Automatic language Identification and discrimination using the modified group delay feature

# Rajesh M. Hegde and Hema A. Murthy

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

Abstract

Automatic language identification (LID) 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. This paper investigates the use of features derived from the Fourier transform phase for implementing LID systems. Features derived from the modified group delay function which we call the modified group delay feature (MODGDF) are used in this study. Performance of the MODGDF and the traditional MFCC for a GMM based LID system for a 3 and 11 language task are discussed. Results of language discriminability analysis are also presented. The MODGDF is found to outperform MFCC in terms of both performance and discriminability of languages.

Keywords

Feature extraction, Group delay functions, Language identification, Dimensionality reduction

1. INTRODUCTION

Automatic language identification (LID) is a challenging task of training a machine to identify the language of a short excerpt of speech uttered by an unknown speaker. Recently there has been a surge in LID research, thanks to its many real time applications like multilingual man machine interfaces and content based indexing of multimedia or audio data. The most successful approach to LID is to use the phonotactic content of the speech signal to discriminate among languages. Phone based LID systems like the parallel phone recognition and language modeling (PPRLM) [1] give the best LID performance but are computationally expensive making them unsuitable for real time applications. High performance LID systems as in [2], using Gaussian mixture models (GMM) which perform on par with phone based approaches have also been used in the recent past. Phone based approaches require some kind of explicit phoneme recognition to be performed prior to language identification. Although new methods have evolved where explicit phoneme recognition can be avoided and classification of languages can be achieved by parallel phone recognition [3], the performance of such methods depends on the quality of speech and the information in the features extracted from the speech signal. The acoustic content of the speech signal is conventionally extracted as features from the short time Fourier transform (STFT) power spectra of the speech signal or even using model based approaches like linear prediction (LP) modeling which essentially capture the formant structure of the speech sounds. Alternative features for LID often depend on energy, pitch, intonation and prosodic information. If such features are treated as an abstract vector and a huge set of such vectors is collected from a large unconstrained database, features of dissimilar languages will cluster apart in the multidimensional feature space. Hence the kind of information in features that are extracted from the speech signal of a particular language is crucial in classifying and identifying languages. A comprehensive literature survey on feature extraction for LID systems reveals that filter bank features, LPC derived features, formant and prosodic features have been used traditionally for classifying languages. The work on LID by Muthusamy [4] uses cepstral features like MFCC, and RASTA-PLP for improved recognition performance. It is evident that attempts to utilize the phase spectrum for deriving features have been minimal. The modified group delay function has been used in [5] to build a phoneme recognizer. It is a variant of the group delay function [6] which has been used widely for speech processing. The modified group delay function has also been used previously in [7] for speaker identification, and, in [8] for syllable recognition. In all these efforts features are derived from the modified group delay function which we call the modified group delay feature (MODGDF). In this paper, we build a GMM based automatic language identification system using the MODGDF as the front end feature and also study its language discriminating properties. The MODGDF is used to identify the language of both read [9], (three language task) and spontaneous speech [10], (eleven language task) using a maximum likelihood classification scheme (GMM). The performance of the system using the MODGDF is compared with that of the traditional MFCC. Joint features are derived from the MODGDF and MFCC using late fusion and the results of performance evaluation listed. 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 first technique is based on Sammon mapping [11] and the second on an unsupervised manifold learning technique called isometric mapping (ISOMAP) [12]. On visualization, the codebooks of languages derived from the MODGDF exhibit a certain level of linear separability in the reduced feature space. One feature selection technique, the sequential forward search using the Bhattacharya distance metric [13], is implemented to illustrate the efficacy of the MODGDF over the MFCC in discriminating languages. The significance of the results of LID performance and the language discriminating properties of the MODGDF are discussed in the final part of the paper.

2. THE MODIFIED GROUP DELAY FUNCTION

Spoken language 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

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

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 [5] using

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

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 [5]. 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 push the zeros well inside 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 [5] 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 [5] is defined as

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

where

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

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 [5] 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 [5].

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).

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

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 $\mathrm{Karhunen\ Loeve\ Transformation}$ (KLT) as in [5]. The DCT acts as a linear de-correlator, which allows the use of diagonal co-variances in modeling the language 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 Database for Indian languages [9], for read speech and the OGI_MLTS corpus [10], for spontaneous telephone speech.

3.1.1 Database for Indian languages (DBIL)

The Database for Indian languages collected by the Speech and Vision lab at IIT, Madras, consists of Doordarshan news bulletins in eight different languages. Out of these we consider news bulletins of three languages Tamil, Telugu and Hindi as they are segmented and labeled. For training we use two male and two female speakers data of fifteen minutes duration. During testing we use three second utterances distributed over two male and two female speakers. The speakers used for training and testing are different.

3.1.2 The OGI_MLTS corpus

The Oregon Graduate Institute Multi-language Telephone Corpus collected by Muthusamy [10], consists of spontaneous speech utterances of eleven languages uttered over real telephone lines by approximately 90 male and 40 female speakers. The 11 languages in the corpus are English, French, Korean, Mandarin, Farsi, German, Spanish, Hindi, Vietnamese, Tamil and Japanese. 40 male and 5 female speakers uttering a particular language are used for training while 18 male and 2 female speakers are used for testing.

The duration of the test utterance is 45 seconds. The speakers used for training and testing are different.

3.2 The Baseline System

A series of GMMs modeling each language for which training data is available and a classifier, that evaluates the likelihoods of the unknown language data against these models make up the likelihood maximization based baseline system used in this study. We tested individual features, the MODGDF and the MFCC and a combination of these features using late fusion. These results are presented in the following sections.

3.3 Experimental Results

The results of the MFCC and the MODGDF on both the DBIL and OGI_MLTS corpora using the GMM scheme are listed in Table 1. For the 3 language task the MFCC and MODGDF gave a 95% and 96% recognition respectively on the DBIL data, but degraded to 83.3% and 87% respectively on OGI_MLTS data. But for the 11 language task the MFCC and MODGDF yielded a recognition performance of 50% and 53% respectively on OGI_MLTS data. 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 90% when MODGDF is combined with MFCC for the 3 language task and 57% for the 11 language task yielding a maximum of 4% improvement in performance. The confusion matrix for the 11 language task is shown in Table 3. The matrix is constructed from the results of testing 20 utterances (18 male and 2 female) each of duration 45 seconds. From the matrix it significant to note that English and French are easily discriminated from the other languages. Confusability between Japanese and Korean is very high.

Table 1. Recognition performance of MFCC and MODGDF for the DBIL and OGI_MLTS database for 3 LT : 3 language task and 11 LT: 11 language task

Feature	Database	Recognition %	
		3 LT	11 LT
MFCC	DBIL	95	–
MODGDF	DBIL	96	–
MFCC	OGI_MLTS	83.3	50
MODGDF	OGI_MLTS	87	53

Table 2. Recognition performance of joint features on the OGI_MLTS database for 3 LT : 3 language task and 11 LT: 11 language task

Feature Name	Recognition %	
	3 LT	11 LT
MFCC+MODGDF	90	57

Table 3. Confusion Matrix for the 11 language task, E:Englsh, Fr: French, K: Korean, M: Mandarin, Fa: Farsi, G: German, S: Spanish, H: Hindi, V: Vietnamese, T: Tamil and J: Japanese

	E	Fr	K	M	Fa	G	S	H	V	T	J
E	16	0	3	0	0	1	0	0	0	0	0
Fr	0	16	2	0	0	0	0	2	0	0	0
K	2	0	10	0	0	0	0	0	1	0	7
M	3	0	4	7	0	3	0	0	0	0	3
Fa	5	0	0	0	11	3	0	0	1	0	0
G	7	0	2	0	3	8	0	0	0	0	0
S	2	2	0	0	0	0	13	0	3	0	0
H	5	0	0	0	2	1	0	6	3	3	0
V	3	0	1	1	1	2	0	0	8	4	0
T	3	0	0	0	3	0	0	2	1	11	0
J	0	2	9	0	2	0	0	0	2	0	5

4. SEPARABILITY ANALYSIS OF LANGUAGES

From the results in the previous section it is evident that some languages have a high degree of separability while the others do not. In this section we make an effort to analyze separability between languages with respect to how well the MODGDF discriminates between languages.

4.1 Visualization of cluster structures

Since we use acoustic features derived from Fourier transform phase as the front end and a classic pattern recognition technique like the GMM as the back end, the separability between languages is well studied by visualizing the cluster structures of different languages. The ideal way to visualize the cluster structures would be to reduce the higher dimensional cluster structure of each language to a lower dimensional cluster structure. We therefore use new and innovative methods like Sammon mapping [11] and a manifold learning technique like Isometric mapping (ISOMAP) [12] for dimensionality reduction of the cluster structure derived from the MODGDF. These methods are optimum for class discrimination problems as they preserve the inherent structure of the underlying distribution. In order to visualize the cluster structure of individual languages, we first compute vector quantization (VQ) codebooks to represent each language. The feature vector dimension selected is 16 and the size of the codebook is 32. Each language codebook is generated by concatenating 45 sentences of that particular language picked from the training set of the OGI_MLTS database [10]. The codebook which consists of sixty four, sixteen dimensional code vectors is transformed into a two dimensional codebook of size 32 using Sammon mapping and ISOMAP. From the results of performance evaluation it is evident that English and French are well distinguished. Japanese and Korean are highly confused. We therefore visualize the codebooks derived from the MODGDF for the languages English, French, Japanese and Korean. Figures 1(a) and 1(b) show the distribution of the code vectors for the languages English and French using Sammon mapping and ISOMAP respectively. We observe that in Figure 1(a) and Figure 1(b) the code vectors corresponding to each of the languages can be separated by a straight line. Figures 2(a) and 2(b) show the distribution of the code vectors for the languages Japanese and Korean using Sammon mapping and ISOMAP respectively. We observe that in Figure 2(a) and Figure 2(b) the code vectors corresponding to each of the languages have a high degree of overlap.

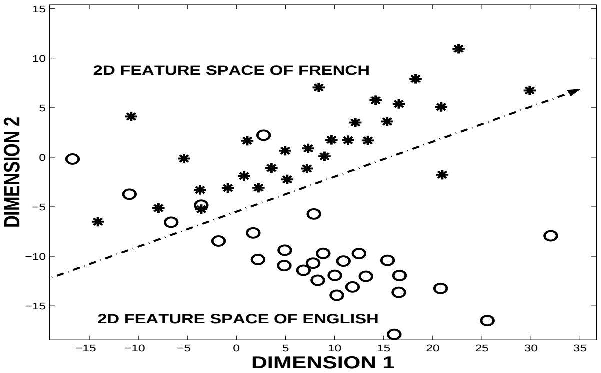

Figure 1(a): English-French language discrimination with the MODGDF using Sammon mapping

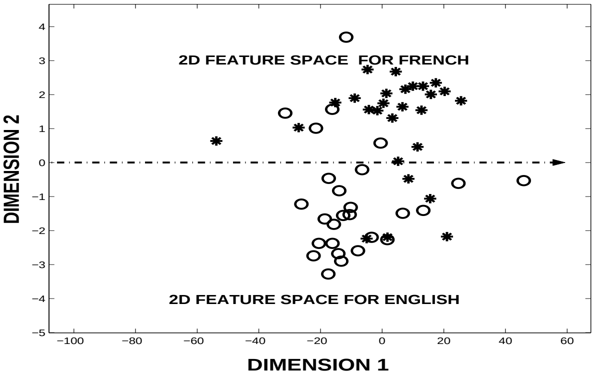

Figure 1(b): English-French language discrimination with the MODGDF using ISOMAP

4.2 Separability analysis via the Bhattacharya distance

Generally class separability can be measured by geometrically intuitive measures like the F ratio and mathematical measures like the Chernoff and the Bhattacharya bound [13]. The Bhattacharya bound which is a special case of the Chernoff bound is a probabilistic error measure and relates more closely to the likelihood maximization classifiers that we use for performance evaluation. We therefore consider all 11 languages from the OGI_MLTS database and compute a 16 dimensional codebook of size 64 for each language. The separability criterion based on the Bhattacharya distance measure is then calculated. The separability criterion and the cumulative separability criterion versus feature dimension for both the MODGDF and the MFCC is illustrated in Figure 3(a) and Figure 3(b) respectively. From Figures 3(a) and 3(b) it is evident that the MODGDF outperforms MFCC with respect to language separability, as the separability curve corresponding to MODGDF is well above that of MFCC.

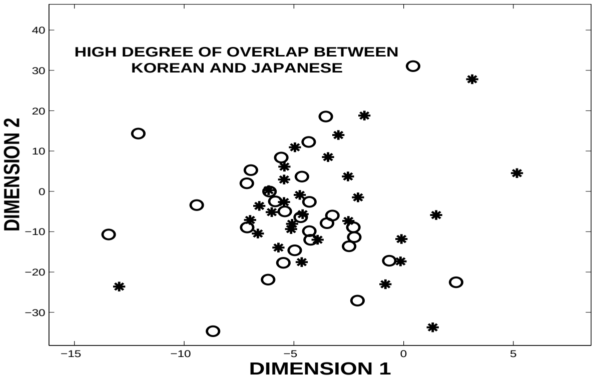

Figure 2(a): Japanese-Korean language discrimination with the MODGDF using Sammon mapping

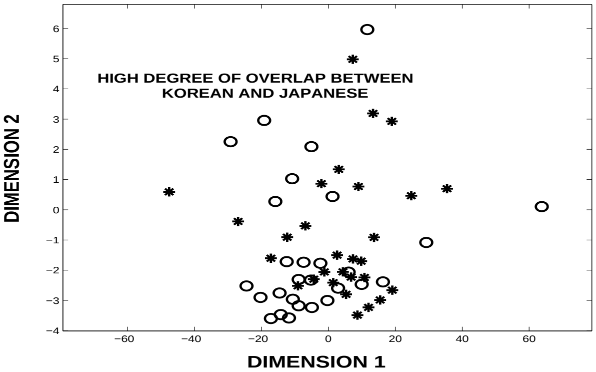

Figure 2(b): Japanese-Korean language discrimination with the MODGDF using ISOMAP

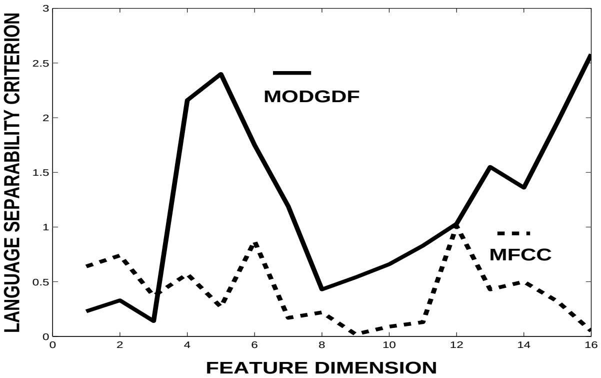

Figure 3(a): Language separability of MODGDF and MFCC feature sets using Bhattacharya distance

5. CONCLUSION

The idea of using features derived from Fourier transform phase like the MODGDF for the task of LID is implemented

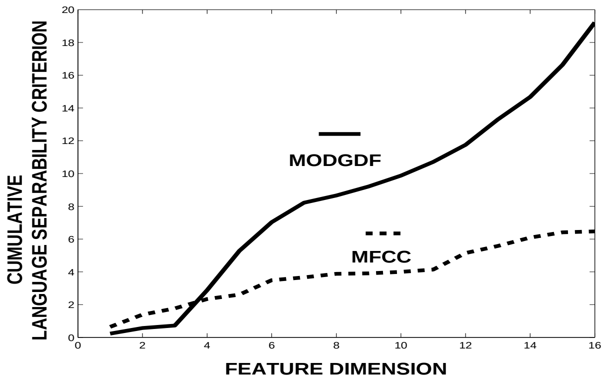

Figure 3(b): Cumulative Language separability of MODGDF and MFCC feature sets using Bhattacharya distance

in this paper. The new feature is found to perform better by 3% than the traditional cepstral feature MFCC. Combining evidences derived from the MODGDF and MFCC also leads to a 3-4% overall improvement in recognition performance. When the MODGDF is transformed using a nonlinear mapping technique like the sammon mapping [11] and ISOMAP, the speaker clusters are almost linearly separable for most languages. Further the MODGDF outperforms the traditional cepstral feature MFCC in terms of class separability, when a distance metric, like the Bhattacharya distance is used. Since the utility of the MODGDF has been proved for speaker [7], phoneme [5], and syllable recognition [8] in our previous efforts, this new effort of applying the MODGDF for language recognition could give credence to the idea of comprehensively proposing the MODGDF as an alternate representation of speech. We also intend to work on deriving a joint feature by combining the MODGDF and the MFCC capable of capturing the complete information in the speech signal, which can be used across all speech and speaker recognition tasks.

REFERENCES

[1] Zissman, M.A., “Comparison of four approaches to automatic language identification of telephone speech, ”IEEE Trans. on Speech, Audio Processing, Vol.4, No.1, pp. 31-44, Jan 1996.

[2] Torres carrasquillo, P.A., Douglas A. Reynolds, and J.R. Deller, “Language identification using Gaussian mixture model tokenization ,” in Proceedings of the ICASSP, Vol.1, pp. 757-760, 2002.

[3] Ramasubramanian, V., A.K.V. Sai Jayaram, and T.V. Sreenivas, “Language identification using parallel phone recognition,” in Proceedings of the WSLP,TIFR, Mumbai, pp. 109-116, Jan 2003.

[4] Muthusamy, Y.K. and R. A. Cole, “Automatic segmentation and identification of ten languages using telephone speech,” in Proceedings of the ICSLP, pp. 1007-1010, 1992.

[5] Hema A. Murthy and VRR Gadde, “The Modified group delay function and its application to phoneme recognition,” in Proceedings of the ICASSP, Vol.I, pp. 68-71, April 2003.

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

[7] Rajesh M.Hegde, Hema A.Murthy and VRR Gadde, ” Application of the Modified Group Delay Function to Speaker Identification and Discrimination,” in Proceedings of the ICASSP, Vol 1, pp. 517-520, May 2004.

[8] Rajesh M.Hegde, Hema A.Murthy and VRR Gadde, ”Continuous Speech Recognition using Joint Features derived from The Modified Group Delay Function and MFCC,” in Proceedings of the INTERSPEECH-ICSLP, Accepted for presentation, Oct 2004.

[9] Database for Indian languages, Speech and Vision Lab, IIT Madras, Chennai, India, 2001.

[10] Muthusamy, Y.K., R.A. Cole, and B.T. Oshika, ”The OGI multilanguage telephone speech corpus,” in Proceedings of the ICSLP, pp. 895-898, Oct 1992.

[11] 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.

[12] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford, “A Global Geometric Framework for Nonlinear Dimensionality Reduction, ” Science, www.science.org, pp.2319-2323, Vol. 290, No.5500, Dec 22 , 2000.

[13] Fukunaga.K, ”Introduction to Statistical Pattern Recognition,” Academic Press, Boston, 1990.