Cluster and Intrinsic Dimensionality Analysis of The Modified Group Delay Feature for Speaker Classification

Rajesh M. Hegde and Hema A. Murthy

Department of Computer Science and Engineering
Indian Institute of Technology, Madras, Chennai.
rajesh,hema@lantana.tenet.res.in

Abstract. Speakers are generally identified by using features derived from the Fourier transform magnitude. The Modified group delay feature(MODGDF) derived from the Fourier transform phase has been used effectively for speaker recognition in our previous efforts.Although the efficacy of the MODGDF as an alternative to the MFCC is yet to be established, it has been shown in our earlier work that composite features derived from the MFCC and MODGDF perform extremely well. In this paper we investigate the cluster structures of speakers derived using the MODGDF in the lower dimensional feature space. Three non linear dimensionality reduction techniques The Sammon mapping, ISOMAP and LLE are used to visualize speaker clusters in the lower dimensional feature space. We identify the intrinsic dimensionality of both the MODGDF and MFCC using the Elbow technique. We also present the results of speaker identification experiments performed using MODGDF, MFCC and composite features derived from the MODGDF and MFCC.

1 Introduction

The most relevant engineering approach to the problem of speaker identification is to represent a speaker by the space which he or she occupies. Indeed there exists a multi-dimensional parameter space in which different speakers occupy different regions. Speakers tend to cluster in this space as points or trajectories at different locations and can occupy more than one region in the entire parameter space. Parameters or features must be chosen such that the clusters are small and well separated. The multidimensional feature space in which speakers position themselves makes pattern recognition difficult, as each observation is made up of a large number of features. Further distances cannot be measured reliably as the covariances of features is difficult to establish. This leads us to investigate effective dimensionality reduction techniques that preserve linear and non linear cluster structures. The issues of cluster analysis and identification of intrinsic dimensionality of the feature set used are crucial. Features like the MFCC which are derived from the Fourier transform magnitude only, by ignoring the phase spectrum, may not be capturing the entire information contained in the signal acquired from each speaker. In this context features derived from phase like the MODGDF [1–3] and composite features derived by combining the MODGDF and MFCC are very relevant. We briefly discuss the MODGDF and use both MODGDF and the traditional MFCC to parametrically represent speakers in

this paper. Further cluster structures of speakers in the lower dimensional space derived using non linear dimensionality reduction techniques like Sammon mapping [4] and unsupervised learning algorithms based on manifold learning like Isometric mapping(ISOMAP) [5] and the Locally linear embedding(LLE) [6], have been investigated in this work. Intrinsic dimensionality analysis is carried out using ISOMAP. The Intrinsic dimensionality is identified using the *Elbow* technique from the residual variance curve and its implications in the context of speaker identification are discussed. Finally the classification results using the MODGDF, MFCC and composite features using a GMM based baseline system are listed.

2 The Modified Group Delay Feature

The group delay function [3], defined as the negative derivative of phase, can be effectively used to extract various system parameters when the signal under consideration is a minimum phase signal. The group delay function is defined as

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

where $\theta(\omega)$ is the unwrapped phase function. The group delay function can also be computed from the speech signal as in [3] using

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

where the subscripts R and I denote the real and imaginary parts of the Fourier transform. $X(\omega)$ and $Y(\omega)$ are the Fourier transforms of $x(n)$ and $nx(n)$, respectively. The group delay function requires that the signal be minimum phase or that the poles of the transfer function be well within the unit circle for it to be well behaved. This has been clearly illustrated in [2] and [3]. It is also important to note that the denominator term $|X(\omega)|^2$ in equation 2 becomes zero, at zeros that are located close to the unit circle. The spiky nature of the group delay spectrum can be overcome by replacing the term $|X(\omega)|^2$ in the denominator of the group delay function with its cepstrally smoothed version, $S(\omega)^2$. Further it has been established in [1] that peaks at the formant locations are very spiky in nature. To reduce these spikes two new parameters γ and α are introduced. The new modified group delay function as in [3] is defined as

$$\tau_m(\omega) = \left(\frac{\tau(\omega)}{|\tau(\omega)|}\right) \left(|\tau(\omega)|\right)^{\alpha} \tag{3}$$

where

$$\tau(\omega) = \left(\frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{S(\omega)^{2\gamma}}\right) \tag{4}$$

where $S(\omega)$ is the smoothed version of $|X(\omega)|$. The new parameters α and γ introduced vary from 0 to 1 where ($0< \alpha \leq 1.0$) and ($0< \gamma \leq 1.0$). The algorithm for computation of the modified group delay function is explicitly dealt with in [3]. To convert the modified group delay function to some meaningful parameters, the group delay function is converted to cepstra using the Discrete Cosine Transform (DCT).

$$c(n) = \sum_{k=0}^{k=N_f} \tau_x(k) \cos(n(2k+1)\pi/N_f) \tag{5}$$

where N_f is the DFT order and $\tau_x(k)$ is the group delay function. The second form of the DCT, DCT-II is used, which has asymptotic properties to that of the *Karhunen Loeve Transformation* (KLT) as in [3]. The DCT acts as a linear decorrelator, which allows the use of diagonal co-variances in modeling the speaker vector distribution.

3 Speaker Cluster Analysis with Sammon mapping

Speaker classification researchers are usually confronted with the problem of working with huge databases and a large set of multidimensional feature vectors, which exerts a considerable load on the computational requirements. Typically Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are used for dimensionality reduction in the speech context, despite the fact that they may not be optimum for class discrimination problems. We therefore use the Sammon mapping technique[4] for dimensionality reduction of the MODGDF and MFCC as it preserves the inherent structure of the underlying distribution. In order to visualize the cluster structure of individual speakers, we first compute 16 dimensional vector quantization (VQ) codebooks of size 64 by concatenating six sentences of that particular speaker picked from the training set of the NTIMIT [7] database. Each codebook is transformed into a two dimensional codebook of size 64 using Sammon mapping [4]. Sammon mapping, which belongs the class of Multidimensional scaling techniques (MDS) minimizes the following error function to extract lower dimensional information from high dimensional data using gradient descent technique:

$$\varepsilon_{sam} = \frac{1}{\sum_{i=1}^{i=N-1} \sum_{j=i+1}^{i=N} D_{ij}} \sum_{i=1}^{i=N-1} \sum_{j=i+1}^{i=N} \frac{(d_{ij} - D_{ij})^2}{D_{ij}} \tag{6}$$

where d_{ij} is the distance between two points i,j in the d-dimensional output space, and D_{ij} is the distance between two points i , j in the D-dimensional input space, N is the number of points in the input or output space. The

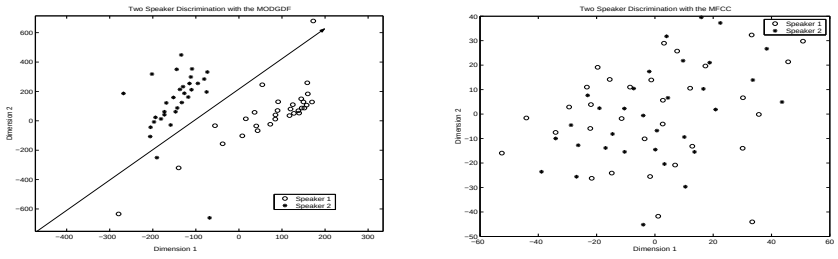

Fig. 1. Two speaker discrimination with the MODGDF (left) and MFCC (right)

results of cluster analysis for two speakers in the two dimensional space is shown in *figure 1(a) and 1(b)*. It is evident that MODGDF clearly separates the two speakers in the low dimensional feature space compared to MFCC.

4 Intrinsic Dimensionality Analysis using Unsupervised learning algorithms

From the results of sammon mapping, one can be tempted to hypothesize that the projections of MFCC resulted in much greater error than those of the MODGDF. We therefore identify the intrinsic dimensionality of the MODGDF and MFCC using unsupervised learning algorithms like the ISOMAP and LLE and then visualize speakers in the lower dimensional space. Although both ISOMAP and LLE can also be used for identifying the intrinsic dimensionality of any feature set by detecting the dimension at which the error bottoms down, LLE may fail for feature sets twisted and folded in the high dimensional input space. But ISOMAP is guaranteed to asymptotically converge and recover the true dimensionality of even such feature sets. Hence we use ISOMAP and the Elbow technique to identify the true dimensionality of the feature set in this work.

4.1 Isometric mapping (ISOMAP) and The *Elbow* technique [5]

The ISOMAP has three steps, The first step determines which points are neighbors on the manifold M, based on the distances $d_x(i,j)$ between pairs of points i, j in the input space X. Two simple methods are to connect each point to all points within some fixed radius e, or to all of its K nearest neighbors. These neighborhood relations are represented as a weighted graph G over the data points, with edges of weight $d_x(i,j)$ between neighboring points. In its second step, Isomap estimates the geodesic distances $d_m(i,j)$ between all pairs of points on the manifold M by computing their shortest path distances $d_g(i,j)$ in the graph G using an appropriate shortest path finding technique like the Dijkstra algorithm. The final step applies classical MDS to the matrix of graph distances $D_G = d_G(i,j)$, constructing an embedding of the data in a d-dimensional euclidean space Y that best preserves the manifolds estimated intrinsic geometry. Further the intrinsic dimensionality of the feature set can be estimated by looking for the *Elbow* at which the curve showing the relationship between residual variance and the number of dimensions of the feature set ceases to decrease significantly which is called the *Elbow* technique. It is important to note that residual variance is the amount of variance in the feature set remaining after the first n principal components have been accounted for. The residual variance curves for two speakers using MODGDF and MFCC are illustrated in *figures 2(a) and 2(b)* respectively. It is interesting to note that MODGDF has a intrinsic dimensionality (2 and 3) while MFCC exhibits an intrinsic dimensionality of 4 with respect to this pair of speakers. The 3 dimensional visualization of codebooks of two speakers with the MODGDF and MFCC using ISOMAP are illustrated in *figures 3(a) and 3(b).*

4.2 Locally linear embedding (LLE) [6]

The LLE is an unsupervised learning algorithm that computes low-dimensional, neighborhood-preserving embeddings of high-dimensional inputs. LLE maps its inputs into a single global coordinate system of lower dimensionality, and its optimizations do not involve local minima. The LLE algorithm, for mapping high dimensional data points, X_i , to low dimensional embedding vectors, Y_i can be summarized in three steps. The first step computes the neighbors of

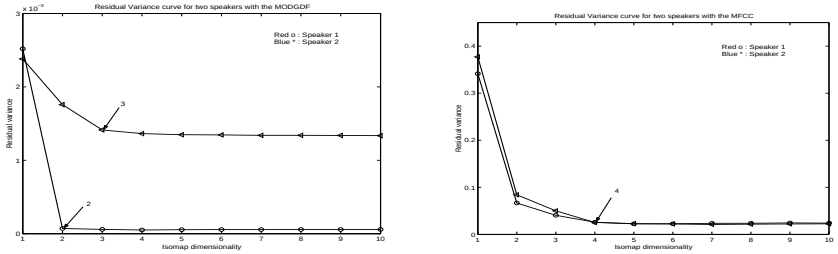

Fig. 2. Residual variance for two Speakers with the MODGDF (left) and MFCC (right) using ISOMAP

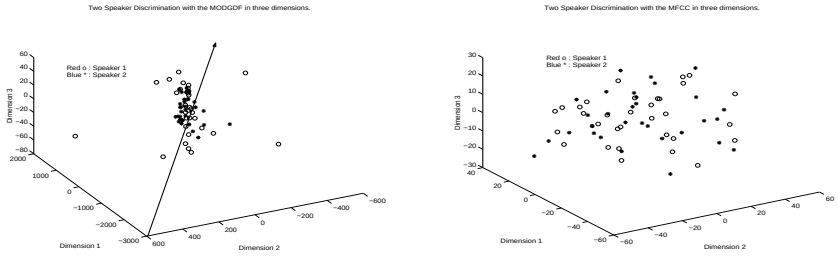

Fig. 3. Two Speaker cluster structure in 3 dimensions with the MODGDF (left) and MFCC (right) using ISOMAP

each data point, X_i. In the next step the weights W_{ij} are computed that best reconstruct each data point X_i from its neighbors, minimizing the cost in

$$\varepsilon(W) = \sum_i |X_i - \sum_j W_{ij} X_j|^2 \tag{7}$$

by constrained linear fits. The final step computes the vectors best reconstructed by the weights , minimizing the quadratic form

$$\Phi(W) = \sum_i |Y_i - \sum_j W_{ij} Y_j|^2 \tag{8}$$

by its bottom nonzero eigenvectors. The 3 dimensional visualization of two speakers with the MODGDF and MFCC using LLE are illustrated in *figures 4(a) and 4(b).*

5 Classification Results and Conclusions

The MODGDF gave a recognition percentage of 98.5% and 96.5% while MFCC gave 98% and 97% for 100 and 200 speakers on the TIMIT database, using a GMM based baseline system. Composite features derived from MFCC and MODGDF performed at 50% for NTIMIT data. We also noticed that the intrinsic dimensionality was around 2 and 3 for MODGDF for 90% of speakers from the NTIMIT [7] database, while MFCC intrinsic dimensionality was equal to or higher than 4. But it can be concluded from the clustering and intrinsic dimensionality analysis that MODGDF is capable of discriminating speakers in a lower

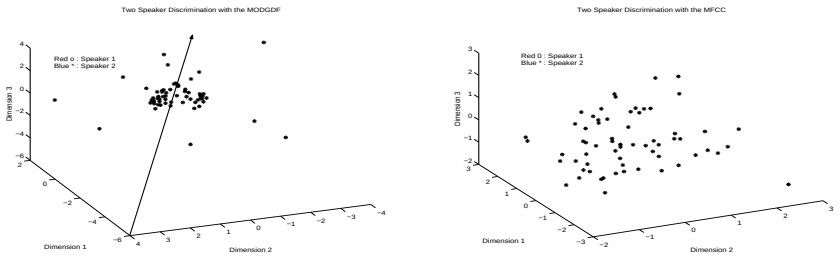

Fig. 4. Two Speaker cluster structure in 3 dimensions with the MODGDF (left) and MFCC (right) using LLE

dimensional space while MFCC requires a higher dimensional representation. We investigated the intrinsic dimensionality of a large number of speakers from the NTIMIT database and noticed from the cluster plots that speaker clusters are well separated only at the intrinsic dimensionality of their parametric representations. We therefore intend to identify the intrinsic dimensionality of speakers first and then use this crucial information for automatic speaker identification tasks in our future efforts. This can reduce the computational overhead and also lead us to various other possibilities is speech recognition tasks.

References

1. Rajesh M.Hegde, Hema A.Murthy and Venkata Ramana Rao Gadde : Application of the Modified Group Delay Function to Speaker Identification and Discrimination. Proceedings of the ICASSP 2004, May 2004, Vol 1, pp. 517-520
2. Rajesh M.Hegde and Hema A.Murthy : Speaker Identification using the modified group delay feature. Proceedings of The International Conference on Natural Language Processing-ICON 2003,December 2003, pp. 159-167
3. Hema A. Murthy and Venkata Ramana Rao Gadde : The Modified group delay function and its application to phoneme recognition. Proceedings of the ICASSP, April 2003, Vol.I, pp. 68-71
4. Sammon, Jr., J. W. : A Nonlinear Mapping for Data Structure Analysis. IEEE Transactions on Computers **C-18(5)** (1969) 401-409
5. Joshua B. Tenenbaum, Vin de Silva, and John C. Langford : A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science *www.science.org* **290(5500)** (2000) 2319-2323
6. Sam T. Roweis and Lawrence K. Saul : Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science *www.science.org* **290(5500)** (2000) 2323-2326
7. Charles Jankowski, Ashok Kalyanswamy, Sara Basson, and Judith Spitz : NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database. Proceedings of ICASSP-90, April 1990.