Dimensionality Reduction Methods Applied to both Magnitude and Phase Derived Features

Andrew Errity, John McKenna and Barry Kirkpatrick

School of Computing
Dublin City University, Dublin 9, Ireland
{andrew.errity, john.mckenna, barry.kirkpatrick}@computing.dcu.ie

10.21437/Interspeech.2007-48

Abstract

A number of previous studies have shown that speech sounds may have an intrinsic low dimensional structure. Such studies have focused on magnitude-based features ignoring phase information, as is the convention in many speech processing applications. In this paper dimensionality reduction methods are applied to MFCC and modified group delay function (MODGDF) features derived from the magnitude and phase spectrum, respectively. The low dimensional structure of these representations is examined and a method to combine these features is detailed. Results show that both magnitude and phase derived features have a low dimensional structure. MFCCs are found to offer higher accuracy than MODGDFs in phone classification tasks. Results indicate that combining MFCCs and MODGDFs gives improvements for phone classification. PCA is shown to be capable of efficiently combining MFCCs and MODGDFs for improved classification accuracy without large increases in feature dimensionality.

Index Terms: modified group delay function, phase, dimensionality reduction, manifold learning, phone recognition.

1. Introduction

1.1. Short-time phase spectrum

Spectral features parametrising speech are conventionally extracted from the magnitude spectrum, derived from the short-time Fourier transform (STFT) of the speech signal. The phase spectrum also resulting from the STFT is conventionally ignored due to the common belief that the phase spectrum does not play a significant part in human auditory perception over the small time frames used in STFT analysis. However a number of recently performed studies have shown that the short-time phase spectrum is useful in human speech perception [1].

It is, however, difficult to extract useful features from the STFT phase spectrum due to problems with phase unwrapping and zeros of the signal's z-transform close to the unit circle [2]. The group delay function [3] has been used to represent the phase spectrum in a number of speech processing applications in the past [2]. A modified group delay function (MODGDF) based feature set has been proposed by Murthy and Gadde [4]. This modification suppresses zeros of the z-transform of the signal that are close to the unit circle and cause the group delay function to become undefined. The MODGDF can be converted to cepstral coefficients, in a similar manner to the conventional MFCC derived from the power spectrum, for use in tasks such as speech recognition. MODGDF features have recently been shown to be useful in various speech processing applications [5].

1.2. Intrinsic low dimensional structure of speech

The degrees of freedom of the speech production apparatus are limited by physiological constraints on articulatory movement. As a result humans are only capable of producing sounds occupying a subspace of the entire acoustic space. Thus, speech data can be viewed as lying on or near a low dimensional manifold embedded in the original high dimensional acoustic space.

The underlying dimensionality of speech has been the subject of much previous research [6–9]. The consensus of this work is that some speech sounds, particularly voiced speech, are inherently low dimensional. Previous studies have analysed the underlying structure of features derived from the magnitude spectrum, ignoring phase information. In this work phase spectrum derived features are examined to determine if the phase spectrum encodes similar information to the magnitude spectrum and to investigate if the phase spectrum has a similar low dimensional structure to the magnitude spectrum.

To accomplish this we apply dimensionality reduction methods, which aim to discover underlying low dimensional structure, to MFCC and MODGDF features. These methods can be categorised as either linear or nonlinear. Linear methods are limited to discovering the structure of data lying on or near a linear subspace of the high dimensional input space. Principal component analysis (PCA), one of the most widely used linear dimensionality reduction methods, is used in this study.

However if speech data occupies a low dimensional submanifold nonlinearly embedded in the original space, as proposed previously [6–9], linear methods will fail to discover the low dimensional structure. A number of manifold learning, also referred to as nonlinear dimensionality reduction, algorithms have been developed [10, 11] which overcome the limitations of linear methods. Manifold learning algorithms have recently been shown to be useful in a number of speech processing applications [7–9]. The isometric feature mapping (Isomap) [11] manifold learning algorithm is used in this investigation.

In this work, features output by these dimensionality reduction methods, along with the baseline MFCC and MODGDF features, are evaluated in phone classification tasks. These classification experiments are primarily used as a means of evaluating how much meaningful discriminatory information is contained in the low dimensional representations produced by each method. These experiments also serve to display the potential value of these methods in speech processing applications.

In addition to using each feature set alone, MFCC and MODGDF features have previously been concatenated and the resulting feature vectors shown to improve speech recognition performance, indicating that they may contain complementary information [12,13]. We examine the performance of these joint features in phone classification tasks and propose a method to

reduce the dimensionality of these joint features in an attempt to improve classification accuracy without increasing the computational cost associated with processing the features.

1.3. Paper layout

This paper is structured as follows. In Section 2 the MODGDF is described. Section 3 details the experimental procedure used. Results are examined and discussed in Section 4, with conclusions presented in Section 5.

2. Modified group delay function

The group delay function (GDF) is the negative derivative of the phase spectrum, $\theta(\omega)$, with respect to frequency, ω:

$$\tau(\omega) = -\frac{d(\theta(\omega))}{d\omega} \tag{1}$$

The GDF can be computed from the speech signal **x** as follows [3]:

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{|X(\omega)|^2} \tag{2}$$

where $X(\omega)$ and $Y(\omega)$ denote the Fourier transforms of $x(n)$ and $nx(n)$, respectively. The real and imaginary parts of the Fourier transform are indicated by the subscripts R and I.

As mentioned previously, the GDF is undefined when the roots of the signal's z-transform are close to the unit circle. The MODGDF [4] overcomes this problem by substituting $S(\omega)$, a cepstrally smoothed version of $|X(\omega)|$, in place of the same:

$$\tilde{\tau}(\omega) = \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{S(\omega)^2} \tag{3}$$

A further two parameters, α and γ, are introduced [4] to reduce the spiky nature of the formant peaks, relative to the magnitude spectrum, giving the final MODGDF definition:

$$\tilde{\tau}_{\alpha,\gamma}(\omega) = \frac{\tilde{\tau}_\gamma(\omega)}{|\tilde{\tau}_\gamma(\omega)|}\,|\tilde{\tau}_\gamma(\omega)|^\alpha \tag{4}$$

where

$$\tilde{\tau}_\gamma(\omega) = \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{S(\omega)^{2\gamma}} \tag{5}$$

A comparison of the magnitude spectrum, GDF, and MODGDF for a single frame of speech is shown in Fig. 1. It can be seen that the magnitude spectrum and MODGDF capture similar information.

To produce a feature set more suitable for applications such as speech recognition cepstral coefficients can be computed from the MODGDF using a discrete cosine transform (DCT), in a similar manner to the conventional MFCC computation. All subsequent references to the MODGDF refer to cepstral coefficients produced by applying a DCT to the output of Equation (4).

In all the MODGDF computations in this study the parameters were set as $\alpha = 0.4$ and $\gamma = 0.9$. A lifter window of length 8 was used for cepstral smoothing. Details of these are discussed in Hegde et al. [5].

3. Experiments

The objective of these experiments is to perform phone classification using MFCCs, MODGDFs, and the low dimensional feature representations resulting from the application of both PCA and Isomap to these baseline features.

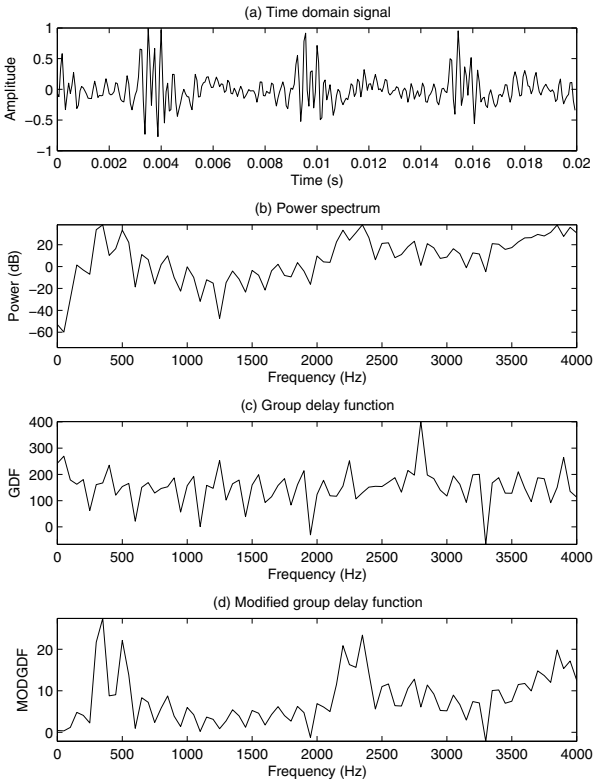

Figure 1: *Comparison of magnitude and phase spectrum representations of a frame of speech taken from an 'iy' vowel sound. The (a) time domain signal, (b) power spectrum, (c) group delay function, and (d) modified group delay function are shown.*

3.1. Data

The speech data used in this study was taken from the TIMIT corpus. This corpus contains 6300 utterances, 10 spoken by each of 630 American English speakers. The speech recordings are provided at a sampling frequency of 16 kHz.

3.2. Classification tasks

Each feature representation was evaluated in two phone classification experiments. The first experiment involved distinguishing between a set of ten vowels ('aa', 'iy', 'uw', 'eh', 'ae', 'ah', 'ay', 'oy', 'ih', and 'ow'). Phones are labeled using TIMIT symbols. The second test involved classifying a set of nineteen phones into their associated phone classes. The phone classes and phones used were: vowels (listed above), fricatives ('s', 'sh'), stops ('p', 't', 'k'), nasals ('m', 'n') and, semivowels and glides ('l', 'y').

3.3. Parameter extraction

Based on the phonetic transcriptions and associated phone boundaries provided in TIMIT all units of a subset of phones, listed in Section 3.2, were extracted from the corpus. Frames of duration 20 ms were extracted with a frame shift of 10 ms. The raw speech frames were preemphasized with the filter $H(z) = 1 - 0.95z^{-1}$ and Hamming windowed. Following this preprocessing, 12-dimensional MFCC and MODGDF features were computed for each frame. Standard delta coefficients were

also computed. These features serve as both baseline features and high dimensional inputs for the PCA and Isomap methods.

3.4. Dimensionality reduction

For each of the classification experiments, 250 units representing each of the required phones were chosen at random from those extracted above to make up the data set. PCA and Isomap were then applied to the equivalent set of MFCC and MODGDF feature vectors. The number of nearest neighbours, k, used in Isomap was set equal to 16. This value was chosen empirically by varying k and examining classification performance.

In order to examine the ability of the feature transformation methods to compute concise representations of the input vectors retaining discriminating information, the dimensionality of the resulting feature vectors was varied from 1 to 12. A separate classifier was subsequently trained and tested using feature vectors with each of the 12 different dimensionalities. Thus the ability of these feature transformation methods to produce meaningful low dimensional features could be evaluated and changes in performance with varying dimension analysed. As a baseline the original MFCC and MODGDF feature vectors were used, also varying in dimensionality from 1 to 12.

3.5. Support vector machine classification

Support vector machine (SVM) [14] classifiers were used in these experiments. SVMs are binary pattern classification algorithms. For our experiments it is necessary to construct a multiclass classifier. This was achieved using a one-against-one training scheme, training one classifier for every possible pair of classes. The final classification result was determined by majority voting.

It is also necessary to choose an appropriate kernel function to be used in the SVMs. In order to select an effective kernel, different SVM models using linear, polynomial, and radial basis function (RBF) kernels were evaluated in a number of phone classification tasks. SVMs with a RBF kernel demonstrated the best classification accuracy and were used for the classification tasks detailed in this work.

In all classification experiments 80% of the data was assigned as training data with the remaining 20% withheld and used as unseen testing data. The data was partitioned such that the training and test sets had no speakers in common, thus ensuring speaker independence.

4. Results

4.1. Baseline classification

Results of each classification experiment using full dimensional MFCC and MODGDF feature sets are shown in Table 1. MFCCs were found to outperform MODGDFs in each test, both with and without the inclusion of delta, Δ, coefficients. Previously published results comparing speech recognition performance using MFCC and MODGDF features are inconsistent with some studies showing better accuracy with MFCC [12] while other studies indicate the opposite [4]. This may be due to various inconsistencies in the corpora, feature extraction procedures, and classification algorithms used.

When the MODGDFs were concatenated with the MFCCs a small improvement over the baseline MFCC result was found. A further improvement was observed when delta coefficients were also included. This is consistent with previously published results [5, 12].

Feature Set	Dim.	Classification Task	
		Phone Class	Vowel
MFCC	12	74.429	51
MODGDF	12	74.000	44
MFCC+Δ	24	75.714	57.4
MODGDF+Δ	24	75.286	52.4
MODGDF+MFCC	24	79.143	51.4
[MODGDF+Δ] + [MFCC+Δ]	48	79.714	61.8

Table 1: *Vowel and phone class classification accuracy (%) using baseline MFCC and MODGDF features. Feature dimensionality (Dim.) is also shown.*

4.2. Reduced dimensionality

In this section we discuss the application of dimensionality reduction methods to magnitude and phase spectrum derived features in an attempt to determine if these representations have underlying low dimensional structure. PCA and Isomap were applied to both MFCCs and MODGDFs and the resulting features evaluated in phone class and vowel classification experiments. In each experiment the classifier was trained and tested on MFCCs, MODGDFs, and features resulting from dimensionality reduction of these baseline features using PCA and Isomap. The dimensionality of the feature vectors used in the experiment vary from 1 to the original dimensionality; this is 12 for static features.

Fig. 2 shows the results of the vowel classification experiments; results of the phone classification experiments were consistent with these results and hence are not discussed in detail. Fig. 2(a) illustrates results using MFCCs and Fig. 2(b) shows results for MODGDFs. The percentage of phones correctly classified is given on the vertical axis. The horizontal axis represents feature vector dimensionality.

MFCCs outperform MODGDFs for all feature dimensionalities. MFCCs also reach a performance plateau more rapidly than MODGDFs indicating that discriminatory information is more compactly represented in MFCCs. This is supported by the fact that the dimensionality reduction methods offer greater relative improvement in accuracy for MODGDFs compared to MFCCs.

In low dimensional feature spaces, with dimensionality $d \leq 5$, Isomap offers the best classification accuracy. This suggests the presence of nonlinear structure in the data which the linear PCA algorithm is incapable of finding but the manifold learning algorithm is able to exploit. The improved performance in low dimensions as a result of dimensionality reduction suggests that both magnitude and phase information may have an intrinsic low dimensional structure.

4.3. Feature combination

Results in Section 4.1 show that classification performance can be improved by including information from both the magnitude and phase spectra. However this was achieved through simple feature vector concatenation that increases the dimensionality and hence the computational cost of any subsequent processing.

In order to reduce this dimensionality PCA and Isomap were applied to the joint features. Isomap was found to yield the best performance for very low dimensional features but does not offer performance comparable to the full dimensional joint features. In contrast, PCA was found to offer minor performance

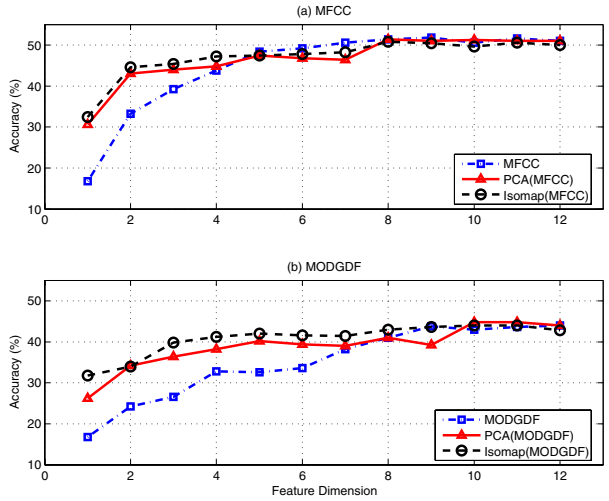

Figure 2: *Vowel classification accuracy using (a) MFCC and (b) MODGDF features. The performance of each feature after dimensionality reduction by PCA and Isomap is also shown.*

increases compared to the original joint features using significantly lower dimensional features. The improvements resulting from PCA are detailed in Tables 2 and 3. Feature dimension (Dim.), classification accuracy (Acc.), and increase in accuracy over the corresponding baseline feature (Inc.) are shown. Applying PCA to joint features with delta coefficients yields less of an increase in accuracy compared with static features alone.

Feature set	Dim.	Acc. (Inc.)
PCA(MODGDF+MFCC)	19	79.857 (0.714)
PCA([MODGDF+Δ] + [MFCC+Δ])	30	80.286 (0.572)

Table 2: *Phone class classification accuracy (%) using joint MFCC and MODGDF features.*

Feature set	Dim.	Acc. (Inc.)
PCA(MODGDF+MFCC)	20	52.2 (0.8)
PCA([MODGDF+Δ] + [MFCC+Δ])	30	61.8 (0)

Table 3: *Vowel classification accuracy (%) using joint MFCC and MODGDF features.*

5. Conclusions

The ability of dimensionality reduction methods to exploit low dimensional structure in both magnitude and phase spectra derived features is demonstrated in this paper. Isomap was found to offer the best phone classification performance in low dimensional space suggesting that this structure may be nonlinearly embedded in higher dimensional space.

We also evaluated magnitude and phase based features in phone classification experiments and found that MFCCs provided better performance than MODGDFs. Joining these features using simple concatenation was found to yield performance increases. Applying PCA to these joint MFCC and MODGDF feature vectors provided increased performance without requiring a large increase in dimensionality and the associated increased computational cost.

6. Acknowledgements

Andrew Errity is supported by the Irish Research Council for Science, Engineering and Technology; grant number RS/2003/114.

7. References

[1] K. Paliwal and L. Alsteris, "On the usefulness of STFT phase spectrum in human listening tests," *Speech Communication*, vol. 45, pp. 153–170, 2005.

[2] B. Yegnanarayana and H. A. Murthy, "Significance of group delay functions in spectrum estimation," *IEEE Transactions on Signal Processing*, vol. 40, no. 9, pp. 2281–2289, Sep. 1992.

[3] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1975.

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

[5] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Significance of the modified group delay feature in speech recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 190–202, January 2007.

[6] R. Togneri, M. Alder, and J. Attikiouzel, "Dimension and structure of the speech space," *IEE Proceedings-I*, vol. 139, no. 2, pp. 123–127, 1992.

[7] V. Jain and L. K. Saul, "Exploratory analysis and visualization of speech and music by locally linear embedding," in *Proc. ICASSP*, vol. 3, 2004, pp. 984–987.

[8] A. Jansen and P. Niyogi, "A geometric perspective on speech sounds," University of Chicago, Tech. Rep., 2005.

[9] A. Errity and J. McKenna, "An investigation of manifold learning for speech analysis," in *Proc. ICSLP*, Pittsburgh PA, USA, September 2006, pp. 2506–2509.

[10] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, December 2000.

[11] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, December 2000.

[12] L. D. Alsteris and K. K. Paliwal, "Evaluation of the modified group delay feature for isolated word recognition," in *Proc. of the Eighth International Symposium on Signal Processing and Its Applications (ISSPA)*, vol. 2, August 2005, pp. 715–718.

[13] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Significance of joint features derived from the modified group delay function in speech processing," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2007, pp. Article ID 79 032, 13 pages, 2007.

[14] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer, N.Y., 1995.