APPLICATION OF THE MODIFIED GROUP DELAY FUNCTION TO SPEAKER IDENTIFICATION AND DISCRIMINATION

Rajesh M. Hegde, Hema A. Murthy

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

Gadde V. Ramana Rao

STAR Laboratory,SRI International,
333,Ravenswood Avenue, Menlo Park,
CA 94025
{rao}@speech.sri.com

ABSTRACT

In this paper, we explore new methods by which speakers can be identified and discriminated, using features derived from the fourier transform phase. The Modified Group Delay Feature(MODGDF) which is a parameterized form of the modified group delay function is used as a front end feature in this study. A Gaussian mixture model(GMM) based speaker identification system is built with the MODGDF as the front end feature. The system is tested on both clean (TIMIT) and noisy telephone(NTIMIT) speech. The results obtained are compared with traditional Mel frequency cepstral coefficients(MFCC) which is derived from the fourier transform magnitude. When both MFCC and MODGDF were combined, the performance improved by about 4% indicating that both phase and magnitude contain complementary information. In an earlier paper [1], it was shown that the MODGDF does possess phoneme specific characteristics. In this paper we show that the MODGDF has speaker specific properties. We also make an attempt to understand speaker discriminating characteristics of the MODGDF using the nonlinear mapping technique based on Sammon mapping [10] and find that the MODGDF empirically demonstrates a certain level of linear separability among speakers.

1. INTRODUCTION

Current state of the art speaker identification systems use features derived from the Fourier transform magnitude like MFCC, its derivatives and also PLP Cepstra. Though half of the underlying spectral information is discarded in these cases, attempts to utilize the phase spectrum for deriving features have been minimal. The modified group delay function [1], which is a variant of the group delay function has been used in [1] to build a phoneme recognizer, and, in [2] for speaker identification. In this paper, we build a GMM based speaker identification system using the MODGDF as the front end feature and primarily study its speaker discriminating properties. The MODGDF is tested on both clean speech (TIMIT) and telephone speech (NTIMIT) using a maximum likelihood classification scheme (GMM). The performance of the system based on MODGDF is compared with that of the traditional MFCC. Since a sixteen dimensional MODGDF is used for recognition, dimensionality reduction is performed to aid in visual perception of the feature in two dimensions. The Classical Sammon mapping technique has been used to reduce a typical codebook of sixteen dimensions to two dimensions. On visualization, the codebooks of speakers derived from the MODGDF exhibit a certain level of linear separability in the reduced feature space.

2. THEORY OF THE MODIFIED GROUP DELAY FUNCTION

Speakers can be characterized by either magnitude or phase information alone [6]. But it is widely perceived that the magnitude spectrum visually represents the system information very well when compared to that of the phase spectrum. It is important to note that unlike the phase spectrum, the group delay function [6], 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. This is primarily because the magnitude spectrum of a minimum phase signal [6], and its group delay function resemble each other. 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 values of the group delay function that deviate from a constant value indicates the degree of non linearity of the phase. The group delay function can also be computed from the speech signal

as in [1] using

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

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. The group delay function becomes spiky in nature due to pitch peaks, noise and window effects. This has been clearly illustrated in [1] and [2]. 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 next task is therefore to suppress the zeros. 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 [1] is defined as

$$\tau_m(\omega) = (\frac{\tau(\omega)}{|\tau(\omega)|})\ (|\tau(\omega)|)^{\alpha} \quad (3)$$

where

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

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 ($0< \gamma \leq 1.0$). It has been emphasized in [1] that in the group delay domain the channel effect can be subtracted out assuming that it is a function of frequency only. The algorithm for computation of the modified group delay function is explicitly dealt with in [1].

2.1. Feature Extraction using the modified group delay function

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).

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

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 [1]. The DCT acts as a linear de-correlator, which allows the use of diagonal covariances in modelling the speaker vector distribution. c(n) shall be referred to as the modified group delay feature (MODGDF) in the forthcoming sections.

3. PERFORMANCE EVALUATION

3.1. Databases

The databases used in this study are the TIMIT [7] for clean speech and the NTIMIT [8] for noisy telephone speech.

3.2. The Baseline System

A series of GMMs modelling the voices of speakers for whom training data is available and a classifier, that evaluates the likelihoods of the unknown speakers voice data against these models make up the likelihood maximization based baseline system used in this study. We tested individual features derived from magnitude and phase and a combination of these features at measurement level. These results are presented in the following sections.

3.3. Experimental Results

The results of the MFCC and the MODGDF on both the TIMIT and NTIMIT corpora using the GMM scheme are listed in *Table 1*. For 400 tests MFCC and MODGDF gave a 97% and 96.5% recognition for clean speech(TIMIT) respectively, but degraded to 40% and 41% respectively for telephone speech(NTIMIT). The recognition performance for the composite feature derived by combining MODGDF with MFCC is listed in *Table 2*. The best net recognition is found to be 44% when MODGDF is combined with MFCC yielding a 3% improvement in performance.

Table 1. R*ecognition performance of MFCC and MODGDF for the TIMIT and NTIMIT database*

Feature	Database	Recognition %	
		200 tests	400 tests
MFCC	TIMIT	98	97
MODGDF	TIMIT	98.5	96.5
MFCC	NTIMIT	41	40
MODGDF	NTIMIT	44	41

Table 2. *Recognition performance of composite features on the NTIMIT database*

Feature Name	Recognition %	
	200 tests	400 tests
MFCC+MODGDF	48	44

4. SPEAKER DISCRIMINATION USING THE MODIFIED GROUP DELAY FEATURE

The goal of feature selection [4] is to find a transformation to a relatively low dimensional feature space that preserves

the information pertinent to the speaker identification problem and to enable acceptable comparisons to be performed. The simplest way to improve the recognition performance of a speaker identification system at feature level is to increase the dimensionality of the extracted feature. But this exerts a huge load on the computational and storage requirements. Hence the selection of an appropriate transformation scheme for dimensionality reduction is a requirement for any feature. Principal Component Analysis (PCA), Factor Analysis (FA) and Linear Discriminant Analysis (LDA) are techniques that may not be neccesarily optimum for class discrimination problems. Keeping in mind that the Automatic speaker identification (ASI) is more of a discrimination problem than a representation problem we reduce the multidimensional modified group delay feature vectors to two dimensional feature vectors using the traditional and widely accepted Sammon mapping technique [10]. Sammon's mapping is an iterative method based on a gradient search [10]. The intent is to map features in n-dimensional space to two dimensions. The algorithm finds the locations in the target space so that the original structure of the measurement vectors in the n-dimensional space is conserved to the maximum extent possible. Classification accuracy based on Sammon's projections is comparable with, and in some cases even superior to that based on other feature extractors [11]. We hence made an effort to visualize two dimensional codebooks for various categories of speakers using Sammon's mapping. Each speaker's codebook of size thirty two is generated by concatenating six sentences of that particular speaker picked from the training set of the NTIMIT database. The codebook which consists of thirty two, sixteen dimensional code vectors is transformed into a two dimensional codebook of size thirty two after sammon mapping [10]. *Figures 1(a) and 1(b)* show the distribution of the code vectors for two female speakers using the MFCC and the MODGDF respectively. We observe that in *Figure 1(a)* no structure for each speaker is visible, while in *Figure 1(b)* the code vectors corresponding to each of the speakers can be separated by a straight line. Similar results are demonstrated for a set of three and four speakers in *Figures 2,3(a) and 3(b)*, respectively. Although in *Figure 3(b)* the code vectors of different speakers begin to overlap, clearly in comparison with *Figure 3(a)*, each speaker's code vectors are clustered close together.

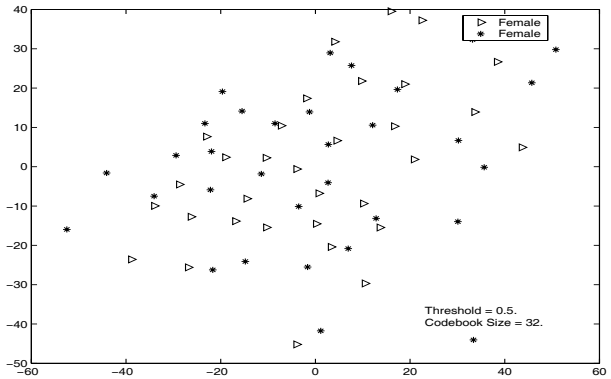

Figure 1(a): *Female-Female Speaker Discrimination with the MFCC Feature*

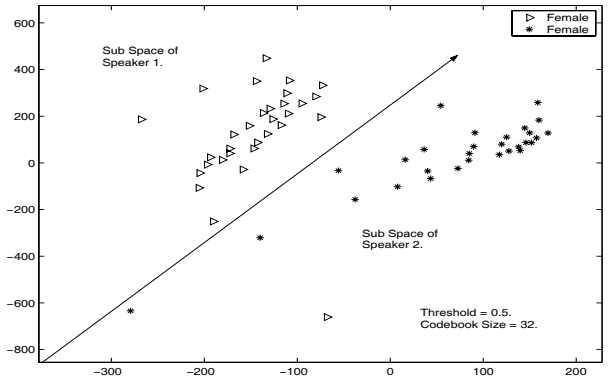

Figure 1(b): *Female-Female Speaker Discrimination with the MODGDF*

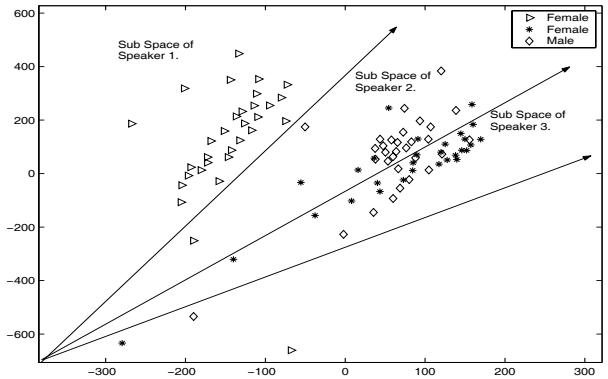

Figure 2: *Three Speaker (2 female - 1 male) Speaker Discrimination with the MODGDF*

5. DISCUSSION

In the experiments we conducted using the baseline system we noted that the performance of the Magnitude based feature MFCC came down drastically from 99% to around 31% respectively for the TIMIT and NTIMIT data. Similar results of recognition performance ranging between 16% to 18% have been reported by researchers on NTIMIT data [5], except in [9] where a 60% recognition performance has been claimed. With respect to experiments conducted on the NTIMIT corpora, the major aspect we noted was that the MODGDF alone gave a better or at least equal performance when compared with MFCC. But a combination of the MODGDF with MFCC gave an overall improvement in performance of 3-4%. It is quite possible that different features capture different speaker characteristics. It is significant to note that the MODGDF is capable of linearly

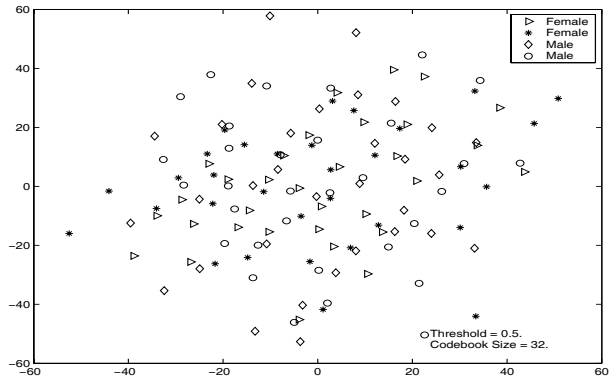

Figure 3(a): *Four Speaker (2 male - 2 female) Discrimination with the MFCC Feature*

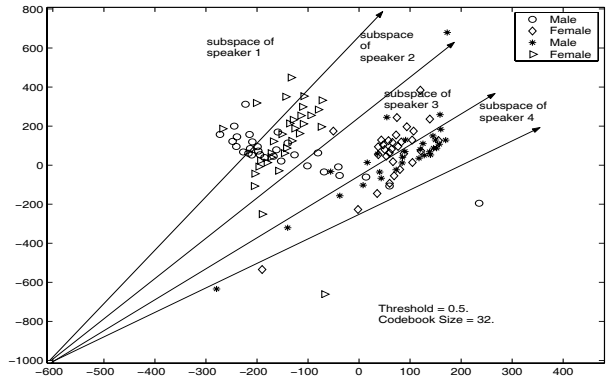

Figure 3(b): *Four Speaker(2 male - 2 female) Speaker Discrimination with the MODGDF*

separating speakers in the lower dimensional speaker space when a non linear mapping technique like the sammon mapping is employed. We are currently exploring new methods of identifying speakers in the lower dimensional space itself. We also intend to transform the MODGDF to a higher dimensional space, where the speaker's features become linearly separable.

6. CONCLUSION

The idea of using features derived from Fourier transform phase like the MODGDF for the task of ASI is implemented in this paper. The new feature is found to perform better than or at least equal to traditional cepstral features like MFCC. Combining evidences derived from MODGDF and traditional features also lead to a 3-4% overall improvement in recognition performance. When the MODGDF is transformed using a nonlinear mapping technique like the sammon mapping [10] the speaker clusters are almost linearly separable. Perhaps a classification scheme like the voting method based on nearest distance may be more appropriate in this context. A sound mathematical insight into this new feature can lead to an optimization where it could be used for a variety of speaker and speech recognition tasks.

7. REFERENCES

[1] Hema A. Murthy and Venkata Ramana Rao Gadde, "The Modified group delay function and its application to phoneme recognition,"*Proceedings of the ICASSP*, Vol.I, pp. 68-71, April 2003.

[2] Rajesh M.Hegde and Hema A.Murthy,"Speaker Identification using the modified group delay feature,"*International Conference on Natural Language Processing-ICON 2003*,Accepted for presentation, December 2003.

[3] Hema A. Murthy, Francoise Beaufays, and Larry P. Heck, "Robust Text-Independent Speaker Identification over Telephone Channels," *IEEE Trans. on Speech and Audio Processing*, Vol.7, No.5, pp554-568 September 1999.

[4] H.Gish and M.Schmidt, "Text Independent Speaker Identification,"*IEEE Signal Processing Mag.*,pp 18-32, October 1994.

[5] L.Besacier and J.F.Bonastre," Time and Frequency Pruning for Speaker Identification," *Proc. ICPR-98*, 1998.

[6] Hema A. Murthy and B.Yegnanarayana, "Formant Extraction from Group Delay function," *Speech Communication*, Vol.10, pp.209-221, 1991.

[7] "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus" (CD-ROM), NIST Speech Disc 1-1.1, NTIS- 1990.

[8] 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.

[9] D.A. Reynolds,"Large Population Speaker Identification using clean and telephone speech,"*IEEE Signal Processing Letters*,Vol.2, pp46-48, March, 1995.

[10] Sammon, Jr., J. W. "A Nonlinear Mapping for Data Structure Analysis,"*IEEE Transactions on Computers*, Vol.C-18, No.5, pp. 401-409, May 1969.

[11] Lerner B, Guterman H, Aladjem M, Dinstein I, Romem Y," On Pattern Classification with Sammon's Nonlinear Mapping- An Experimental Study", *Pattern Recognition*, Vol. 31, pp. 371- 381.