Feature Evaluation for Speaker Identification in Radio Communications Channel

R.Padmanabhan
Department of Computer Science and Engineering
Indian Institute of Technology Madras
Chennai 600036
Email: padmanabhan@lantana.tenet.res.in

Hema A. Murthy
Department of Computer Science and Engineering
Indian Institute of Technology Madras
Chennai 600036
Email: hema@lantana.tenet.res.in

***Abstract*—This paper performs feature evaluation in a speaker identification task. The identification task was performed on a collection of utterances of various strengths over a simulated military radio communications channel. A GMM-based speaker identification system using two different types of features, one from spectral magnitude and the other from phase is used. From the experiments, it is observed that for matched conditions, magnitude-based features perform better, whereas for mismatched conditions, phase-based features perform better. We also visualise the separability amongst speakers in these feature spaces. The paper attempts to evaluate the two features, and describes the some preliminary results in mismatched channel conditions.**

I. Introduction

Speaker identification in mismatched and noisy conditions is still a challenging task. Although identification in noise-free and matched train-test conditions is almost 100% accurate, the error rate increases drastically when any of these conditions are not met.

Typically, speaker identification systems are made up of the following blocks: (i) a feature extraction system, which extracts meaningful information (typically spectral information) from the raw signal samples and (ii) a classifier, which uses information from training data to assign the feature vector to a particular class (in this case, the most probable speaker.)

Performance of speaker identification systems can be improved by either fine-tuning the feature extraction or the classifier. It has been established that a single (set of) feature(s) is not optimal across various environmental conditions [1]. We take two different types of features, and evaluate their performances on the identification task under different conditions.

This paper is organised as follows. In section II we briefly describe the dataset. In section III we describe the identification system, the feature extraction and classifiers. We summarize the robustness of phase-based features to noise in section IV. The experimental setup is described in section V and our observations in section VI. Finally, we conclude in section VII.

II. Description of the dataset

The dataset used for the identification task is a sequence of simulated military voice communications over VHF (very high frequency) radio. The quality of the recordings vary from good (strength 5) to poor (strength 3). There are different number of speakers for each set of strengths. The recordings were collected in a variety of environments including field recordings in a city street and with communicating units at various distances from the recording station. In general, the quality of recordings deteriorates as the distance increases. Also, many utterances (for eg. voice-procedure phrases like "roger") are short in duration (about 1 sec.) Since the data is from the field, there is lack of systematic methodology in its collection. For our experiments, we have tried to organise the recordings uniformly, with respect to duration and number of speakers. In spite of this, all the speakers available for a particular strength are not present in other strengths. The data is sampled at 44kHz.

Speaker recognition experiments were performed on various train-test combinations. The focus of the investigation is on mismatched train-test conditions.

III. Speaker recognition system

A. Classification system

A GMM-based speaker identification system [2] is used as the classifier. The GMM-based system has demonstrated almost 100% accuracy on clean speech (for eg. on the TIMIT corpus) and about 36% for noisy speech (for eg. on the NTIMIT corpus.) as described in [3].

B. Feature extraction

Two complementary types of features, one derived from Fourier transform magnitude and the other from phase were used as the front-end of the recognition system:

- Standard Mel-frequency cepstral coefficients (MFCC) [4] were used as the magnitude-based feature.
- Modified group delay feature (MODGDF) [3], which is derived from the phase of the signal is used as the other feature.
- Joint features which were formed by concatenating the MFCC and MODGDF were also evaluated. The joint features capture complementary information from both magnitude and phase, as described in [5].

It has been shown that features derived from the group delay is robust towards additive noise [6]. Based on this robustness property, we hypothesise that the MODGDF gives better recognition performance under mismatched conditions.

IV. Robustness of group delay features

In [6], it was shown that in additive noise, the group delay $\tau(\omega)$ for a minimum phase signal is related to the signal power and noise power in the following manner:

For $\forall\omega$ such that $P_X(\omega) \ll \sigma^2(\omega)$

$$\tau(\omega) \approx \frac{1}{\sigma^2(\omega)} \sum_{k=1}^{\infty} k\, d_k \cos(\omega\, k) \tag{1}$$

for $\forall\omega$ such that $P_X(\omega) \gg \sigma^2(\omega)$,

$$\tau(\omega) \approx \sum_{k=1}^{\infty} k\, (d_k + \sigma^2(\omega)\, e_k) \cos(\omega\, k) \tag{2}$$

and for $\forall\omega$ such that $P_X(\omega) \approx \sigma^2(\omega)$

$$\tau(\omega) \approx \sum_{k=1}^{\infty} k\, d_k \cos(\omega\, k) \tag{3}$$

where $P_X(\omega)$ is the power spectrum of the clean signal, $\sigma^2(\omega)$ is the power spectrum of additive noise. In equation (1), d_k are the coefficients of the Fourier series expansion of $P_X(\omega)$, in equation (2), d_k and e_k are the coefficients in the Fourier expansion of $\ln(P_X(\omega))$ and $\frac{1}{P_X(\omega)}$ respectively, and in equation (3), d_k are the Fourier series coefficients of $\ln(P_X(\omega))$.

From equations (1-3), we note that the group delay function of a minimum phase signal is *inversely* proportional to the noise power at frequencies corresponding to low SNR regions in the power spectrum. Similarly, for high SNR regions, the group delay function becomes *directly* proportional to the signal power. In other words, its behaviour is similar to that of the signal magnitude spectrum under clean conditions. This shows that the group delay function of a minimum phase signal preserves the peaks and valleys in the magnitude spectrum well even in the presence of additive noise.

The MODGDF is an approximation of the minimum phase group delay and hence this analysis is applicable here.

V. Experiments and results

A. Summary of experiments

Broadly, speaker recognition experiments were performed on three types of data:

- **Clean speech for train-test:** In this set of experiments, the models are trained and tested on good quality (strength 5) radio speech. The GMM-based recognition system gave 100% accuracy in identification of 18 speakers using either MFCC or MODGDF features.
- **Noisy but matched train-test:** In this set of experiments, the speech was noisy (strengths 3 and 4), but identical conditions were used for both train and test. In this case, there were 11 speakers for strength 3 and 13 speakers for strength 4.
- **Noisy and mismatched train-test:** In this set, the speech was noisy, and different strengths were used for training and testing. As expected, the performance of this set of the experiments was the poorest. There were 11 speakers in this case.

As was done in [3], a line search was performed to determine optimal parameters for the number of mixtures, number of cepstral coefficients and other parameters in the recognition system. Individual speaker models were built for each speaker, for each strength. For each experiment, 15 seconds of speech per speaker was used for training. The test utterances varied from 1 to 2 seconds. A summary of the experiments for MFCC, MODGDF and the joint features are tabulated in tables I and II. These results are obtained for an optimal set of parameters.

In this paper, "Train SX-Test SY" refers to the experiment performed by training with speech of strength X and testing with speech of strength Y.

TABLE I
SUMMARY OF EXPERIMENTS IN MATCHED CONDITIONS USING VARIOUS FEATURES. THE TRAIN AND TEST COLUMNS SPECIFY THE STRENGTH OF DATA USED. THE THIRD COLUMN GIVES THE NUMBER OF SPEAKERS IN EACH CASE.

Train	Test	Num spk	Accuracy
MFCC			
Str 3	Str 3	11	90%
Str 4	Str 4	13	90%
Str 5	Str 5	18	100%
MODGDF			
Str 3	Str 3	11	85%
Str 4	Str 4	13	83%
Str 5	Str 5	18	100 %

TABLE II
SUMMARY OF EXPERIMENTS IN MISMATCHED CONDITION USING VARIOUS FEATURES. THE TRAIN AND TEST COLUMNS SPECIFY THE STRENGTH OF DATA USED. THERE WERE 11 SPEAKERS.

Train	Test	Accuracy
MFCC		
Str 3	Str 4	42%
Str 4	Str 3	43%
MODGDF		
Str 3	Str 4	48%
Str 4	Str 3	57%
Joint		
Str 3	Str 4	50%
Str 4	Str 3	61%

B. Observations

From the above tables, it is observed that on an average, MFCC performs better in matched conditions, whereas in mismatched conditions MODGDF performs better. The joint features give the best average performance in the mismatched case. The bulk of the improvement in performance is seen when MFCC is replaced by MODGDF. The improvement over this on using joint features is lesser but still significant.

Surprisingly, it was found that cepstral mean subtraction [7] was detrimental to performance, even in mismatched conditions, for both types of features. This could be due to the possibility that different radio sets were used by the same speaker at different times. Therefore, in all experiments, cepstral mean subtraction is not done.

VI. Analysis of experiments

A. Separability of speakers

This is illustrated in the separability of a pair of confusable speakers in 2-dimensional space. Sammon mapping [8], a non-linear dimension reduction technique is used to plot the 18-dimensional MODGDF features in 2-dimensional space. Figures 1 and 2 show the cluster centroids of two speakers (speaker 6 and speaker 10) in the MFCC and MODGDF feature spaces respectively in experiment Train S3-Test S4. It can be seen that the clusters are better separable in MODGDF space when compared to the MFCC space.

Also, it can be noted from the table that Train S4-Test S3 gives better identification accuracy when compared to Train S3-Test S4. This is due to the fact that models are better built with the higher quality (strength 4) speech.

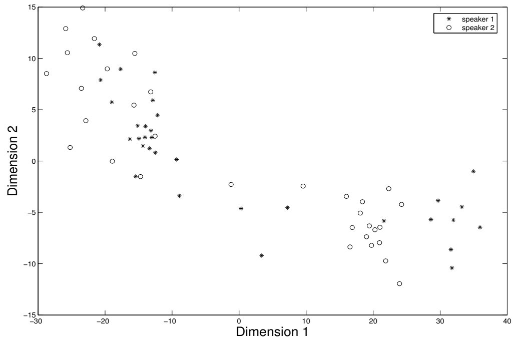

Fig. 1. 2-D plot of MFCC feature space for speakers 6 and 10

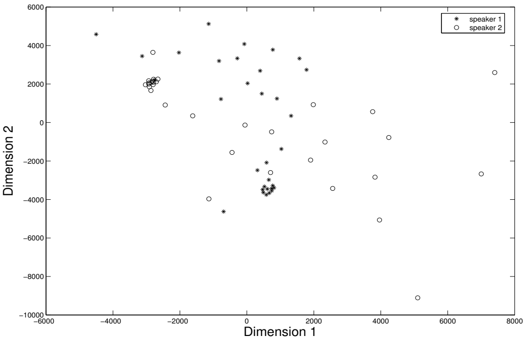

Fig. 2. 2-D plot of MODGDF feature space for speakers 6 and 10

The Bhattacharya distance [9] is a measure of similarity amongst distributions and can be used for analysing separability of the corresponding classes. Figure 3 gives the plot of Bhattacharya distance versus feature dimension for MODGDF and MFCC for all eleven speakers. The plot was done on strength 3 speech using the TOOLDIAG pattern recognition toolbox [10]. It can be seen that as the dimension of the feature vector increases, the MODGDF features gives better separability.

On considering the two-best results, the identification accuracy improved considerably as shown in table III. This shows that in many test cases, the correct speaker is confused with some other speaker just one rank above it.

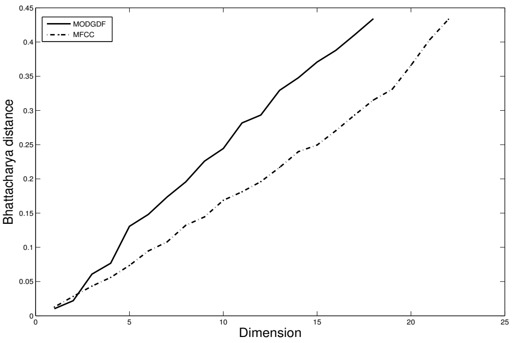

Fig. 3. Bhattacharya distance versus feature dimension for MFCC and MODGDF for strength 3 speech.

TABLE III
Summary of experiments in mismatched condition using two-best results.

Train	Test	Accuracy
MFCC		
Str 3	Str 4	68%
Str 4	Str 3	72%
MODGDF		
Str 3	Str 4	60%
Str 4	Str 3	75%
Joint		
Str 3	Str 4	69%
Str 4	Str 3	73%

B. Using feature switching

Figure 4 shows Bhattacharya distance versus feature dimension for speakers 7 and 10 for strength 3 speech. It is seen that in this case, MFCC features shows better separability than MODGDF features. This fact is also confirmed from the experimental results: MFCC features have higher identification accuracy for these two speakers. This prompts us to hypothesise dynamic *feature switching*, as some speakers are better recognised by one (set of) features than another. Implementing this scheme is non-trivial, because the range of probabilities for both features are not directly comparable. Further studies are being made in this direction.

VII. Conclusions

This paper presents some preliminary results on performing speaker identification on radio speech. Radio speech collected from the field presents challenges different from telephone speech. Experiments were performed on clean, matched and mismatched conditions. Three types of features, MFCC, MODGDF and joint features were evaluated. From

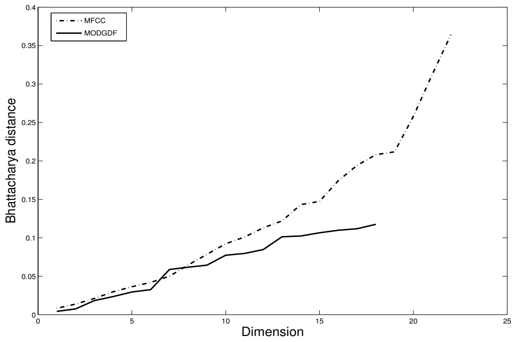

Fig. 4. Bhattacharya distance versus feature dimension for MFCC and MODGDF for speakers 7 and 10.

the experiments, it is observed that in matched conditions MFCC features performed better, whereas in mismatched conditions, MODGDF features performed better. Separability amongst confusable speakers is improved using MODGDF features. On taking the 2-best result, there is an improvement in identification accuracy. Finally, we also mentioned the idea of feature switching based on separability.

References

[1] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, pp. 1711–1723, 2007.

[2] D. Reynolds and R. Rose, "Robust text independent speaker identification using gaussian mixture speaker models," *IEEE Trans. on Speech and Audio Processing*, vol. 3, pp. 72–82, 1995.

[3] 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, pp. 190–202, 2007.

[4] P. Mermelstein and S. B. Davis, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 28, pp. 357–366, 1980.

[5] 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 of Applied Signal Processing*, 2007.

[6] R. Padmanabhan, S. H. Krishnan, and H. A. Murthy, "A pattern recognition approach to VAD using modified group delay," in *Proceedings of the fourteenth National Conference on Communications*. Indian Institute of Technology Bombay, 2007.

[7] D. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 639–643, 1994.

[8] J. W. Sammon, "A non-linear mapping for data structure analysis," *IEEE Trans. on Computers*, vol. 18, pp. 401–404, 1969.

[9] C. Bishop, *Pattern Recognition and Machine Learning*. Cambridge University Press, 2006.

[10] "TOOLDIAG pattern recognition toolbox," http://www.inf.ufes.br/~thomas/home/tooldiag.html.