On the Role of Spectral Dynamics in Unit Selection Speech Synthesis

Barry Kirkpatrick, Darragh O'Brien, Ronán Scaife and Andrew Errity

Research Institute for Networks and Communications Engineering
Faculty of Engineering and Computing, Dublin City University
Dublin 9, Ireland

{bkirkpatrick, dobrien, aerrity}@computing.dcu.ie, scaifer@eeng.dcu.ie

10.21437/Interspeech.2007-543

Abstract

Cost functions employed in unit selection significantly influence the quality of speech output. Although unit selection can produce very natural sounding speech the quality can be inconsistent and is difficult to guarantee due to discontinuities between incompatible units. The join cost employed in unit selection to measure the suitability of concatenating speech units typically consists of sub costs representing the fundamental frequency and spectrum at the boundaries of each unit. In this study the role of spectral dynamics as a join cost in unit selection synthesis is explored. A number of spectral dynamic measures are tested for the task of detecting discontinuities. Results indicate that spectral dynamic measures correlate with human perception of discontinuity if the features are extracted appropriately. Spectral dynamic mismatch is found to be a source of discontinuity although results suggest this is likely to occur simultaneously with static spectral mismatch.

Index Terms: speech synthesis, join costs, auditory perception, spectral dynamics, feature extraction.

1. Introduction

Unit selection synthesis is currently considered state-of-the-art in text-to-speech synthesis. Synthetic speech is generated by concatenating units of speech which are selected from a large speech database. Cost functions are employed to select the optimum sequence of units. The quality of speech generated can be quite inconsistent; natural sounding speech is generated when the join between successive speech units is inaudible, much lower quality speech results when the transition between units sounds discontinuous. An audible discontinuity occurs when two units are not appropriately matched. Specific criteria for a perceptually continuous join remain undefined to date. Join costs currently employed in unit selection typically consist of f0 and spectral measures usually represented by mel-frequency cepstral coefficients (MFCC).

1.1. Background

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 [1–6]. Many of these studies have presented conflicting results, with measures that ranked highly in one study performing poorly in another. It is difficult to make direct comparisons between studies as each used a different database and different criteria to rank each measure. A consistent element in each of the studies is that the degree of correlation with human perception is often quite weak, and many studies report improvement in results with the inclusion of basic perceptual modelling.

The aforementioned studies predominantly focused on static spectral features as measures of spectral continuity. In Wouters and Macon [3] and Vepa and King [4] the static spectral measures were combined with corresponding delta coefficients to represent spectral change in the overall measure. The addition of delta coefficients was found to contribute minor improvements, of the order of 1-2% for the study of Wouters and Macon, depending on the feature set. Vepa and King found that adding delta features sometimes decreased performance. The resulting performance achieved by including delta features was inconsistent and relatively small when an improvement was achieved. This is in contrast to the successful use of delta features in other applications such as ASR [7] and HMM-based speech synthesis [8].

The perceptual importance of spectral dynamics is well documented in the literature, for example Furui [9]. The auditory system amplifies spectral bands that exhibit significant spectral change; a number of models have been proposed to mimic this behaviour both from physiological [10] and psychoacoustic perspectives [11]. Joins contained in diphthongs, which contain significantly more spectral dynamics than monophthongs, have been identified as having an increased likelihood of containing discontinuities [12]. Including spectral dynamic information in the join cost accounts for spectral movement on either side of the join. Such behaviour may result from co-articulation with neighbouring phones and is beyond the scope of local static spectral costs at unit boundaries.

It is understood that a number of different sources give rise to discontinuities in concatenated speech. From the studies reported to date it is unclear if spectral dynamic mismatch has a significant role in the perception of discontinuities. Furthermore it is unclear how to effectively incorporate spectral dynamic measures into join costs for unit selection. The objective in this paper is to identify the role of spectral dynamic behaviour in the perception of discontinuities and to investigate spectral dynamic measures for use in defining perceptually salient join costs.

In section 2 the spectral dynamic measures considered in this study are presented. The results for detecting discontinuities with the proposed spectral dynamic measures are presented in section 3, alongside a procedure to combine the static and dynamic spectral measures with the corresponding results. A further experiment and analysis are presented in section 4. The analysis is to quantify the degree of correlation between static and dynamic distance measures in the test database. The experiment is based on generating stimuli with a Klatt synthesiser to further explore the perceptual effect of spectral dynamic mis-

match when other sources of discontinuity are known to be absent. The conclusions are given in section 6.

2. Spectral dynamic cost functions

In order to investigate the potential of spectral dynamic information as a source of discontinuity, candidate spectral dynamic measures were tested for the task of detecting discontinuities in concatenated speech using the test database from [6]. The temporal variations of both resonant frequencies across the frequency axis and amplitudes within a particular spectral band were both investigated.

2.1. Feature extraction

The amplitude-based features were MFCCs and the frequency based features were Line Spectral Frequencies (LSFs). Other standard feature sets were also tested with similar results and are not presented here.

- LSFs were computed using the Burg algorithm to compute a 16^{th} order LPC model.
- MFCCs were computed as in Rabiner and Juang [7] using a total of 19 coefficients; the first cepstral coefficient was discarded.

For both feature sets the raw speech was pre-emphasised with the filter $H(z) = 1 - 0.95z^{-1}$ and the parameters were computed pitch synchronously with a 50 ms Hanning window. The features were computed on each pitch pulse throughout the voiced region for each of the words in the inventory, of which there is a total of 300. This results in parameter trajectories through the vowel centre. The spectral dynamic features were obtained by calculating the derivatives of the trajectories at the unit boundaries.

2.2. Trajectory modelling

Each parameter trajectory, Figure 1, was modelled using a polynomial of order 4. The polynomial coefficients were computed that best fit the parameter trajectories with respect to the mean squared error. Empirically it was found that nine data points gave an effective polynomial approximation which corresponds with nine pitch periods. Fitting too many data points, covering a longer time span, was found to have a negative impact on results due to over modelling in regions distant from the join and too few points did not effectively represent the spectral changes about the join.

The first and second derivatives with respect to time were computed at the unit boundaries from the estimated polynomials and were employed to represent the spectral dynamics of the corresponding units. As a baseline for comparison delta coefficients are also tested. The delta coefficients were computed using pitch synchronous analysis. A window length of one pitch period was employed to extract the features in the two pitch periods preceding the unit edge. The delta coefficients were generated by computing the difference between the feature vectors. Pitch synchronous feature extraction with a window length of one pitch period was found to be the optimum strategy for the task of detecting discontinuities with static spectral measures on the test database [6].

3. Experimental results

Each measure was tested for its ability to detect discontinuities independently and subsequently as part of a combined spectral measure with both static and dynamic features. The objective was to determine the degree of correlation of the dynamic measures, with human perception independently, and to further investigate if these measures could be successfully incorporated with static spectral measures to enhance performance. The results presented were computed by generating receiver operating characteristic (ROC) [13] curves that relate the perceptual results of human listeners with the proposed measures. The performance metric employed is the area under the ROC curve (AUC). The AUC represents the separability of the sets of continuous and discontinuous joins for each measure tested. Further details of the perceptual experiment and test procedure are contained in [6].

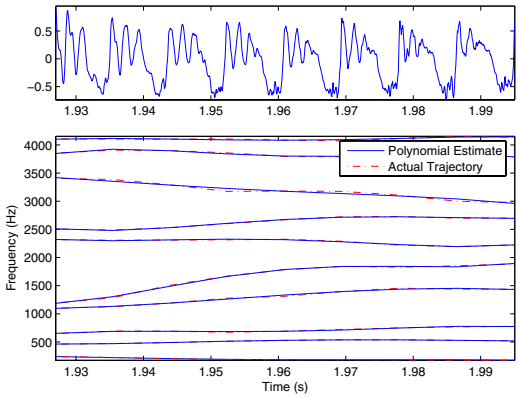

Figure 1: *Illustrating the LSF trajectories in the vowel centre of the word 'went'; the computed LSF parameters, the polynomial fit and the corresponding speech waveform in the time domain are shown.*

3.1. Detecting discontinuities with spectral dynamics

The results for the task of detecting discontinuities using spectral dynamic measures are presented in Table 1. The dynamic measures evaluated from the polynomial based derivatives are denoted by $d\mathbf{x}/dt$ and $d^2\mathbf{x}/dt^2$, the static measures by $\mathbf{x}$ and the delta features by $\Delta\mathbf{x}$. The Euclidean distance was used to quantify the degree of mismatch between feature vectors.

Features	$\mathbf{x}$	$\Delta\mathbf{x}$	$d\mathbf{x}/dt$	$d^2\mathbf{x}/dt^2$
MFCC	0.77458	0.52688	0.70860	0.65568
LSF	0.74238	0.50997	0.70801	0.67225

Table 1: *Comparison of results for each candidate measure of spectral dynamics; the table entries indicate the AUC value for each of the measures.*

It was found that adopting a longer window yielded better results with respect to the proposed spectral dynamic measures, this is counter to static measures for which a single pitch period was found to yield the best results. Longer window lengths gave rise to less variation between successive spectral estimates and in turn gave rise to smoother trajectories, which was found to yield spectral dynamic measures with higher AUC values.

The best spectral dynamic measure was obtained from the first derivative of the MFCC trajectories with an AUC value of 0.70860. The AUC values obtained for the delta coefficients was surprisingly low; just above the value of pure chance, which

Features	MFCC	LSF
$[\mathbf{x},\ d\mathbf{x}/dt]$	0.77540	0.74046
PCA $[\mathbf{x},\ d\mathbf{x}/dt]$	0.77540	0.74152
$[\mathbf{x},\ d^2\mathbf{x}/dt^2]$	0.77399	0.73964
PCA $[\mathbf{x},\ d^2\mathbf{x}/dt^2]$	0.77410	0.73964
$[\mathbf{x},\ d\mathbf{x}/dt,\ d^2\mathbf{x}/dt^2]$	0.77409	0.73858
PCA $[\mathbf{x},\ d\mathbf{x}/dt,\ d^2\mathbf{x}/dt^2]$	0.77516	0.73858

Table 2: *Comparison of results for combining static and dynamic spectral measures with and without the application of PCA; the table entries indicate the AUC values for each of the measures.*

corresponds with an AUC value of 0.5. The lower performance of the delta coefficients may be due to sensitivity to noise in numerical differentiation or perhaps that the measure reflects spectral change on a shorter time scale, due to shorter window lengths, which may not be relevant for the detection of discontinuities. The spectral dynamic measures based on the second derivatives were also found to correlate significantly with human perception, the results of which are also contained in Table 1. The results suggest that dynamic measures correlate with human perceptual results provided the features are appropriately extracted.

3.2. Combining static and dynamic features

The static and dynamic features are combined by concatenating the static and dynamic feature vectors.

$$\mathbf{x} = \left[\mathbf{x}_{static}^T,\ \mathbf{x}_{dynamic}^T\right]^T \tag{1}$$

When combining feature sets it is important to effectively utilise the discriminating information contributed by each feature vector. The individual feature vectors may be on considerably different scales and there may be significant correlation between the measures introducing redundancy. To overcome this problem we propose using a vector to represent a join [14], herein referred to as a join vector, $\mathbf{x}_{join}$ (2), constructed by subtracting the right unit feature vector from the left unit feature vector. This allows each join to be represented by a single vector and enables the application of a transform, $\mathbf{A}$, on the join vector to rescale and decorrelate the features with the intent of increasing the separability between the distributions of continuous and discontinuous joins. PCA [15] was applied to transform join vectors for each of the test stimuli in the database. This typically results in a reduced dimensionality representation with redundant information removed from the measure.

$$\mathbf{x}_{join} = \mathbf{x}_{left} - \mathbf{x}_{right} \tag{2}$$

$$\mathbf{x}_{pca} = \mathbf{A}_{pca}\mathbf{x}_{join} \tag{3}$$

$$\mathbf{D} = \|\mathbf{x}_{pca}\| \tag{4}$$

The distance measure, D, is the norm of the transformed join vector, which is the distance from the origin and is a measure of the spectral mismatch between two units.

The results from the combined spectral measures are illustrated in Table 2. The gain in performance from adding the spectral dynamic measure is low and can even cause a decrease in performance in some cases. PCA is found to have a marginal

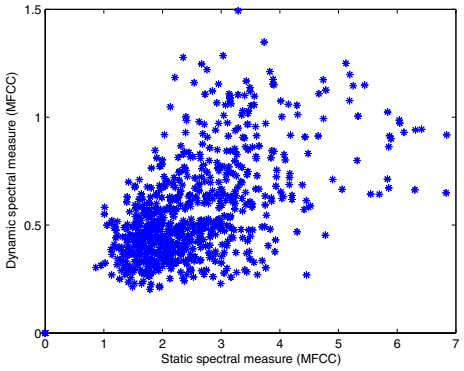

Figure 2: *Scatter plot of the dynamic spectral measure versus the static spectral measure for MFCCs, the dynamic measure is from the* 1^{st} *derivative.*

improvement in some cases and no impact on the result in other cases. The dimensionality of the transformed features was chosen empirically for each feature set to maximise the AUC value. In some cases reducing the dimensionality did not improve the AUC value and the original dimension was maintained, this arose with MFCCs. When this occurs PCA effectively rotates the location of the feature vector about the origin in the feature space; this has no impact on the norm of the feature vector which is employed as the distance measure and subsequently has no impact on the AUC value.

The increases in results from the combined spectral measures were found to be quite small in comparison to their individual performance. The application of PCA was found to have little impact on the results.

4. Spectral dynamics as a source of discontinuity

In the previous section it was found that measures of spectral dynamics correlated with human perception of discontinuity. The gain in performance when static and dynamic measures were combined was disappointing. In this section we investigate spectral dynamics as an independent source of discontinuity.

4.1. The correlation between static and dynamic features

Motivated by the disappointing increase in results due to the addition of spectral dynamic measures an analysis was conducted to investigate if a correlation exists between static and dynamic spectral mismatch in the test database. If static and dynamic mismatches are correlated then they are likely to occur simultaneously, it is unlikely that a significant increase in performance will result from combining these measures in such a scenario.

The degree of correlation between static and dynamic measures was quantified using the correlation coefficient (5).

$$r_{xy} = \frac{\sum_{n=1}^{N}(x - m_x)(y - m_y)}{(N-1)\sigma_x\sigma_y} \tag{5}$$

The dynamic and corresponding static distances are represented by the vectors x and y, were m_x, m_y, σ_x and σ_y represent the mean and standard deviation of the distance vectors.

Features	MFCC	LSF
$d\mathbf{x}/dt$	0.7748	0.7345
$d^2\mathbf{x}/dt^2$	0.7522	0.7272

Table 3: *Correlation coefficients between static and dynamic spectral measures for both LSFs and MFCCs.*

The correlation between static and dynamic spectral measures are presented in Table 3. The table entries indicate the correlation coefficients computed between the static and dynamic measures for both MFCCs and LSFs using measures of dynamics computed from both the first and second derivatives. The results indicate that there is significant correlation between static and dynamic spectral mismatch for the test database, with correlations coefficients of the order of 0.7. A scatter plot of static spectral distance versus dynamic spectral distance is illustrated in Figure 2 for MFCCs. This may explain why the improvement for combining the measures is small compared to their individual performance.

4.2. Perception of formant dynamics

In order to further clarify the role of spectral dynamics a number of exploratory experiments were conducted using stimuli generated with a Klatt synthesiser. With this framework it was possible to create stimuli that contained spectral dynamic mismatch with continuity maintained with respect to all other potential sources of discontinuity, such as f0 and static spectral mismatch. This effectively allows the spectral dynamic behaviour to be isolated as the only possible source of discontinuity and enables human listeners to observe the resulting phenomena. Such a scenario gives rise to a small static spectral distance and a large spectral dynamic distance, any resulting discontinuities can only be detected from spectral dynamic measures.

Stimuli were created to mimic potential spectral dynamic mismatch in concatenated speech. Each of the formant trajectories were manipulated individually to introduce a spectral dynamic mismatch in one formant trajectory at a time, while all the remaining formant trajectories were kept at a constant value. Informal listening tests revealed that the test stimuli contained audible abnormalities about the point of spectral dynamic mismatch consistent with what may be expected from a mismatch in co-articulation. Minor degrees of mismatch in spectral dynamics did not produce such artifacts. It was also found that such artifacts where more prominent in F1 and became perceptually less noticeable with higher formants. The stimuli used in this experiment were significantly different to natural speech, although they effectively served the intended purpose of isolating spectral dynamic behaviour.

5. Conclusions

In this study it was found that measures of spectral dynamics correlated with human perception of discontinuity in concatenated speech. The degree of correlation with human perception was found to be sensitive to basic parameters used in feature extraction, the general trend indicated that smooth trajectories that do not contain the fine detail of temporal variations are more suitable for the task at hand. Delta coefficients computed using short window lengths were found to perform poorly. This is in contrast to results for static spectral measures on the same database [6] in which temporal resolution was found to have a significant impact for the detection of discontinuities.

The gain in performance for the task of detecting discontinuities when standard measures were combined with dynamic measures was quite small. Synthetic stimuli generated using a Klatt synthesiser supports the claim that spectral dynamic mismatch is an independent source of discontinuity, although analysis revealed that there is significant correlation between measures of static and dynamic spectral distances in the test database. This suggests that discontinuities due to spectral dynamics are likely to occur simultaneously with discontinuities resulting from static spectral mismatch. This may explain why combining these measures does not result in a significant gain in performance.

6. Acknowledgements

This work is funded by Science Foundation Ireland, grant number 04/BRG/E0111. Andrew Errity is supported by the Irish Research Council for Science, Engineering and Technology; grant number RS/2003/114.

7. References

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

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

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

[4] J. Vepa and S. King, "Subjective evaluation of join cost and smoothing methods for unit selection speech synthesis," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 1763 – 1771, 2006.

[5] E. Klabbers, J. P. H. van Santen, and A. Kain, "The contribution of various sources of spectral mismatch to audible discontinuities in a diphone database," *IEEE Transactions on audio speech and language processing*, vol. 15, pp. 949 – 956, March 2007.

[6] B. Kirkpatrick, D. O'Brien, and R. Scaife, "Feature extraction for spectral continuity measures in concatenative speech synthesis," in *Proc. ICSLP*, Pittsburgh, USA, 2006.

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

[8] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from hmm using dynamic features," in *Proc. ICASSP*, Detriot, USA, 1995.

[9] S. Furui, "On the role of spectral transitions for speech perception," *J. Acoust. Soc. Am.*, vol. 80, pp. 1016–1025, 1986.

[10] R. Meddis, M. Hewitt, and T. Shackleton, "Implementation details of a computational model of the inner hair-cell/auditory-auditory nerve synapse," *J. Acoust. Soc. Am.*, vol. 87, pp. 1813 – 1816, 1990.

[11] T. Dau and D. Puschel, "A qualitative model of the "effective" signal processing in the auditory system," *J. Acoust. Soc. Am.*, vol. 99, pp. 3615–3622, 1996.

[12] A. K. Syrdal, "Prosodic effects on listener detection of vowel concatenation," in *Proc. EUROSPEECH*, Scandinaivia, 2001.

[13] R. Duda and R. E. Hart, *Pattern Classification*, 2nd ed. John Wiley and Sons, 2001.

[14] B. Kirkpatrick, D. O'Brien, and R. Scaife, "A comparison of spectral continuity measures as a join cost in concatenative speech synthesis," in *Proc. of the IET Irish Signals and Systems Conference (ISSC)*, Dublin, Ireland, 2006.

[15] I. Jolliffe, *Principal Component Analysis*. Springer-Verlag, 1986.