Feature Transformation Applied to the Detection of Discontinuities in Concatenated Speech

Barry Kirkpatrick, Darragh O'Brien and Ronán Scaife

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

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

Abstract

The quality of concatenated speech depends on the degree of mismatch between successive units. Defining a perceptually salient join cost to represent the degree of mismatch has proven to be a difficult task. Such a join cost is critical in unit selection synthesis to ensure that the optimum sequence of speech units is selected from the units available in the speech inventory. In this study the problem of defining a join cost is extended to include a feature transformation stage. Two feature transformations are considered, principal component analysis and a neural network-based approach. Each transformation was investigated for its ability to improve the detection of discontinuities in concatenated speech for a given feature set. The results indicate that a feature transformation combining principal component analysis as a preprocessing stage to a neural network-based transformation can increase the rate of detection of discontinuities. The neural network was trained using perceptual data obtained from a subjective listening test indicating if a join is continuous or discontinuous. The highest scoring measure based on this strategy provided a correlation with perceptual results of 0.8859 compared with a value of 0.7576 over the baseline MFCC measure on the same test data set.

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, also many studies report improvement in results with the inclusion of basic perceptual modelling.

The aforementioned studies predominantly focused on a comparison of standard speech parametrisations as measures of spectral continuity typically based on representations found to be useful in automatic speech recognition and coding. Both Bellegarda [7] and Vepa and King [8] have tailored specific strategies for the problem of defining spectral join costs. Bellegarda developed an alternative transform approach based on a singular value decomposition of speech frames extracted about the points of concatenation in the speech inventory. Vepa and King developed a Kalman filter based strategy that measured the degree of mismatch between idealised trajectories predicted by the Kalman filter and the actual trajectories about the point of concatenation.

In this study feature transformations are investigated to enhance the ability of existing spectral measures to detect discontinuities, specifically principal component analysis (PCA) and neural networks. This extends the existing distance measure framework to a feature space based framework and enables the application of feature space transformations. The objective is to maximally exploit the discriminating information in the features extracted with the proposed transformations and as a result determine a spectral join cost that correlates better with human perception of discontinuity.

1.2. Motivation

In unit selection systems the spectral join cost is computed by extracting spectral features from speech frames adjacent to the unit boundaries and calculating the Euclidean distance between the features. In this computation the level of spectral mismatch between corresponding features is treated equally for all features. Perceptually it is unlikely that all features are equally significant. Mismatch below a certain threshold is likely to be perceptually irrelevant and should be discarded with no contribution to the overall distance measure. Certain spectral bands may be of more significance, for example mismatch coinciding with the location of a formant would be expected to be of more perceptual importance than mismatch in other regions of

the spectrum. It has been reported that an abrupt increase in an acoustic component is more perceptually significant than a sudden drop in amplitude [9], this indicates that mismatch due to the introduction of a new component should be weighted more heavily than mismatch due to a drop in component energy. With the application of neural networks a mapping can be learned from data provided from subjective listening tests relating continuous and discontinuous joins with input feature vectors representing a join. The appropriate weighting of the features is data driven and does not require advanced knowledge of auditory processing.

The testing procedure to quantify the performance of the proposed techniques is outlined in section 2. Section 3 introduces the background associated with generalising the distance measure approach to a feature space representation and the application of feature transformations. The application of PCA and neural networks as feature transformations are also discussed in section 3. Section 4 contains the results from employing PCA and neural networks to transform features for the task of detecting discontinuities in concatenated speech. Section 5 contains discussion and conclusions.

2. Testing

To test each proposed technique it is necessary to correlate the perceptual response of a human listener with each candidate measure. This enables a comparison of the standard spectral distance for a given feature set with the proposed measure after the feature transformation has been applied.The evaluation of each measure was conducted using the database from [5] and the corresponding perceptual results. The perceptual stimuli consisted of 1800 monosyllabic words. Each of these words was generated by concatenating two half words with the same vowel nucleus. The inventory of units consisted of 300 words recorded from an adult male. The inventory of 300 words consisted of 50 sets of 6 words. Within each set the words share the same vowel nucleus and differ in the final or initial consonant. The perceptual test required the listeners to make a forced decision for each test word: continuous or discontinuous. Twelve listeners in total contributed perceptual results with coverage of three listeners per subtest. A majority scoring system was employed to indicate if a test word was continuous or discontinuous.

In order to test the performance of the neural network-based measures the database was split into training and testing subsets. The training set contained 50% of the database and the remaining 50% made up the testing set. The database was divided to have an equal number of discontinuities in both the training and testing sets. The database contained a total of 434 discontinuities. The database was split such that joins contained in each vowel type represented in the database are equally spread between training and testing subsets.

3. Feature transformations

Many studies have been conducted in the automatic speech recognition literature investigating the use of feature transformations to improve the discriminating qualities of the features for speech recognition [10, 11]. In this study PCA [12] and neural networks [13] are applied to transform features representing the spectral join cost in concatenated speech. The objective is to investigate the ability of these techniques to enhance existing measures for objective detection of discontinuities.

3.1. Defining a join vector

In order to apply feature transformations that fully exploit the discriminating information within each feature, it is necessary to define a suitable vector to represent a join. Existing methods employ a distance to represent a join and feature vectors to represent individual units of speech. In this study the error vector is used to represent a join, hereby referred to as the join vector, which is computed by subtracting the left and right unit feature vectors, $\mathbf{x}_{left}$ and $\mathbf{x}_{right}$. Each feature in the join vector represents the degree of mismatch between the corresponding features in the left and right units.

$$\mathbf{x}_{join} = \mathbf{x}_{left} - \mathbf{x}_{right} \tag{1}$$

Different strategies to construct join vectors motivated by the standard, l_p norms and the symmetric Kullback-Leibler were investigated in [14]. The join vector resulting from the subtraction of the left and right features was found to be suitable. This generalises the standard distance measure approach for the l_p norms. Classification of the join vectors in the feature space without further processing for the join vectors constructed using equation (1) corresponds exactly with calculating the l_p distance between the original left and right feature vectors, for a given p, equation (2).

$$l_p(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{N} |x(i) - y(i)|^p \right)^{1/p} \tag{2}$$

The ideal join should correspond with the origin in the feature space and the quality of the join can be quantified as the distance of the join vector from the origin. Thus joins can be classified as continuous or discontinuous with respect to distance from the origin. When classification is based on a distance from the origin, the subsequent choice of norm for the feature space establishes the geometry of the classifier. The classifiers corresponding to the l_1, l_2, l_4 and l_∞ norms are illustrated in Fig. 1. This illustrates how the standard distance measure can be interpreted in the feature space.

With the join vector representation it is possible to apply a transformation, $\mathbf{A}$, on the join vector before computing the final measure of mismatch, equation (3).

$$\mathbf{X} = \mathbf{A}(\mathbf{x}_{join}) \tag{3}$$

With this approach standard techniques can be applied to increase the separability between join vectors representing continuous and discontinuous joins. The application of a linear feature transformation is equivalent to stretching or contracting the individual axes and rotating the classifiers in Fig. 1, with a possible reduction in dimensionality. The final measure of mismatch, D, can be computed from the transformed vector, $\mathbf{X}$.

$$\mathbf{D} = \|\mathbf{X}\| \tag{4}$$

Two techniques were investigated; PCA and a neural network-based approach. For the neural network-based approach PCA was used as a preprocessing stage for dimensionality reduction of the input data. Ideally a feature transformation will remove redundant information and weight perceptually important information resulting in improved discrimination between continuous and discontinuous joins.

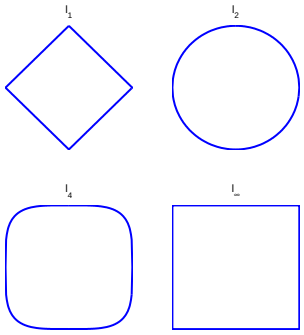

Fig. 1: Classifier shape in 2 dimensions employing l_1, l_2, l_4 and l_∞ norms.

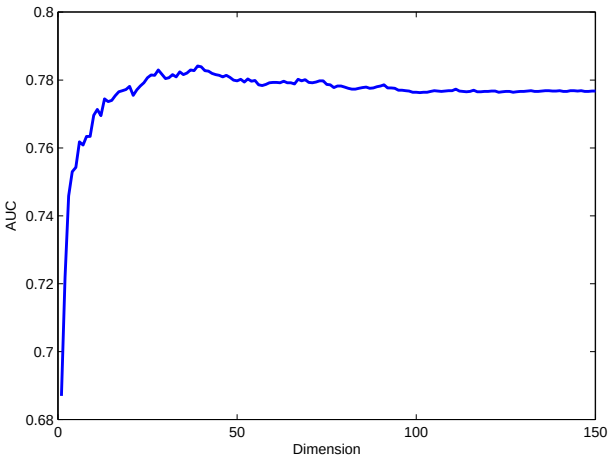

Fig. 2: Plot of AUC value as the output dimension from PCA is varied using Log PS.

3.2. Principal component analysis

Principal component analysis is an unsupervised learning technique and does not require splitting the database. In this study PCA is investigated for two roles; firstly for its ability as an unsupervised learning technique to improve the detection of discontinuities and secondly as a dimensionality reduction technique to remove redundant information preceding the application of neural networks. The removal of redundant information with PCA often leads to an improvement in performance in many pattern recognition tasks [12].

In the implementation of PCA the data is centred in the feature space about the origin by subtracting the mean vector computed over the complete database. The data is also normalised with respect to variance such that the standard deviations are equal to one. The normalised data is transformed using PCA, this produces transformed join vectors whose components are uncorrelated and ordered according to the magnitude of their variance.

3.3. Neural networks

For each of the feature sets considered and for each possible combination of feature sets a corresponding neural network is trained from the training set of the database. PCA is applied as a preprocessing step to reduce the dimensionality of the input vectors before training the networks. When the join vector is passed through the neural network a distance measure is output. To train the neural networks join vectors corresponding with discontinuities are assigned an output value of 1 and continuous joins are assigned an output value of 0.

A number of neural network architectures were investigated for the task of detecting discontinuities. General regression neural networks (GRNN) [15] were found to be the most suitable for the task. Feedforward neural networks were investigated but were found to be less consistent than GRNNs. GRNNs do not suffer from the problem of getting trapped in local minima, which can be a problem with iteratively trained neural networks.

4. Results

The results presented were computed by generating receiver operating characteristic (ROC) curves [16] that relate the perceptual results of human listeners with the proposed measures. Two probability density functions, $p(\tau|1)$ and $p(\tau|0)$, are estimated for each distance measure, τ, based on the perceptual results for continuous (0) and discontinuous (1) joins. The ROC curves were calculated from the probability density functions and provide information regarding the separability of $p(\tau|1)$ and $p(\tau|0)$, for each distance measure.The ROC curves are generated by plotting the hit rate, P_H, against the false alarm rate, P_{FA}.

$$P_H(\tau_0) = \int_{\tau_0}^{\infty} p(\tau|1)d\tau \quad (5)$$

$$P_{FA}(\tau_0) = \int_{\tau_0}^{\infty} p(\tau|0)d\tau \quad (6)$$

The performance metric employed was the area under the ROC curve (AUC). The AUC represents the separability of the sets of continuous and discontinuous joins for each measure tested. The AUC values are presented for before and after the application of the proposed transforms for each feature set tested.

4.1. Features

The features employed were the log power spectra (Log PS) computed from the fast Fourier transform (FFT), MFCCs and Line spectral frequencies (LSF). They were all extracted using a frame of one pitch period in length with a Hanning window. The MFCCs were computed from FFT spectra and the LSFs were computed from a 16th order LPC analysis on a Mel scale .

4.2. PCA

The results comparing the AUC values computed before and after the application of PCA are presented in Table 1. These results were computed across the entire dataset as the database did not require the separation into training and testing for the application of PCA.

Features	$\mathbf{x}$	$PCA[\mathbf{x}]$
MFCC	0.75	0.7696
LSF	0.7381	0.6966
Log PS	0.7615	0.7841

Table 1: Comparison of results with and without the application of PCA for each feature set; the table entries indicate the AUC value.

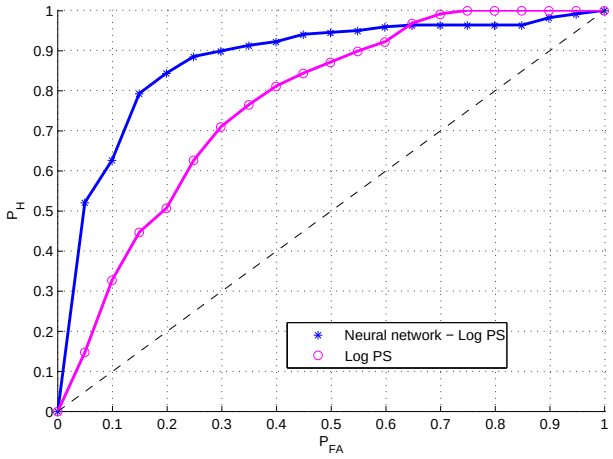

Fig. 3: Illustrating the ROC curves computed from the Log PS before and after applying the neural network.

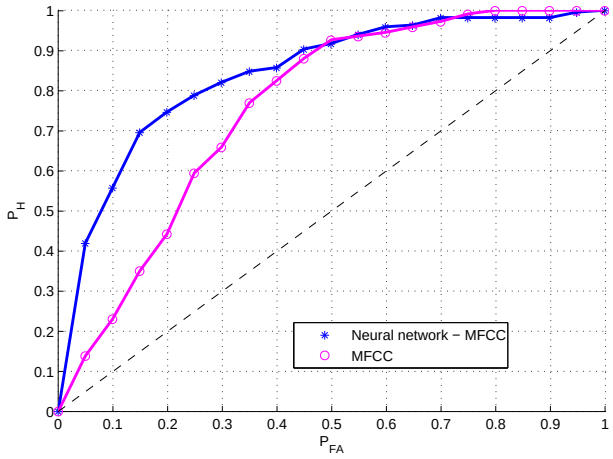

Fig. 4: Illustrating the ROC curves computed from MFCCs before and after applying the neural network.

Features	$\mathbf{x}$	$ANN[\mathbf{x}]$
Log PS	0.7673	0.8744
MFCC	0.7565	0.8413
LSF	0.7468	0.7955

Table 2: Comparison of results before and after the application of the neural network for each feature set; the table entries indicate the AUC value.

For both MFCCs and Log PS the application of PCA was found to improve the rate of detection of discontinuities. For LSFs, PCA was found to result in a decrease in the AUC value. The dimension of the transformed vector was chosen to maximise the AUC value for each of the feature sets. Figure 2 illustrates the resulting AUC values as the dimension of the transformed vector is varied for the case of join vectors constructed from Log PS. The maximum AUC value in Figure 2 occurs for a dimension of 39. This indicates how effective PCA is at retaining the discriminating information in relatively few dimensions; the original dimension was 256. This justifies the use of PCA as a preprocessing stage prior to applying neural networks. For MFCCs the maximum AUC value was obtained at a dimension of 3; the original dimension was 19. For LSFs the maximum AUC value corresponded with the maximum possible dimension of 16, although the AUC value essentially plateaued at a dimension of 4 (AUC = 0.6953 for dimension 4).

4.3. Neural networks

The results for each of the feature sets before and after the application of the proposed neural network-based transformation are presented in Table 2. The neural network is trained on the training set and tested in a separate testing set. PCA is employed as a preprocessing stage to reduce the dimensionality of the input vectors. The results presented are for GRNN type networks.

Table 2 indicates that the neural network-based approach significantly enhances the performance, this is most notable for Log PS in which the AUC value increased from 0.7673 with the standard distance measure approach to a value of 0.8744 with the proposed approach. The ROC curves comparing each of these measures before and after they were passed through their respective neural networks are illustrated in Figures 3, 4 and 5.

4.4. Combined measures

To combine the measures each join vector is concatenated and subsequently PCA is applied, at this point the neural network is trained. The results computed from the standard distance measures with no transformation based on the concatenated join vectors and those computed from the neural network-based measures are presented in Table 3.

The neural network-based measure based on both MFCC and Log PS features is the best performing measure tested with an AUC value of 0.8859. For each of the combined measures the neural network-based measure outperforms the corresponding standard measure. The ROC curves comparing the performance of MFCCs combined with Log PS features before and after they were passed through the neural network are illustrated in Figure 6.

Features	$\mathbf{x}$	$ANN[\mathbf{x}]$
MFCC + LSF	0.7468	0.8581
MFCC + Log PS	0.7673	0.8859
LSF + Log PS	0.7517	0.8753
LSF + MFCC + Log PS	0.7517	0.8829

Table 3: Comparison of results with and without the application of the ANN for possible combination of feature sets; the table entries indicate the AUC value.

5. Discussion and conclusions

This paper discusses a framework for applying feature transformations to spectral features for join cost optimisation in concatenative speech synthesis. PCA and a neural network-based strategy were investigated. The results indicate that PCA can be employed as an effective mechanism for dimensionality reduction without losing critical information in the detection of discontinuities. PCA does not always provide an increase in the performance as illustrated in the results for the LSF-based measure. The potential gain in performance is relatively small when it does occur. Perceptual data is required to optimally select the output dimension. The neural network-based measures were found to outperform the corresponding standard distance

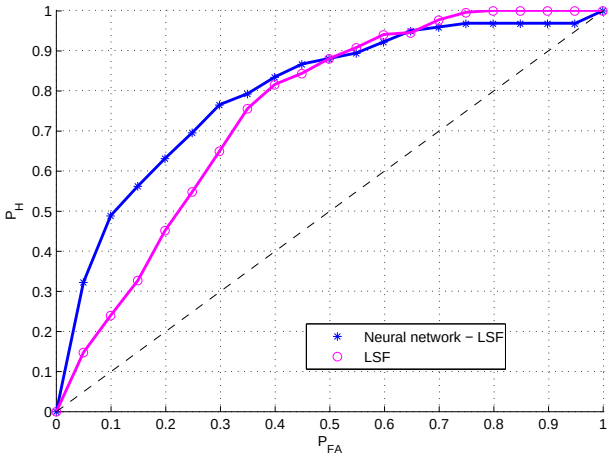

Fig. 5: Illustrating the ROC curves computed from LSFs before and after applying the neural network.

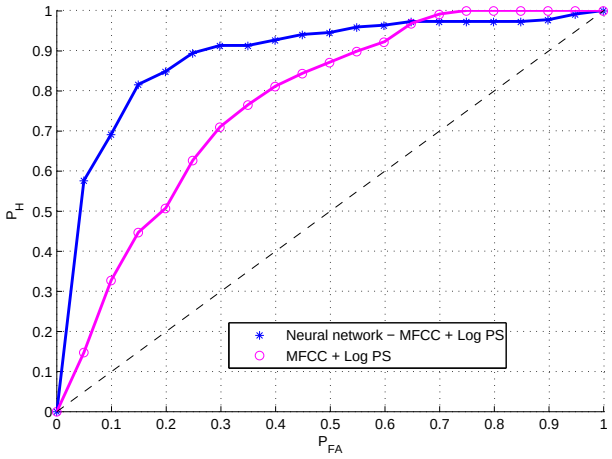

Fig. 6: Illustrating the ROC curves computed from the combined features from the Log PS and MFCCs before and after applying the neural network.

measure approach for each feature set tested and can be employed to enhance an existing feature set for its ability to detect discontinuities. The neural network-based strategy provided the best results of all measures tested and produced the highest detection rates on the test database to date. This suggests that the proposed feature transformation framework used in conjunction with neural networks is an effective strategy to learn the levels of mismatch that give rise to discontinuities for a given feature set. A critical issue with the proposed strategy is that training the neural network requires perceptual data which requires conducting perceptual experiments. This is a laborious and difficult task; most studies that involved listening experiments for the detection of discontinuities reported that the listeners found the task difficult.

6. Acknowledgements

This work is funded by Science Foundation Ireland, grant number 04/BRG/E0111.

7. References

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

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

[6] 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.

[7] J. Bellegarda, "A global, boundary-centric framework for unit selection text-to-speech synthesis," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 990 – 997, 2006.

[8] J. Vepa and S. King, "Kalman-filter based join cost for unit-selection speech synthesis," in *Proc. EUROSPEECH*, Geneva, Switzerland, 2003.

[9] Q. Summerfield, A. Sidwell, and T. Nelson, "Auditory enhancement of changes in spectral amplitude," *J. Acoust. Soc. Am.*, vol. 81, pp. 700 – 708, 1986.

[10] P. Somervuo, B. Chen, and Q. Zhu, "Feature transformations and combinations for improving ASR performance," in *Proc. EUROSPEECH*, Geneva, Switzerland, 2003.

[11] P. Somervuo, "Experiments with linear and nonlinear feature transformtions in hmm based phone recognition," in *Proc. ICASSP*, Hong Kong, 2003.

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

[13] C. Bishop, *Neural Networks for Pattern Recognition*. Oxford, 1995.

[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] D. F. Specht, "A general regression neural network," *IEEE Trans. on Neural Networks*, vol. 2, 1991.

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