ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 2, February 2013

Speaker Identification Using Combined MFCC and Phase Information

Nisha.V.S[1] , M.Jayasheela[2]

PG Scholar, Department of ECE, SNS College of Technology, Coimbatore, India[1]

Associate Professor, Department of ECE, SNS College of Technology, Coimbatore, India[2]

ABSTRACT : **Speaker recognition is the identification of the person who is speaking by the characteristics of their voices. To improve the performance of speaker recognition systems, an effective and robust method is proposed to extract speech features, capable of operating in noisy environment. For capturing the characteristics of the signal, the Mel-Frequency Cepstral Coefficients (MFCC) are calculated. Gaussian Mixture Models (GMMs) are used for the recognition stage as they give better recognition for the speakers' features. In conventional speaker recognition methods based on MFCC, phase information has been ignored. The proposed method integrated the phase information with MFCC on the speaker recognition method. Comparison of the proposed approach with the MFCCs conventional feature extraction method shows that the proposed method improves the recognition rate.**

Keywords : Speaker Identification, Mel-Frequency Cepstral Coefficients (MFCCs), Phase Information, GMM, EM Algorithm.

I.INTRODUCTION

Recent development has made it possible to use speech in the security system. Speaker recognition can be classified into speaker identification and speaker verification. Speaker identification is the process of determining which registered speaker provides a given utterance. Speaker recognition methods can be divided into text independent and text dependent methods[4]. In a text independent system, speaker models capture characteristics of somebody's speech which show up irrespective of what one is saying. In a text dependent system, the recognition of the speaker's identity is based on his or her speaking one or more specific phrases or words.

Speaker recognition systems contain two main modules: feature extraction and feature matching [3]. Feature extraction is the process of extracting a small amount of data from the voice signal that can be later used to represent each speaker. Feature matching involves the actual procedure to identify the unknown speaker by comparing extracted feature from his/her voice input with the ones from a set of known speakers.

The speaker recognition systems are presented in two phases – training phase and testing phase. In the training phase, each registered speaker has to provide samples of their speech so that the system can build or train a reference model for that speaker. In the testing phase, the input speech is matched with stored reference models and a recognition decision is made.

Speaker recognition is a difficult task. The principle source of variance is the speaker himself/herself.

Speech signals in training and testing sessions can be greatly different due to many facts such as people voice changes with time, health conditions, speaking rates, and so on. There are also other factors, beyond speaker variability that present a challenge to speaker recognition technology [7].

One of the first decisions in any pattern recognition system is the choice of what features can be used and how exactly to represent the basic signal that is to be classified, in order to make the classification task easiest and accurate. Through many years of research, many different feature extraction techniques have been suggested and tried.

MFCC is the best known and most popular feature extraction technique [8], and this feature has been used in this paper. MFCC's are based on the known variation of the human ear's critical bandwidths with frequency. The technique makes use of two types of filter, namely, linearly spaced filters and logarithmically spaced filters. To capture the phonetically important characteristics of speech, signal is expressed in the Mel frequency scale. This scale has a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. Normal speech waveform may vary from time to time depending on the physical condition of speakers' vocal cord. MFCCs are less susceptible to the said variations.

Current speech recognition system can achieve high recognition accuracy rates (>90%). The system performance can be further improved by using a more complicated recognition model; one which takes in and processes more information. Most state-of-the-art speech recognition systems only utilize the magnitude of the Fourier transform of the time-domain speech segments. This means that the corresponding Fourier transform phases are discarded. Several studies have indicated that it may be a fruitful effort to directly model and incorporate the phase into the recognition process. Also, several studies have shown the importance of phase in speech coding.

In this paper, a highly robust speech recognition system which incorporates the phase information is proposed in order to improve the recognition rate of the system. The system described is text independent speaker recognition system since its task is to identify the person who speaks regardless of what is saying.

With regards to speaker recognition, various types of speaker models have been studied over time. The Gaussian mixture model (GMM) has been widely used as a speaker model [2], [4]. The use of GMM for modeling speaker identity is motivated by the fact that the Gaussian components represent some general speaker-dependent spectral shapes and by the capability of Gaussian mixtures to model arbitrary densities.

The rest of the paper is organized as follows : Section II gives a description about feature extraction technique. Section III investigates the importance of phase for speaker recognition and formulates the phase information. Section IV briefly describes the speaker recognition method used. The experiments and the results obtained are given in section V. Finally, Section VI summarizes the paper and describes future work.

II. FEATURE EXTRACTION USING MFCC TECHNIQUE

Speech is a complicated signal produced as a result of several transformations occurring at several different levels – semantic, linguistic, articulatory and acoustic. Differences in these transformations appear as differences in the acoustic properties of the speech signal. An important problem in speech recognition systems is to determine a representation that is well adapted for extracting the information content of speech signals. Generally, transformation of a signal to a different domain is done to get a better representation of the signal. Better recognition techniques having more ability to separate signals which belong to separate categories in the new domain than in the original domain are required.

A block diagram of the structure of an MFCC processor is given in Fig 1. The speech input is recorded at a sampling rate above 8000 Hz. This sampling frequency is chosen to minimize the effects of *aliasing* in the analog-to-digital conversion process. Fig 1 shows the block diagram of an MFCC processor .

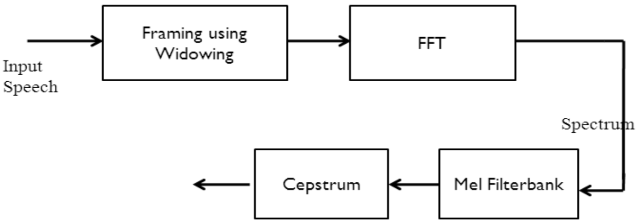

Fig 1 Block Diagram of the MFCC Processor

The speech input is typically recorded at a sampling rate above 10000 Hz. This sampling frequency was chosen to minimize the effects of *aliasing* in the analog-to-digital conversion. These sampled signals can capture all frequencies up to 5 kHz, which cover most energy of sounds that are generated by humans.

In frame blocking, the continuous speech signal is divided into frames of N samples, with adjacent frames being separated by M ($M < N$). The first frame consists of the first N samples. The second frame begins M samples after the first frame, and overlaps it by $N - M$ samples and so on. This process continues until all the speech is accounted for within one or more frames.

The next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. The concept here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame. Hamming window is used in this paper.

The next processing step is the Fast Fourier Transform, which converts each frame of N samples from the time domain into the frequency domain. The result after this step is often referred to as spectrum or periodogram.

Psychophysical studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the ‘mel’ scale. A filter bank spaced uniformly on the mel scale is used to obtain the spectrum. The filter bank has a triangular bandpass frequency response, and the spacing as well as the bandwidth is determined by a constant mel frequency interval. The number of mel spectrum coefficients, K, is typically chosen as 20.

In the final step, the log mel spectrum is converted back to time. The result is called the mel frequency cepstrum coefficients (MFCC). The cepstral representation of the speech spectrum is a good representation of the local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients and their logarithm are real numbers, they can be converted to the time domain using the Discrete Cosine Transform (DCT).

Thus each input utterance is transformed into a sequence of vectors which can be used to represent and recognize the voice characteristic of the speaker.

III. PHASE INFORMATION ANALYSIS

In this section, we first investigate the effect of phase on speaker recognition, and then formulate the phase information.

A. Investigating the Effect of Phase

The conventional MFCCs ignore the phase information and so they cannot capture all the speaker characteristics contained in a voice source with the same power spectrum, but a different phase. In other words, speaker characteristics in the voice source are not captured completely by the MFCC since the phase information is ignored. The phase is greatly influenced by voice source characteristics. Of course, the phase is also influenced by pitch.. In this paper, the distribution of phase for a speaker was modeled by GMM [4]. The phases are similar and the power spectra are influenced greatly by the vocal tract with the same voice source. To capture the speaker characteristics in the voice source and vocal tract exactly, both power spectrum and phase information of the input speech are required. The combined usage of MFCC and phase information can help to distinguish the speaker characteristics.

B. Formulation of Phase Information

The short-term spectrum S(ω,t) for the ith frame of a signal is obtained by the DFT of an input speech signal sequence

$$S(\omega,t) = X(\omega,t) + jY(\omega,t)$$
$$= \sqrt{(X^2(\omega,t) + Y^2(\omega,t))} \text{ x } e^{j\theta(\omega,t)} \qquad (1)$$

For conventional MFCCs, the power spectrum $X^2(\omega,t) + Y^2(\omega,t)$ is used, but the phase information $\theta(\omega,t)$ is ignored. In this paper, phase $\theta(\omega,t)$ is also extracted as one of the feature parameter set for speaker recognition. The GMMs used in this paper are insensitive to the temporal aspects of speech, and do not capture the dependence of features extracted from each frame [1]. Phase information of the same person with the same voice extracted from different frames may be $\theta(\omega,t)$ and $2\Pi + \theta(\omega,t)$. They express different phase value and the different speaker characteristics using phase-based GMMs.

In this paper, the phase value is constrained to $[-\Pi, \Pi]$. Thus $\theta(\omega,t)$ and $2\Pi + \theta(\omega,t)$ are converted to the same phase value. Therefore, it is no problem to use GMMs to model the speaker characteristics using phase information [9], [10], [11].

IV. SPEAKER RECOGNITION METHOD

A. *Combination Method*

A Gaussian mixture model (GMM) has been widely used as a speaker model [2], [4], [5], [6]. The use of GMM for modeling speaker identity is motivated by the fact that the Gaussian components represent some general speaker-dependent spectral shapes and by the capability of Gaussian mixtures to model arbitrary densities.

In this paper, the GMM based on MFCCs is combined with the GMM based on phase information. When a combination of two methods is used to identify the speaker, the likelihood of MFCC-based GMM is linearly coupled with that of the phase information-based GMM to produce a new score $L_{comb}{}^{n}$ given by [4], [18]

$$L^n{}_{comb} = (1- \alpha)\, L^n{}_{MFCC} + L^n{}_{Phase} \text{ ; } n=1,2,3.. \qquad (2)$$

where $L^n{}_{MFCC}$ and $L^n{}_{Phase}$ are the likelihood produced by the nth MFCC-based speaker model and phase information-based speaker model, respectively. N is the number of speakers registered and αdenotes weighting coefficients. A speaker with the maximum likelihood is decided as the target speaker.

B. Decision Method

In speaker identification, the speaker with the maximum likelihood is chosen as the target speaker. Therefore, likelihood normalization is crucial in dealing with real-world data for speaker identification. Expectation maximization algorithm is used to find the maximum likelihood function.

V. EXPERIMENTS AND RESULTS

The speech samples are stored in Microsoft wave format files with 8000 Hz sampling rate. Using MFCC technique, feature vectors are obtained from the speech sample. For MFCC, the Mel filter bank is designed with 24 frequency bands. In the calculation of all the features, the speech signal is partitioned into frames; the frame size of the analysis is 256 samples with 100 samples overlapping.

GMM was used to model statistically the characteristic features of the phonemes that are present in the utterances. GMM was trained using Estimation and Maximization algorithm for finding the maximum likelihood solution for a model with latent variables, to test the later speeches against the database of all speakers who enrolled in the database. For each unknown person who is to be recognized, features are extracted from his voice sample, followed by calculation of model likelihood of all models, and followed by the selection of the person whose model likelihood is highest.

NTT database was used for the experiment. The NTT database consists of recordings of 35 speakers (22 males and 13 females) collected in 5 sessions over 10 months in a sound proof room. For training the models, 5 same sentences for all speakers from one session were used. They were uttered by a normal speaking style mode. Five other sentences every the other four sessions were also uttered at normal speed and used as test data. The average duration of the sentences is about 4 seconds. GMMs with 32 mixtures having diagonal covariance matrices were used as speaker models.

Fig 2 shows the cepstral representation of a speech sample from the database. Using the extracted features, GMM models were created and Fig 3 shows the GMM model of the speech sample. The speakers were tested using MFCC and GMM , ignoring the phase information and the accuracy was calculated. The experiment was done after combining the phase information with the conventional MFCC and the accuracy was obtained.

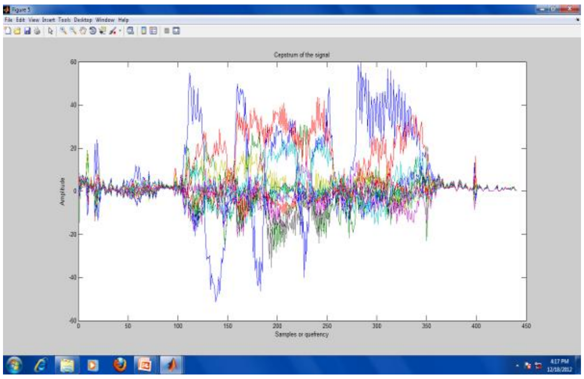

Fig 2 Cepstrum of a speech sample

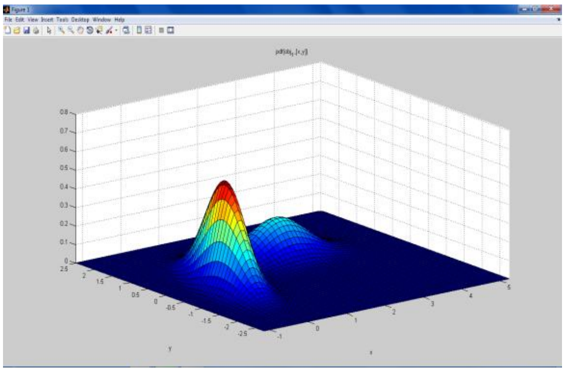

Fig 3 GMM Model of the speech sample

By using the combination of MFCC and phase information on clean speech training data , the identification rate was improved to 99.3 % in comparison with 97.7 % using only MFCC.

VI. CONCLUSION

In this paper, an effective and robust technique was used for the speaker identification systems. This technique integrated the phase information with MFCC. By using the combination of MFCC and phase information on clean speech training data, the identification rate was improved to 99.3 % in comparison with 97.7 % using only MFCC.

The future work is to analyze the performance of the speaker identification system in the presence of noise and echo.

ACKNOWLEDGMENT

We would like to thank all the staff members for their help in conducting the speech recognition experiments and the reviewers who provided constructive comments for this paper.

REFERENCES

[1] Seiichi Nakagawa, Longbiao Wang and Shinji Ohtsuka,"Speaker Identification and Verification by combining MFCC and Phase Information", IEEE transactions on Audio, Speech, and Language Processing, Vol. 20, No 4, May 2012.

[2] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," Speech Communication., Vol. 17, No. 1–2, pp. 91–108, 1995.

[3] J. P. Campbell, "Speaker recognition: A tutorial," Proc. IEEE,Vol. 85, No. 9, pp. 1437–1462, Sep. 1997.

[4] D. A. Reynolds, T. F. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," Digital Signal Processing,Vol. 10, No. 1–3, pp. 19–41, 2000.

[5] K. P. Markov and S. Nakagawa, "Integrating pitch and LPC-residual information with LPC-cepstrum for text-independent speaker recognition," J. ASJ (E), Vol. 20, No. 4, pp. 281–291, 1999.

[6] L.Wang, K. Minami, K. Yamamoto, and S. Nakagawa, "Speaker identification by combining MFCC and phase information in noisy environments," Proceedings. ICASSP, pp. 4502–4505, 2010.

[7] Fatma zohra Chelali, Amar.Djeradi, Rachida.Djeradi,"Speaker Identification System based on PLP Coefficients and Artificial Neural Network",Proceedings of the World Congress on Engineering, London, U.K., Vol II WCE, July 6 - 8, 2011

[8] Md.Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani Md. Saifur Rahman,"Speaker Identification Using Mel Frequency Cepstral Coefficients",3rd International Conference on Electrical & Computer Engineering ICECE, Dhaka, Bangladesh , 28-30 December 2004.

[9] R. Padmanabhan, S. Parthasarathi, and H. Murthy, "Robustness of phase based features for speaker recognition," in Proceedings Interspeech, pp. 2355–2358, 2009.

[10] K. K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception," in Proceedings. Eurospeech'03, pp. 2117–2120, 2003

[11] G. Shi et al., "On the importance of phase in human speech recognition," IEEE Trans. Audio, Speech, Lang. Process., Vol. 14, No. 5, pp. 1867–1874, Sep. 2006.