A Hybrid Approach to Speaker Recognition in Multi-speaker Environment

Jigish Trivedi, Anutosh Maitra, and Suman K. Mitra

Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, Gujarat, India
{jigish_trivedi, anutosh_maitra, suman_mitra}@da-iict.org

Abstract. Recognition of voice in a multi-speaker environment involves speech separation, speech feature extraction and speech feature matching. Though traditionally vector quantization is one of the algorithms used for speaker recognition; its effectiveness is not well appreciated in case of noisy or multi-speaker environment. This paper describes the usability of the Independent Component Analysis (ICA) technique to enhance the effectiveness of speaker recognition using vector quantization. Results obtained by this approach are compared with that obtained using a more direct approach to establish the usefulness of the proposed method.

Keywords: Speech recognition, ICA, MFCC, Vector Quantization.

1 Introduction

The automatic recognition process of human voice by a machine involves both speech recognition and speaker recognition. Speech recognition is the process by which a computer maps an acoustic speech signal to text. Speaker recognition [1] is the process of recognizing the speaker on the basis of individual information present in the speech waves. In a multi-speaker environment, speech signal may be corrupted by the speech of other speakers, by presence of noise and reverberation, or a combination of both. The quality of the degraded speech will affect the performance of feature extraction for various applications like tracking a moving speaker, speech recognition, speaker recognition, and audio indexing. Thus, it is imperative to enhance the voice from the degraded speech before using it in any application. In this paper, a hybrid approach for speaker recognition is presented where the speech recognition and separation issue has been addressed by a trial application of Independent Component Analysis (ICA); and the speaker recognition technique involves Mel Frequency Cepstrum Coefficient (MFCC) representation of the speech signals and a Vector Quantization (VQ) approach to the feature matching to identify individual speakers. Interestingly, an earlier work in the field of computational auditory scene analysis that focused on the problem of speech segregation concluded that blind source separation techniques could be remarkably powerful if certain requirements on the properties of the source signal are met [2]. This observation was one of the motivating factors in applying the ICA in the speech separation process.

S.K. Pal et al. (Eds.): PReMI 2005, LNCS 3776, pp. 272 – 275, 2005.

2 Problem Formulation

The speech separation and speaker recognition problems are modeled individually.

Speech Separation: The Independent Component Analysis (ICA) [3,4] has been used for separating the source signals from the information of the mixed signals observed in each input channel. Assuming that we observe n linear mixtures $X=\{x_1, x_2..., x_n\}$ of n independent components $S= \{s_1, s_2..., s_n\}$; the basic problem of source separation could be viewed as modeling the above information in the form of X=AS, where A is the mixing matrix. The task here is to obtain an unmixing matrix W such that $W^TX=S$. An improved variant for ICA, known as the FastICA [5] algorithm, is used for the specific task of speech separation in the current work. FastICA uses a learning rule to find the unmixing matrix to separate mixed signals [5]. The source speech signals are assumed to be nongaussian and the nongaussianity is measured by the approximation of negentropy [5].

Speaker Recognition: This comprises of speaker identification and speaker verification. Speaker identification is the process of determining which registered speaker provides a given utterance and it involves speech feature extraction. Speaker verification is the process of accepting or rejecting the identity claim of a speaker with the help of feature matching. The aim of speech feature extraction is to convert the quasi-stationary speech waveform to a parametric representation. A range of techniques exists to suit this purpose; the principal ones being Linear Prediction Coding (LPC) and Mel-Frequency Cepstrum Coefficients (MFCC). MFCCs are based on the filters spaced linearly at low frequencies and logarithmically at high frequencies to capture the phonetically important characteristics of speech. The structure of an MFCC processor consists of 5 well-established different steps, *viz.* Frame blocking, Windowing, FFT, Mel Frequency Wrapping and Cepstrum computation [1]. Feature matching follows the feature extraction process. Amongst the many available feature matching techniques, a vector quantization (VQ) based approach for feature matching with LBG binary split [6] has been preferred here due to its reported high accuracy with relative ease of implementation.

3 Proposed Methodology

For the speech recognition process, only clean speech features are required. In the proposed model, it was conceived that instead of denoising the noisy speech signal in the pre-processing step, it could be computationally more efficient and accurate to directly separate the clean speech features from noisy speech or separate the speech uttered by different speakers.

In a conventional direct approach as depicted in Figure 1, the source signals are randomly mixed to generate an equal number of mixed signals. In real environment, this mixing is achieved using equal number of speakers and microphones. The word 'Direct' signifies that the mixed signals are directly fed to the MFCC-VQ algorithm. In contrast, in the proposed hybrid approach as shown in Figure 2, the mixed signals are processed using the FastICA technique and only the estimated signals are used by the MFCC-VQ algorithm for speaker identification.

Fig. 1. Direct Approach

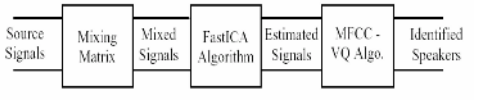

Fig. 2. The Proposed Hybrid Approach

4 Experimental Results

The proposed hybrid approach was tested in a multi-speaker, limited vocabulary recognition scenario. The vocabulary was available from the SR4X sample speech corpus [7] with the speech recorded on four different channels of 5 speakers repeating nine words. The speech was recorded at a sampling frequency of 8KHz and stored at 16 bits per sample. The mixed speech signals are generated from original speech signals by multiplying with random square mixing matrix. These mixed signals are used as testing set for speaker recognition. The FastICA is used for the speech separation task. MFCC-VQ is used for feature extraction and matching.

The implementation of the MFCC algorithm was based on the following setup:

a) The speech signal is blocked into frames of N samples, with adjacent frames being separated by M (M < N) samples and with an overlap of N – M samples. Typical values used for N and M are 256 and 100 respectively, with an 8KHz sampling rate.
b) A hamming window function is used to minimize spectral distortion.
c) A standard FFT routine is used to convert the signals into frequency domain.
d) The approximation used to compute the mel for a given frequency *f* in Hz is:

$$\text{mel}(f) = 2595 * \log_{10}(1 + f/700)$$

The number of mel spectrum coefficients, K, is chosen as 20.
e) A set of MFCC, called the acoustic vector, is calculated from the log mel spectrum for each speech frame of around 30msec with overlap.

The results of the MFCC-VQ approach for five speakers without using FastICA are shown in Table 1. The training set consists of 31 sound files and the testing set

Table 1. Results of MFCC VQ without using FastICA

Actual Speaker	Training Words	No. of Samples	Testing Words	Identified Speaker
1	71523	4	abracadabra	1
	Computer	4	generation	1
	Nebula	4	processing	1
2	Supernova	4	sungeeta	2
			tektronix	2
3	abracadabra	4	71523	3
			computer	3
4	Sungeeta	3	supernova	4
	Tektronix	4	tektronix	4
5	71523	4	71523	5
			abracadabra	1

Table 2. Comparison of recognition accuracy

No. of Speakers	Recognition Accuracy (%)	
	Direct Approach	Hybrid Approach
2	50	100
3	33 to 66	66 to 100
4	25 to 50	75
5	20 to 40	60

comprises of 11 sound files. During the training and the testing session, the average number of samples per sound file was approximately 22158 and 30714 respectively. After MFCC processing, the dimension for 31 generated set of acoustic vectors for each sound file was 20 × Number of Frames. For each sound file, these acoustic vectors are transformed into the codebook of size 20 × 16 using VQ algorithm.

The performance of the proposed hybrid (FastICA followed by MFCC-VQ) algorithm is compared with that of a direct approach as discussed in Section 3. The comparison result is shown in Table 2. It is observed that the hybrid approach outperformed the direct approach in all cases. Even with a small number of speakers, the accuracy of the direct approach is poor. An overall improvement of 43% in the recognition accuracy is noted in case of the proposed hybrid approach.

5 Conclusion

The study of MFCC – VQ algorithm shows that in a multi-speaker environment, the task of reliable speaker recognition using conventional techniques becomes difficult. However, one can exploit the state of the art of a single speaker environment in a multi-speaker scenario by using the technique of speech separation. The proposed hybrid approach eventually combines the technique of FastICA for speech separation and MFCC-VQ in multi-speaker environment. A fair degree of improvement in recognition accuracy is achieved. The investigation may be further extended to measure the performance of the hybrid approach in noisy environments.

References

1. S. Furui, “An overview of speaker recognition technology”, ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp. 1-9, 1994.
2. Andre J.W.van der Kouwe, DeLiang Wang, Guy J. Brown, “A comparison of auditory and blind separation techniques for speech segregation”, IEEE Transaction of Speech and Audio Processing, Vol.9, No.3, pp. 189-195.
3. J. F. Cardoso, “Eigenstructure of the 4th-order cumulant tensor with application to the blind source separation problem”, Proc. ICASSP ’89, pp.2109-2112, 1989.
4. A. Bell, T. Sejnowski, “An information maximization approach to blind separation and blind deconvolution”, Neural Computation, vol.7, pp.1129-1159, 1995.
5. A. Hyvärinen, E. Oja, “A fast fixed-point algorithm for Independent Component Analysis”, Neural Computation, vol. 9, no. 7, pp. 1483-1492, 1997.
6. Y. Linde, A. Buzo, R. Gray, “An algorithm for vector quantizer design”, IEEE Transactions on Communications, Vol. 28, pp.84-95, 1980.
7. http://cslu.cse.ogi.edu/corpora/sr4x/