curves are constructed for each measure from the results of a perceptual experiment and are used to rank the performance of each measure. Results indicate that phase spectra is comparable to magnitude spectra as a join cost for spectral continuity. Measures based on wavelet transform coefficients outperform all other measures tested.

Index Terms: speech synthesis, unit selection, join cost, wavelet transform, phase spectra.

1. Introduction

Unit selection based concatenative synthesis is currently considered state-of-the-art in text-to-speech (TTS) synthesis and is capable of producing highly natural-sounding synthetic speech. Synthetic speech is produced by concatenating units of speech which are selected from a large speech database containing many possible instances of each unit, each of which exhibit varied prosodic and spectral characteristics. The selection of the best unit sequence is based upon a cost criterion, which is composed of a target cost and a join cost [1]. The target cost measures the difference between the target unit and the unit under consideration in terms of prosodic and phonetic parameters, the join cost measures how suitable two neighboring units are for concatenation. In order to select the optimum sequence of units, the database of units is modeled as a state transition network with the target cost representing the cost of state occupancy and the join cost representing the cost of state transition. The optimum unit sequence is determined by a Viterbi search through the network.

An ideal join cost should accurately reflect human perception of discontinuity. A number of studies have attempted to determine which distance measures are most successful at predicting audible discontinuities in concatenated speech. Klabbers and Veldhuis [2] found the Kullback-Leibler distance between LPC power spectra to be the best predictor, Stylianou and Syrdal [3] found the Kullback-Leibler distance between FFT-based power spectra to be the best predictor. In similar studies Wouters and Macon [4] found the Euclidean distance between Mel-scale LPC based cepstra to outperform other measures while Vepa and King [5] found Line Spectral Frequencies (LSF) to be good predictors of audible discontinuities. Many studies have presented conflicting results with measures that ranked highly in one study performing poorly in another. For example, Stylianou and Syrdal found the Euclidean distance between LSFs to be the worst predictor of audible discontinuity, a finding that is inconsistent with those of Vepa and King. Klabbers and Veldhuis reported that the Euclidean distance between Mel-Frequency Cepstral Coefficients (MFCC) ranked poorly as a predictor of audible discontinuities, although this measure ranked highly in many of the other studies. It is difficult to make direct comparisons between studies as each used a different database and different criteria to rank each measure.

This paper, while following similar previous studies, compares standard distance measures but also investigates phase spectra and the limitations of current measures. The wavelet transform is proposed as a means to overcome some of the limitations associated with the standard measures. The paper is organised as follows. In section 2, a perceptual experiment is described, the results of which were used to evaluate the performance of each measure. Section 3 explains how each measure is related to the results of the perceptual experiment using ROC curves. Section 4 outlines the standard feature sets compared in the study and introduces feature sets based on phase spectra. The limitations associated with these measures are investigated. Section 5 introduces the wavelet transform as an alternative strategy for feature extraction that addresses some of the limitations of the standard measures.

2. Database and perceptual experiment

2.1. Database

A database of test stimuli was constructed adopting the approach of Stylianou and Syrdal [3]. The inventory of units consisted of 300 words recorded from an adult male. The test words were concatenated by pitch synchronous overlap and add that exploited knowledge of the pitch marks to maintain F0 continuity across the join. The inventory of 300 words was recorded in a hemi-anechoic recording studio at a sampling frequency of 16 kHz.

2.2. Perceptual experiment

The perceptual test was divided into subtests, each containing 36 concatenated words. At the start of a subtest the listener was presented with examples of audible discontinuities. The test required the listener to make a forced decision for each test word, continuous or discontinuous. The listener was provided with the original recorded words that were used to create each of the concatenated words for comparison. Each test word could be listened to as often as the listener requested. Each subtest contained six control words to validate each listener's results. Each listener undertook the test in a quiet environment using headphones. Twelve listeners in total contributed perceptual results with coverage of three listeners per subtest. A majority scoring system was employed to decide if a test word was continuous or discontinuous.

3. Evaluation of distance measures

Each measure was evaluated by generating an ROC curve [6] based on the experimental results. The area under the ROC curve (AUC) was used to rank measures. Two probability density functions, $p(\kappa|0)$ and $p(\kappa|1)$, were estimated for each distance measure based on the perceptual results for continuous and discontinuous joins respectively. ROC curves were calculated from the probability density functions and provided information regarding the separability of $p(\kappa|1)$ and $p(\kappa|0)$, for each distance measure. The ROC curves were computed by calculating the hit rate, P_H (1), the probability of correctly detecting a discontinuity and the false alarm rate, P_{FA} (2), the probability of classifying a continuous join as discontinuous, for varying distance thresholds κ.

$$P_H(\kappa_0) = \int_{\kappa_0}^{\infty} p(\kappa|1) d\kappa \tag{1}$$

$$P_{FA}(\kappa_0) = \int_{\kappa_0}^{\infty} p(\kappa|0) d\kappa \tag{2}$$

The ROC curve is constructed by plotting the pairs P_H and P_{FA} for each threshold value, κ_0, from 0 to ∞. The AUC can be interpreted as the probability of correctly classifying a join.

3.1. F0 analysis

Test words judged to contain audible discontinuities from the perceptual test may have contained discontinuities that were not due to spectral mismatch. One common source of discontinuity is due to $F0$ mismatch across a join. In this study we were only concerned with spectral mismatch so audible discontinuities due to other sources may skew results. To address this issue the database of test words was analysed to identify test words containing an audible discontinuity that was potentially due to $F0$ mismatch across the join. The threshold for acceptable $F0$ mismatch was set at 10 Hz, so joins with $F0$ mismatch above this threshold were not included in the the final results.

4. Feature sets

4.1. Standard features

In this study the following standard feature sets were considered:

- MFCCs [7] computed from FFT and LPC spectra.
- Power Spectra (PS) and Log Power Spectra (LPS) computed from FFT, LPC and Perceptual Linear Prediction (PLP) [8].
- Cepstral coefficients computed from PLP (PLPCC) and LPC (LPCC) spectra.
- LSFs computed from LPC, on both linear and Mel-frequency scales and LSFs computed from PLP coefficients.

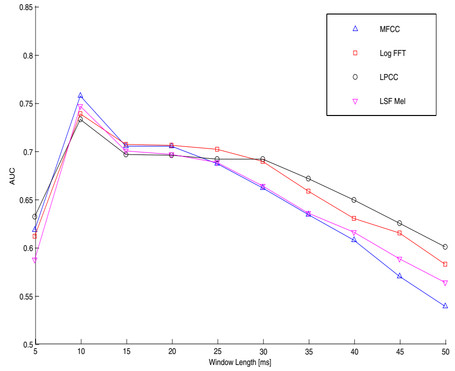

Figure 1: *The AUC for varying window length for selected feature sets with the l_2 distance.*

All features were extracted using a pitch synchronous window one pitch period in duration. The raw speech was pre-emphasised with the filter $H(z) = 1 - 0.95z^{-1}$ and Hanning windowed. The order of LPC analysis was 16 and was implemented using the autocorrelation method. The order of PLP analysis was 6. The number of cepstral coefficients used was 20 with the exception of PLP based cepstra in which case 6 coefficients were used. The first cepstral coefficient was not used in the distance calculation. All representations were power normalised. A 512-point FFT was used with the same number of frequency samples used for LPC based spectra.

In order to quantify the degree of similarity between two feature vectors the distance between the vectors is calculated. A number of distance measures are considered; the absolute distance (l_1), the Euclidean distance (l_2) the Cos distance and the Symmetric Kullback-Leibler distance (D_{skl}) [9].

4.2. Windowing and feature extraction

The size of window was found to have a significant impact on the results. The AUC is plotted against window length for the l_2 distance between selected feature sets in Fig. 1. This indicates the importance of selecting the appropriate window length and also that shorter window lengths are preferred, which demonstrates that temporal resolution is significantly important. For non-pitch synchronous windows the maximum AUC is achieved with a window length of 10 ms across all feature sets, this corresponds approximately with the average pitch period for the database. The AUC value decreases rapidly when the window length is smaller than a pitch period in length, this is the window length at which the FFT can no longer resolve individual harmonic components and appears to be the lower threshold for frequency resolution. The optimum windowing strategy for all feature sets was found to be pitch synchronous windowing with a window length of one pitch period.

4.2.1. *Results*

The Euclidean distance between MFCCs (FFT based) was found to have the highest AUC value of all the standard measures considered in this study. The AUC values for all the standard distance measures are presented in Table 1.

Features	l_1	l_2	Cos	D_{skl}
MFCC (FFT)	0.752623	0.758867	0.722920	-
LPC LSF Mel	0.752031	0.749812	0.747719	0.741308
MFCC (LPC)	0.739146	0.749416	0.716419	-
PLP LSF	0.722690	0.727755	0.721518	0.732785
LPC LSF Linear	0.736001	0.726512	0.723908	0.733440
FFT LPS	0.743525	0.738806	0.736100	-
FFT PS	0.741910	0.695029	0.718724	0.730279
PLP LPS	0.727081	0.727844	0.727733	-
PLP PS	0.695374	0.678421	0.676899	0.690103
LPC LPS	0.732215	0.729698	0.738738	-
LPC PS	0.739635	0.661300	0.708576	0.730068
LPCC	0.736193	0.733753	0.724141	-
PLPC.	0.713751	0.721172	0.693551	-

Table 1: *Results for each of the standard feature sets and distance measure combination, the table entries indicate the AUC.*

4.3. Features derived from phase spectra

The Fourier spectrum can be expressed in terms of its magnitude and phase spectra. To date, most distance measures reported are computed from the magnitude spectrum. In this study the phase spectrum was investigated in the form of the group delay function, $\tau(\omega)$, which is the negative of the derivative of the phase spectrum with respect to frequency, (3).

$$\tau(\omega) = -\frac{d}{d\omega}\{arg[X(j\omega)]\} \tag{3}$$

Accurately estimating the group delay function is sensitive to noise, window shape and length [10]. Blackman windows produced the best results employing a pitch synchronous analysis with a window length of one pitch period. No pre-emphasis filter was employed for estimating the group delay function.

A number of methods for computing the $\tau(\omega)$ were investigated: FFT, LPC and the Modified Group Delay Function [11]. The FFT based computation employed a 512-point FFT. The phase spectrum was extracted and subsequently unwrapped. The unwrapped phase spectrum was numerically differentiated by computing the difference between successive phase values. The LPC based GDF was computed employing LPC analysis of order 16. The phase spectrum of the LPC model was computed, unwrapped and differentiated. The MODGDF was implemented as in [11].

4.3.1. Results

The LPC based GDF was found to provide the highest AUC value when used in conjunction with the absolute distance as indicated in Table 2. Window length and the type of window used for feature extraction were found to significantly impact on the AUC values.

Features	l_1	l_2	Cos
GDF LPC	0.729783	0.686111	0.713704
MODGDF	0.669804	0.670570	0.696050
GDF FFT	0.660042	0.645949	0.597345

Table 2: *Results for GDF based measures computed from the phase spectrum, the table entries indicate the AUC.*

4.4. Limitations of standard measures

Each of the measures in this section employs a fixed length analysis window. Time resolution can only be improved by decreasing the window length but at the expense of frequency resolution and vice versa. For the Fourier transform this can be stated in terms of the axis scaling theorem, with scaling factor a.

$$x(at) \leftrightarrow \frac{1}{|a|} X\left(\frac{j\omega}{a}\right) \tag{4}$$

More generally, the Heisenberg uncertainty principle explicitly places a fundamental lower limit for the product of the time resolution and frequency resolution. The primary drawback of the Fourier transform is that both time and frequency resolution are constant across all frequency bands.

5. Wavelets

In order to overcome the limitations of the standard measures which adopt an analysis procedure with a fixed window length the wavelet transform was adopted. Wavelet analysis allows the time-frequency resolution to vary with respect to frequency. This allows spectral estimation with a time-frequency resolution adapted to each frequency band.

For a given wavelet mother function, ψ, the wavelet transform at scale a and position u, (5), is defined in (6) [12], were ψ^* denotes the complex conjugate of ψ.

$$\psi_{u,a}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-u}{a}\right) \tag{5}$$

$$Wx(u,a) = \int_{-\infty}^{+\infty} x(t)\psi^*_{u,a}(t)dt \tag{6}$$

Many possible wavelet functions, ψ, exist [13]. A number of these were tested for the task of detecting discontinuities in the test stimuli. In general it was found that most wavelets produced similar results.

A feature vector of 50 wavelet coefficients was used to represent each unit, these coefficients correspond with an analysis centered on the pitch pulse. Fig. 2 illustrates a wavelet based scalogram and a spectrogram across a concatenated join. The periodic nature of the scalogram illustrates the importance of employing feature vectors from the same relative position within a pitch period.

5.1. Results

The AUC values for most wavelets occupy a relatively narrow numerical range indicating that the success of the method is relatively independent of the choice of wavelet basis function. The Complex Morlet wavelet in conjunction with the Cos distance was found to have the highest AUC of all measures tested in this study. Fig. 3 compares ROC curves for a number of distance measures representative of standard, phase, wavelet based measures and a measure

Wavelet	l_1	l_2	Cos
Complex Morlet	0.771198	0.773598	0.794222
Symlet 4	0.778703	0.775279	0.788458
Morlet	0.770942	0.767512	0.784368
Complex Gaussian	0.778053	0.773773	0.776907
Gaussian	0.777922	0.775347	0.755023
Coiflet 2	0.776848	0.774638	0.772798
Daubechies 8	0.766776	0.762850	0.774815

Table 3: *Results for wavelet based measures, the table entries indicate the AUC.*

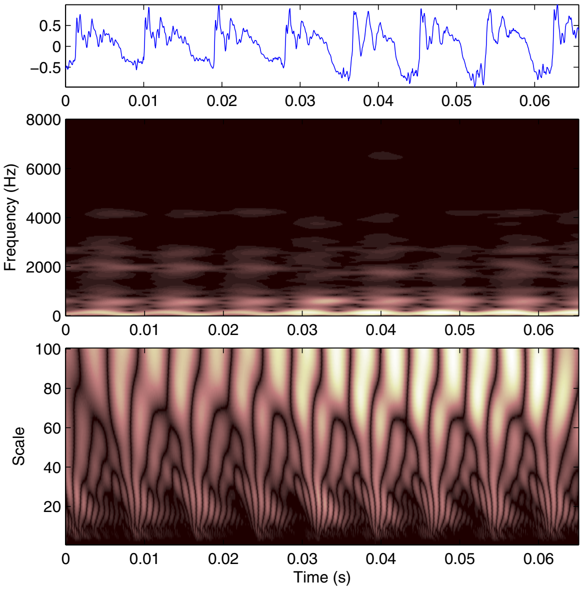

Figure 2: *Time domain signal, spectrogram and scalogram across a concatenated join respectively from top to bottom. With a frame size of 10 ms for the spectrogram and a frame shift of one sample for both spectrogram and scalogram.*

based on MFCCs extracted with a 40 ms analysis window, all with the l_2 distance.

6. Conclusion

This paper compared a set of standard measures used to detect spectral discontinuities in concatenated speech. Alternative measures based on phase spectra were introduced some of which were found to perform similarly to measures derived from the magnitude spectrum but ultimately share the same limitations. Wavelet based measures were introduced in order to overcome the limitations of standard measures and were found to outperform all standard measures. The results presented indicate that the strategy adopted for feature extraction has a more significant impact on the results than on selecting a specific acoustic feature set. The sensitivity of the results to window length may account for some of the inconsistencies in previous studies. This study suggests that the human auditory system is sensitive to subtle time variations in the spectral content of speech in vowel centres, regions which are typically assumed stationary.

7. Acknowledgments

This work is funded by Science Foundation Ireland, grant number 04/BRG/E0111. The authors would like to thank the participants in the perceptual listening test.

8. References

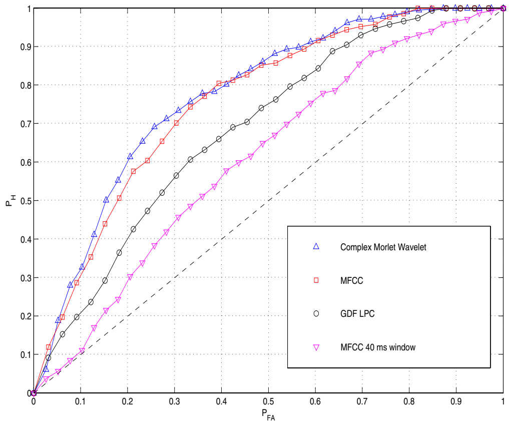

Figure 3: *A comparison of selected ROC curves.*

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

[2] E. Klabbers and R. Veldhuis, "Reducing audible spectral discontinuities," *IEEE Trans. on Speech and Audio Processing*, vol. 9, pp. 39–51, 2001.

[3] Y. Stylianou and A. K. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," in *Proc. ICASSP*, Salt Lake City, USA, 2001.

[4] J. Wouters and M. Macon, "Perceptual evaluation of distance measures for concatenative speech synthesis," in *Proc. ICSLP*, Sydney, Australia, 1998, vol. 6.

[5] J. Vepa and S. King, "Subjective evaluation of join cost functions used in unit selection speech synthesis," in *Proc. ICSLP*, Jeju, Korea, 2004.

[6] R. Duda and R. E. Hart, *Pattern Classification*, John Wiley and Sons, 2 edition, 2001.

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

[8] H. Hermansky, "Perceptual Linear Prediction (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, pp. 1738–1752, 1990.

[9] E. Klabbers and R. Veldhuis, "On the computation of the Kullback-Leibler distance measure for spectral distances," *IEEE Trans. on Speech and Audio Processing*, pp. 100–103, Jan. 2003.

[10] B. Bozkurt, B. Doval, C D'Alessandro, and T. Dutoit, "Appropriate windowing for group delay analysis and roots of z-transform of speech signals," in *Proc. EUSIPCO*, Vienna, Austria, 2004.

[11] H. A. Murthy and V. R. R. Gadde, "The modified group delay function and its application to phoneme recognition," in *Proc. ICASSP*, Hong Kong, 2003.

[12] Stéphane Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 1998.

[13] Ingrid Daubechies, *10 Lectures on Wavelets*, SIAM, Philadelphia, PA, 1992.