A Novel Feature Extraction Technique for Speaker Identification

Amita Dev
Bhai Parmanand Institute of Business Studies, Department of Training & Technical Education
India - New Delhi-110092

ABSTRACT
This paper presents a novel feature extraction approach for speaker identification when the speech is corrupted by additive noise. The environmental mismatch between training and testing data degrades the performance of speaker identification system. The performance degradation is primarily due to presence of background noise when try to match a given speaker to the set of known speakers in a database. Mel frequency cepstral coefficients (MFCCs) are perhaps the most widely used front ends in the state of the art speaker identification systems. One of the major issues with MFCCs is that they are very sensitive to additive noise. To overcome this bottleneck, a temporal filtering procedure on the autocorrelation sequence is proposed to minimize the effect of additive noise. The proposed feature is called Relative Autocorrelation Mel Frequency Cepstral Coefficients (A-MFCC) which is derived based on filtering the temporal trajectories of short time one sided autocorrelation sequence. This filtering process minimizes the effect of additive noise. No prior knowledge of noise characteristics is required. The additive noise can be a colored noise. For speaker identification, Hindi database was constructed from the speech samples of each known speaker. Feature vectors (MFCCs and A-MFCCs) were extracted from the samples by short-term spectral analysis, and processed further by vector quantization for locating the clusters in the feature space. Experimental results indicated that A-MFCCs significantly improved the performance of speaker identification system in noisy environment.

KEYWORDS
Speaker identification, vector quantization, relative autocorrelation

1. INTRODUCTION
Speaker recognition is a generic term used for two related problems: speaker identification and verification. In the identification task the goal is to recognize the unknown speaker from a set of N known speakers. In verification, an identity claim (e.g., a username) is given to the recognizer and the goal is to accept or reject the given identity claim. In this work we concentrate on the identification task.

The input of a speaker identification system is a sampled speech data, and the output is the index of the identified speaker. There are three important components in a speaker identification system: the feature extraction component, the speaker model and the matching algorithm. Feature extractor derives a set of speaker-specific vectors from the input signal. Speaker model is then generated from these vectors for each speaker. The matching procedure performs the comparison of the speaker models. It is expected that the feature extraction is the most critical component of the system but it is also much more difficult part to be designed than the matching procedure. The environmental mismatch between training and testing data drastically degrades the performance of speech or speaker identification system. The mismatch between training and testing environment is often due to background noise and channel distortion. The robustness of the speaker identification system can be accomplished in three ways. Firstly using speech enhancement technique to increase signal to noise ratio. Secondly extracting the robust parameters of speech signals to minimize the effect of the noise on the speech signal. Thirdly using model compensation technique to dynamically adapt clean speech model to the noisy environment. Many techniques have been proposed to overcome this degradation problem [1], such as Parallel Model Compensation (PMC) [2], stochastic matching (SM) [3], combining channel identification with proper spectrum estimation etc. Although a variety of techniques can demonstrate the comparable performance but some weakness may limit their practical application. For example spectral subtraction and Parallel Model Compensation (PMC) need a priori knowledge of the noise characteristics. In this paper we propose a method to remove noise effect based on the idea of temporal filtering in autocorrelation domain and the Mel frequency cepstral coefficient are derived from it and named as A-MFCCs. When a speech is corrupted by additive noise, the noise component is additive to the speech not only in the power spectral domain, but also in the autocorrelation domain. Instead of subtracting the noise in the power spectral domain, we remove the noise in the autocorrelation domain based on the temporal trajectory filtering.

2. SPEAKER IDENTIFICATION SYSTEM
The structure of a VQ-based speaker identification system is illustrated in Fig. 1. There are two phases in the speaker identification: training and recognition. In the training phase, a mathematical model (VQ codebook in our case) is constructed for each speaker from their speech samples and the models are stored in the database. In recognition phase, the speech data of an unknown speaker is analyzed and the best matching model is searched from the database.

The analysis of the speech signals is based on short-term spectral analysis and computing of MFCCs and A-MFCCs. The speech signal is decomposed into short fixed-length speech frames, which form the feature vectors. The extracted feature vectors are processed further by vector quantization for locating the clusters in the feature space and for reducing the amount of data. The input of vector

quantization is the set of feature vectors X and the output is a codebook C that consists of the cluster centroids, denoted as code vectors. The codebook represents the speaker model by approximating the distribution of the feature vectors in the feature space.

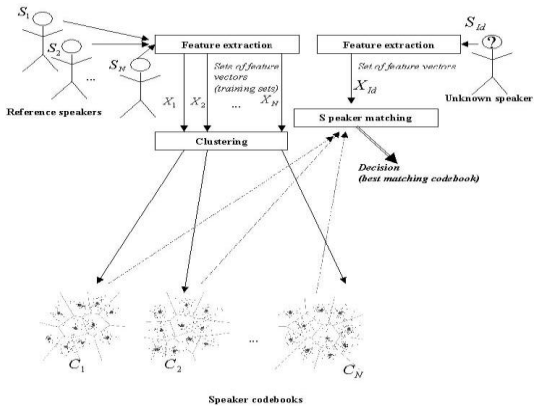

Fig.1 Vector Quantization base Speaker Identification System

The identification procedure is formulated as follows:

1. Compute the set of feature vectors X = { xi }
2. FOR EACH speaker model Ci DO Compute the distortion Di = d(X, Ci) between X and Ci.
3. Identify the index of the unknown speaker Id as the one with the smallest distortion, i.e.

$$Id = \arg\min\{D_i\}, \quad i = 1, 2, \ldots, N$$

The distortion measure d in the second step approximates the dissimilarity between the codebook Ci = {ci1, ci2,..., ciK } and the vector set X ={x1, x2,..., xL } . We use the most intuitive distortion measure; map each vector in X to the nearest code vector in Ci and compute the average of these distances which is known as Mean square Error (MSE):

$$d(X, C_i) = \frac{1}{L}\sum_{j=1}^{L} \min_{k=1}^{K} \; d_E(x_j, c_{ik})$$

Where d_E is the Euclidean distance.

Note that in training phase we generated codebooks for the speakers, but in the recognition phase we performed a direct comparison between the set of feature vectors and the codebooks of the known speakers. There are two good reasons for this: memory and time requirement. Computational load of the identification process becomes too high if we do not reduce the amount of data. It is pertinent to remove this kind of bottleneck from a real-time speaker identification system. Memory consumption could also be a restricting factor in case of very large databases.

We assume that the feature vectors discriminate well the different acoustical units in the speech signal; similar phonemes (vectors) are located near to each other in the feature space while different phonemes are far away from each other. When we perform the clustering of the feature vectors, we obtain efficient mean values of these different short-term acoustical units. The codebooks of different speakers may have some vectors very close to each other, but it is expected that there are enough dissimilar vectors so that the matching process can differentiate between codebooks of different speakers.

3. FEATURE EXTRACTION

3.1 MEL FREQUENCY CEPSTRAL COEFFICIENTS (MFCC)

The purpose of feature extraction is to convert the speech waveform to some type of parametric representation (at a considerably lower information rate) for further analysis and processing, which is referred as the signal-processing front end. The speech signal is a slowly time-varying signal (called quasi-periodic) when examined over a sufficiently short period of time (5 ~ 100 ms), its characteristics are fairly stationary. However, over long periods of time (on the order of 1/5 seconds or more) the signal characteristic change to reflect the different speech sounds being spoken. Therefore, the short-time spectral analysis is the most common way to characterize the speech signal. MFCC is perhaps the best known and most popular, and has been used in this paper. MFCC is based on the known variation of the human ear's critical bandwidths with frequencies, with filters spaced linearly at low frequencies and logarithmically at high frequencies. This is expressed in the Mel-frequency scale, linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. MFCC provides a substantial data reduction, because a few coefficients are sufficient to represent the cepstrum of the acoustic signal. Computation of MFCCs has been explained in Fig. 2. The speech signal is divided into frames of 256 samples each, and a pre-emphasis filter is applied on each frame. Pre-emphasis coefficient used is 0.9375. A Hamming window is used to minimize the signal discontinuities at the beginning and end of each frame and then FFT is computed. Finally MFCC coefficients are derived by passing the resultant magnitude through the mel-frequency filter bank followed by discrete cosine transform (DCT).

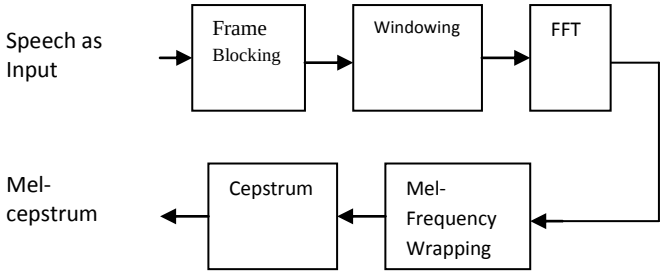

Fig. 2 Block Diagram of MFCC Processor

3.2 ROBUST FEATURE EXTRACTION [A-MFCC]

In this paper we propose a new approach, utilizing peaks obtained from the autocorrelation spectrum of the speech signal. This approach preserves the autocorrelation spectral peaks [4]. Computation of A-MFCCs has been explained in Fig. 3. Firstly, we calculate the autocorrelation of the noisy signal. As the temporal autocorrelation of noise is a DC or slowly varying signal, its effect is suppressed by a high-pass filter. The autocorrelation sequence of the frame signal is obtained using a biased estimator.. A temporal filtering is then applied to the autocorrelation sequence to obtain the relative autocorrelation sequence in order to suppress the additive noise. A set of robust mel-frequency cepstral coefficients are derived from the magnitude of the relative autocorrelation power spectrum by applying it to a conventional mel- frequency filter-bank and finally passing its logarithm to the DCT block.

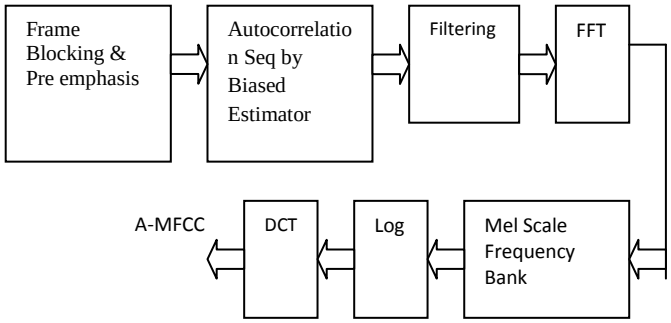

Fig. 3 Block Diagram of A-MFCC Processor

4. DATA SET AND EXPERIMENTAL RESULTS

A digital database of 200 Hindi words spoken by 30 speakers (Table 1) has been used for the experiment of speaker identification system. The spoken samples are recorded by 15 male, 10 female and 5 child speakers in the studio environment using the Sennheiser microphone model MD421 and a tape recorder model Philips AF6121. Each speaker pronounced 5 repetitions of words. The resulting database was partitioned for the use of training and testing.

1. Language : Standard Hindi (Khari Boli)

2. Vocabulary Size : A set of 200 most frequently occurring Hindi words

3. Speakers : 30 speakers

4. Utterances : (15 male, 15 female and 5 children) 5 repetitions each

5. Audio Recording: Recording on a cassette tape in studio SNR>40dB

6. Digitization : 16KHz. Sampling, 16 bit quantization.

Table 1. Specifications of Hindi Database

4.1 Testing on Clean Speech

The purpose of this experiment is to evaluate the performance of MFCC, A-MFCC when training data and the testing data are in a clean environment, i.e., assuming 40 dB signals to noise ratio (SNR). We observe that recognition rates are approximately identical for MFCC and A-MFCC. With the use of MFCC front end, the speaker identification rate was 98.24% and with A-MFCC it was 99.27% as given in Table 2.

Feature type	Recognition rate (%)
MFCC	98.24
A-MFCC	99.27

Table 2. Speaker identification rate (%) for clean speech

Testing on noisy speech

The polluted testing utterances are generated by adding the artificial noises at five SNR levels. The white noise is generated by a random number generator program, and other colored noises such as factory noise, F16 noise are extracted from the NATO RSG-10 corpus [5]. The noises are added to the clean speech signal at 20, 15, 10, 5 and 0 dB of SNR. Both MFCC and A-MFCC are evaluated and the speaker identification rates are compared with the traditional MFCC front-end. Table 3(a)-(c) show the results obtained by using MFCC, A-MFCC front-ends respectively. From the result it is obvious that A-MFCC are quite robust to the additive noise.

Feature type	Noise levels (dB)					
	40	20	15	10	5	0
MFCC	98.2	83.8	55.8	29.3	10.5	3.7
A-	99.2	85.8	58.9	34.8	15.0	7.4

Table 3 (a) Speaker Identification rate (%) for testing speech corrupted by white noise.

Feature type	Noise levels (dB)					
	40	20	15	10	5	0
MFCC	99.24	84.10	56.17	30.18	11.50	4.20
A-	99.31	86.11	58.90	35.16	16.15	8.20

Table 3 (b) Speaker Identification rate (%) for testing speech corrupted by factory noise.

Feature type	Noise levels (dB)					
	40	20	15	10	5	0
MFCC	98.0	83.2	57.1	31.4	10.8	3.7
A-	98.9	85.9	59.1	35.7	14.9	7.2

Table 3(c) Speaker Identification rate (%) for testing speech corrupted by F 16 noise.

5. CONCLUSION

We evaluated the performance of VQ-based speaker identifier using two different front ends in this paper, cepstral features derived from autocorrelation spectral domain are proposed in order to improve the robustness of the speaker identification systems. In the case of noisy environment, A-MFCC contributes to better performance in terms of speaker identification. Experimental results show that the proposed approach is more effective in overcoming additive noises which are stationary in nature at low SNR's. Furthermore, this proposed method works well for different types of noises including white, F16 and factory noise.

6. REFERENCES

[1] Y. GONG, Speech recognition in noisy environments: A survey. *Speech Communication*, Vol. 16 (1995), pp. 261–291.

[2] S. F. BOLL, Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustic Speech and Signal Processing*, 27 (2), (1979), pp. 113–120.

[3] HERMANSKY AND MORGAN, RASTA processing of speech. *Speech Communication*, Vol. 41, (2003), pp. 469–484.

[4] J. HERNANDO AND C. NADEU, Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition. *IEEE Trans. Speech Audio Processing*, Vol. 2, No. 5, (1994) pp. 578-586.

[5] A. VARGA AND H. J. M. STEENEKEN, Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, Vol. 12, (1993), pp. 247–251.