Detection of Speaker Identities from Cochannel Speech Signal

PALLAVI INGALE[1], SANJAY NALBALWAR[2]
Department of Electronics and Telecommunication Engineering[1, 2]
Dr. Babasaheb Ambedkar Technological University[1, 2]
Lonere[1, 2]
INDIA[1, 2]
pallaviing@gmail.com[1], nalbalwar_sanjayan@yahoo.com[2]

Abstract: - Supervised speech segregation for cochannel speech signal can be made easier if we use predetermined speaker's models instead of taking models for all the population. Here we propose a signal to signal ratio (SSR) independent method to detect speaker identities from a cochannel speech signal with unique speaker specific features for speaker identification. Proposed Kekre's Transform Cepstral Coefficient (KTCC) features are the robust acoustic features for speaker identification. A text independent speaker identification system is utilized for identifying speakers in short segments of test signal. Gaussian mixture modeling (GMM) classifier is used for the identification task. We compare the proposed method with a system utilizing conventional features called Mel Frequency Cepstral Coefficient (MFCC) features. Spontaneous speech utterances from candidates are taken for experimentation instead of utterances that follow a command like structure with a unique grammatical structure and have a limited word list in speech separation challenge (SSC) corpus. Identification is performed on short segments of the cochannel mixture. Two Speakers who have been identified for most of segments of the cochannel mixture are selected as two speakers detected for the same cochannel mixture. Average speaker detection accuracy of 93.56% is achieved in case of two speaker cochannel mixture for of KTCC features. This method produces best results for cochannel speaker identification even being text independent. Speaker identification performance is also checked for various test segment lengths. KTCC features outperform in speaker identification task even the length of speech segment is very short.

Key-Words: - Detection of speaker identities, text independent speaker identification, cochannel speech, KTCC.

1 Introduction

Detecting Speaker Identities in cochannel speech signal, is a task of identifying the two speakers present in the given test signal. In Speech segregation, one have to obtain clean speech signal from the given mixture signal [1]. In broadcast news or meeting recordings we want to obtain clean speech signal of any single speaker. When the given mixture signal is a cochannel speech where at least two speakers are present, the segregation becomes harder as the intruder or the noise is another person talking in the vicinity [2]. Speech segregation can be supervised or unsupervised. In supervised speech segregation, speaker specific models are created and stored. Later at the time of test, pre-trained models can be used for speech segregation. But as the speaker population go on increasing speaker distortion or confusion occurs. This will not happen if identities of active speakers are detected in advance. Many supervised speech separation systems like [3, 4] assume the speaker identities and go for denoising.

Environment optimized algorithms of speech segregation perform segregation that can be speaker and/or masker dependent [5]. The same concept can be applied for cochannel speech separation. Motivated with this, here we propose a method to detect speaker identities from a cochannel speech signal that can further be used for speech segregation. Such *single channel speech separation* systems were designed in [6-8] which use speaker identities by performing speaker identification as a first step and then go for speech separation. System proposed in [6] is originally a speech segregation and robust speech recognition system, which performs speaker identification subtask in cochannel condition. They mention the results of cochannel speaker identification for identifying target speaker and identifying both speakers using speech separation challenge (SSC) corpus [10]. Identifying both speakers is important in cochannel speaker identification. *Iroquois* system [7] trains speaker models on gain normalized speech features. It uses gain estimation algorithm and model based analysis to narrow down the speakers list. Signal to signal

ratio (SSR) from -9 dB to 6 dB with an interval of 3 dB is used in above mentioned systems. An SSR dependent method is presented in [8], which use all SSR levels of speech to train models using Gaussian mixture modeling (GMM) [9] technique. Performance of this system is also examined on the SSC corpus. They use the conventional MFCC feature extraction technique. They follow an algorithm to find two speaker identities from the top-three scoring speaker list using the SSR levels for combination of speakers. A Deep Neural Network (DNN) based approach for cochannel speaker identification is presented in [11]. DNN is trained with cochannel training data for each target to interferer ratio (TIR) in anechoic and reverberant conditions.

We can observe two things while going through literature survey for cochannel speaker identification. First is that the database used in most of the papers is having fixed text and unique grammatical structure. Database should contain natural speech from the speakers and have large vocabulary for being able to be used for text independent speaker identification. Second thing is that training is done on cochannel speech for all the available speaker pairs in all SSRs or TIRs. This kind of training is very particular and time consuming.

Here we propose a method to detect speaker identities from a cochannel speech signal with unique speaker specific features called Kekre's Transform Cepstral Coefficient (KTCC) for speaker identification. Proposed KTCC features are robust acoustic features for speaker identification because acoustic features represent vocal track information. Acoustic features perform better than prosodic and other types of features [12].

SSC corpus is specially designed for speech separation task. Here we use a database which contains spontaneous speech from the speakers. This makes the identification task text independent which is more challenging. We break the test utterance into very small segments. Speaker identification is done for these small segments of test signal. A text independent speaker identification system is utilized for identification process. Two Speakers that have been identified for most of segments of the cochannel speech are selected as two speakers detected for the same cochannel speech test signal. Gaussian mixture modeling (GMM) classifier is used for the identification task.

The remainder of the paper is presented in following order. Section 2 provides description of the proposed system for detecting speaker identities. This section also includes the procedure to extract Kekre's Transform Cepstral Coefficient (KTCC) features from the speech signal. Experiments and results are presented in section 3.

2 Proposed System

As it is mentioned earlier that detection of the speaker identities involves identification process for short segments of the test speech signal. Here, we consider that the test speech signal (cochannel speech signal) contains two speakers i.e. one is target speaker and other is intruder. Our aim is to find identities of two speakers present in the cochannel speech. We accumulate the speaker identities for all the segments of the test speech signal and find the two speakers that are identified for most of the segments. Basic block diagram of the proposed system is shown in Figure 1. The blocks are explained subsequently in this section.

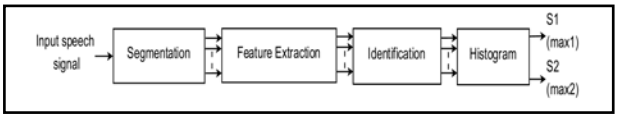

Fig. 1 : Basic block diagram of the proposed system

Figure 2 shows cochannel speech signal constructed with two single speaker speech signals. Identification process is carried out for very short interval segments of the test cochannel speech signal. Each segment is tested to find out the active speaker in that segment.

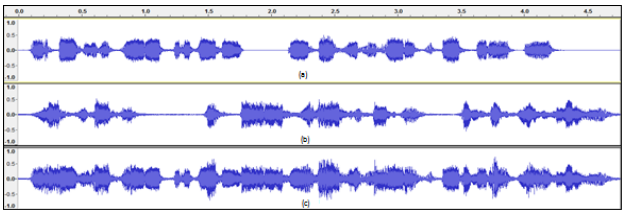

Fig. 2 : (a) and (b) shows single speaker speech signals of two different speakers,(c) shows the mixture of signals (a) and (b) i.e. cochannel speech.

Considering the short segments of speech is important because we can say that the segment may be having a single active speaker, silence or overlapped speech. If we take larger segments then each and every segment will be having overlapped speech. Silent parts of the test signal are removed by performing voice activity detection. Identification process is explained in the next subsection.

Figure 3 explains the segmentation of the speech signal. There are some challenges in the identification process. Here, identification is carried out for very short duration segments of the

cochannel speech. As the duration of segment goes on decreasing, the identification accuracy also decreases. Overlapped speech is as good as noisy speech. Identification becomes harder for such overlapped speech segments. So we need robust features that will produce accurate results. Speaker identification for single speaker speech segments is achieved accurately.

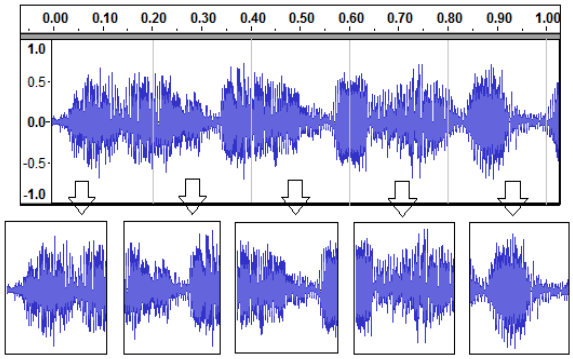

Fig. 3: Segmentation of the speech signal

In most of the overlapped speech segments, a single speaker is dominant. This happens because of different onsets/offsets of the utterance, changing intensity of voice and different speaking styles of the different speakers. The speaker is said to be dominant if its speaker specific acoustic features are retained. Many such segments are found with single dominant speaker. When both speakers are dominant, then identification goes wrong. Generally, these kinds of segments are less in count. Now we will go through major steps involved in the process.

2.1 Front end Processing

In front end processing, speech and silence parts are separated. This is called as Voice Activity Detection (VAD). We remove silent part and take the speech part for further processing. This is a very important step because it avoids unnecessary modelling of background environment. An energy based voice activity detector [15] is used for the same. Signal is broken into frames of same duration. Energy is calculated for each frame and then it is compared with a reference value. The frames having energy value below the reference value are discarded.

2.2 Feature Extraction: Kekre's Transform Cepstral Coefficient (KTCC) Features

Speech signal is a non-stationary signal. But if we assume very short duration of the speech signal, these can be considered to be stationary. That is why we frame the signal into 20ms frames for extracting features. Such a short time frames are formed for the given speech signal with an overlap of 50%. These frames are arranged column wise to form a matrix. Discrete Fourier Transform (DFT) is applied on the columns of this matrix. Here, we obtain the spectrogram of the given speech signal. Squared magnitudes of this spectrogram are considered for further processing.

Spectrogram gives three dimensional information i.e. time, frequency and amplitude. In spectrogram, we get a very wide range of amplitude levels (very small values accompanied with very high values).While processing these wide ranges of values; generally we will not be able to produce faithful results. Because when these values are taken linearly (as it is.), the larger values will affect the output more and smaller values will affect the output less. But, instead of this, if we first apply log to the spectrogram values, then smaller values are emphasized i.e. given more importance. This is somehow similar to the human ear behaviour; which is giving more response to lower frequencies compared to the higher frequencies. By taking log, the range of values of result will be shortened. It becomes more manageable. So, log is applied on the spectrogram of the speech signal.

We use Kekre's transform [13] here, which has been used for various applications in image processing. Kekre's Transform matrix (K) can be of any size NxN, which need not have to be in powers of 2. All upper diagonal and diagonal values of Kekre's transform matrix are one, while the lower diagonal part except the values just below diagonal are zero. Generalized N×N Kekre's Transform Matrix can be given as in (1).

$$K_{N\times N} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 & 1 \\ -\mathrm{N}+1 & 1 & 1 & \cdots & 1 & 1 \\ 0 & -\mathrm{N}+2 & 1 & \cdots & 1 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 1 \\ 0 & 0 & 0 & \cdots & -\mathrm{N}+2 & 1 \end{bmatrix} \quad (1)$$

The formula for generating the term K_{xy} of Kekre's transform matrix is given by (2).

$$K_{xy} = \begin{cases} 1, & x \le y \\ -N + (x-1), & x = y+1 \\ 0, & x > y+1 \end{cases} \quad (2)$$

Kekre's transform on a column vector f is given by,

$$F = [K]f \quad (3)$$

Kekre's Transform is applied on each column the log magnitude squared spectrogram of the speech

signal. This transform matrix forms the KTCC features for the speech signal. MFCC features are widely used for speaker identification task. The first KTCC feature represents energy i.e. it takes all the frequencies. The second feature emphasizes all the frequencies except the first, which is the highest frequency. The third feature emphasizes all the frequencies except first and second high frequencies. In the same way, last feature consider only the lowest frequency. Lower frequencies are very important for human hearing. -N + (x-1) factor is for normalization. Here, we can see the all the frequencies are considered from the spectrogram, but in every feature, lower frequencies are given more importance. Because of this, the KTCC features are good at representing the speaker specific characteristics. Figure 4 shows the block diagram for extracting KTCC features. First, KTCC features for different speakers are calculated and then these features are used to create models for every speaker in the database.

MFCC features are used in the baseline system [8] for detecting identities of the speakers from cochannel speech. MFCC features are very well known features for speaker identification. To obtain MFCC features, short time frames of 20 ms duration of the speech signal are taken. For each short time frame a spectrum is obtained using FFT. Spectrum is passed through Mel-filters to obtain Mel-spectrum. These filters are non-uniformly spaced on the frequency axis i.e. more filters in the low frequency regions and less number of filters in high frequency region. Cepstral analysis is performed on Mel-Spectrum to obtain Mel-Frequency Cepstral Coefficients. Identification performance of MFCC features is compared with Kekre's Transform Cepstral Coefficient (KTCC) features.

Fig.4: Extracting Kekre's Transform Cepstral Coefficient (KTCC) features from speech signal

2.3 Speaker Identification

Text independent speaker identification must be done here as fixed phrases are not taken into consideration. Speaker specific acoustic features are used to create models for the corresponding speakers. Gaussian mixture modelling is conducted for creating models. In [9, 15], benefits of using the Gaussian mixture density for speaker identification are mentioned. First, the individual component Gaussian in a speaker-dependent GMM is interpreted to represent some broad acoustic classes. These acoustic classes reflect some general speaker-dependent vocal tract configurations that are useful for modelling speaker identity. Second, a Gaussian mixture density is shown to provide a smooth approximation to the underlying long-term sample distribution of observations obtained from utterances by a given speaker. Speaker models are represented with λ_i. probability density function is given by the equation,

$$p(\mathrm{x}/\lambda) = \sum_{i=1}^{M} w_i \, g(\mathrm{x}/\mu_\mathrm{i}, \Sigma_\mathrm{i}) \qquad (4)$$

where x is a D-dimensional continuous-valued data vector (i.e. measurement or features), w_i, i = 1, . . . , M, are the mixture weights, and $g(\mathrm{x}/\mu_\mathrm{i}, \Sigma_\mathrm{i})$, i = 1, . . . , M, are the component Gaussian densities. In enrolment phase, models are created for each speaker. These models are used for reference. For speaker identification, a group of S speakers S = (1, 2,. . . , s} is represented by GMM's $\lambda_1, \lambda_2, \lambda_3 \ldots . \lambda_\mathrm{s}$. The objective is to find the speaker model which has the maximum a posteriori probability for a given observation sequence. Identity of speaker $\hat{s}$ is determined by using the following equation.

$$\hat{s} = \arg \max_{1 \le k \le S} = \sum_{t=1}^{T} log \, p(\vec{\mathrm{x}}/\lambda_\mathrm{k}). \qquad (5)$$

When we are detecting speakers from a cochannel speech, we follow the identification procedure for all segments.

2.4 Detection of Speaker Identities

Here we propose a method to detect speaker identities from a cochannel speech signal utilizing KTCC features. We form short time segments of the cochannel speech signal as shown in figure 3. KTCC features are extracted using the procedure mentioned in subsection 2.2, for the identification process. Identification is carried out for these short segments as described in the subsection 2.3. We accumulate the speaker identities by histogram for all the segments of the cochannel speech signal and find the two speakers that are identified for most of the segments. Two speakers who have been identified for most of segments are selected as two speakers detected for the same cochannel speech signal.

3 Experiments and Results

We perform two kinds of experiments. First experiment is of checking the effect test segment duration for speaker identification in case of clean speech signals. This is for evaluating the performance of MFCC and KTCC features for identification for minimum test segment duration. Test segment of cochannel speech should be very small, so that we can consider the dominance of a single speaker in that segment. Performance of detecting speaker identities in cochannel speech mainly depends on the accuracy of identification i.e. on discriminative nature of features. In the second experiment, we detect identities of the two speakers from cochannel speech and compare the performance of MFCC and KTCC features for the same.

Previous systems [6, 7, 8] are text not text independent. Proposed system is text independent. System explained in [8], use all SSR levels of cochannel speech to train models using GMM technique. A performance of this system is examined on the SSC corpus. They use conventional MFCC feature extraction technique. They follow an algorithm to find two speaker identities from the top-three scoring speaker list using the SSR levels for combination of speakers. This system is proven to be better than the Iroquois [7] system. We consider [8] as the baseline system i.e. a system which uses MFCC feature and GMM technique is used as classifier for the identification task without the SSR selection algorithm.

As we have mentioned earlier in this paper, SSR dependency needs the system to be trained with cochannel speech of all possible pairs for all SSR or TMR ratios. This creates a huge burden on the system at the time of training. And as the number of speakers enrolled in the system goes on increasing, it becomes a tedious task. Finding the SSR ratios at the time of test also increases the computational complexity. Therefore, we perform experimentation for SSR independent case. We train the models for the speakers using their clean speech only, and not the cochannel speech. Due to this, the whole training procedure becomes faster.

We can observe in natural speech that the intensity and energy in an utterance is never the same for the complete utterance. It always changes its level. So when we consider cochannel speech, the ratio between the energy of the two speakers involved is not the same throughout the utterance. So SSR is always changing, naturally. So there is no need to change it externally in the experiments. In cochannel speaker identification, we are getting benefit of this naturally changing SSR in identifying the dominant or active speaker in the test segment.

3.1 Database

Most of the systems use SSC database which is originally recorded for speech separation task and not for speaker identification. We use Hindi speech database[1], in which spontaneous speech from speakers is recorded from Indian national Hindi news channels. We have 31 speakers in the database of which 15 are female and 16 are male speakers. These spontaneous speech utterances from speakers are taken for experimentation instead of utterances that follow a command like structure with a unique grammatical structure and have a limited word list in SSC corpus. In SSC corpus each sentence is formed by a unique structure like “command, color, letter, number and code”. So it has limited word list.

We created mixtures i.e. cochannel speech signals by mixing speech signal of two different speakers and of same talker. We have 240 different gender, 256 same gender and 100 same talker test signals. Train and test speech sets are different. GMM classifier is used for identification process. A speaker model is created by using its clean speech utterance only.

3.2 Effect of varying test segment duration on identification accuracy for clean speech signals

First, we examine the performance of both of the features for speaker identification on varying lengths of test segments. This is done for evaluating the performance of MFCC and KTCC features for identification for the minimum test segment duration. Speaker identification is done for the clean speech signals.

We tested all the features for a maximum 10 seco,nds to a minimum duration of duration 0.2 second duration of test signal. For test signal duration of 10 seconds, both MFCC and KTCC have good performance. As the test signal duration goes on decreasing, accuracy goes on decreasing. KTCC features produce good performance over MFCC features even for minimum test signal duration i.e. 0.2 second. Effect of varying test signal duration on identification accuracy is shown in figure 5. This shows that KTCC features will also produce good performance for cochannel identification.

[1]The Hindi speech database can be downloaded at https://drive.google.com/open?id=0B4usACU21mFoMkJ3ZVRZYlFJWGs

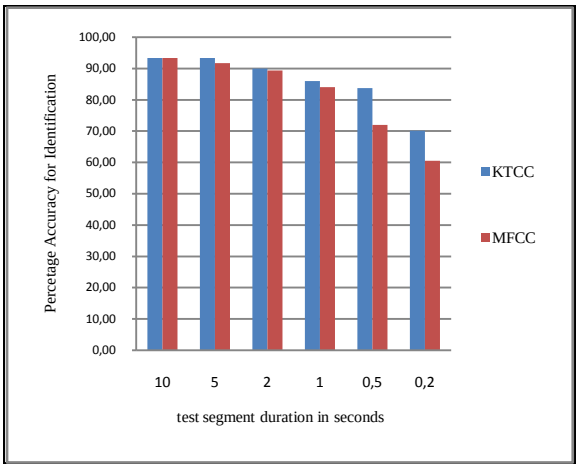

Fig.5: Effect of varying test segment duration on identification accuracy.

3.3 Detection of Speaker Identities from Cochannel Speech

For the detection of identities of speakers from cochannel mixture, we formed segments of the test cochannel speech signal of 0.2 second duration segments. If we have a 5 second long test cochannel speech signal, then 25 non overlapping segments can be obtained. Identification is carried out for all the segments. We trained GMM with 128 Gaussian component densities. We accumulate the speaker identities for all the segments of the test speech signal and find the two speakers that are identified for most of the segments. This is done through plotting histogram as shown in figure 6. It is the histogram plot for cochannel speech signal of 'speaker 1' and 'speaker 15' in the database.

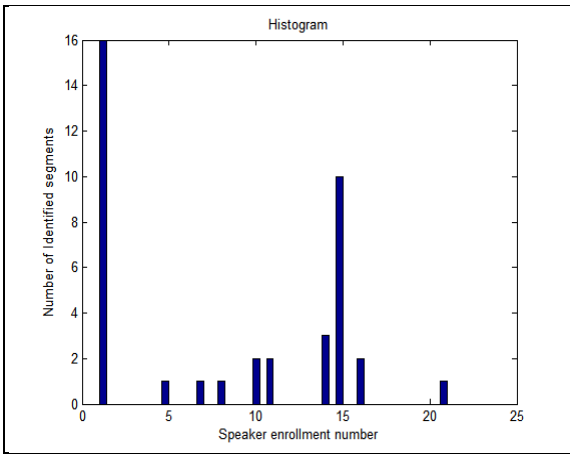

Fig. 6: Histogram plot of identification result for all the segments in cochannel test signal.

We can see from the plot that 'speaker 1' is found over total 16 segments and 'speaker 10' is found over total 10 segments of the cochannel speech signal. For these 26 segments identification is done accurately. So we choose 'speaker 1' and 'speaker 15' as the detected speakers for the given cochannel speech signal.

For performance evaluation, the parameter used is Identification accuracy in percentage. The accuracy of the identification [14] system is calculated as given by equation (6).

$$\text{Percetage of Identification accuracy} = \frac{\text{No.of test segments for which speakers are correctly Identified}}{\text{Total no.of test segments}} \times 100 \quad (6)$$

Table I presents the comparative results for the baseline system using MFCC features and our proposed system using KTCC features. As we have taken the cochannel mixtures of same gender speakers, different gender speakers and same talker; we refer them as SG, DG and ST respectively.

Our proposed system of detecting speaker identities from cochannel speech mixtures achieved average accuracy of 93.56% in detecting speakers correctly in cochannel speech signals utilizing KTCC features. Results for baseline system using MFCC features are calculated here without using SSR algorithm. Baseline system with SSR selection algorithm produces identification accuracy about 97%. But our system reduces training burden as well as computational complexity at the time of test. Considering D speakers M Gaussians and G SSR levels, the numbers of Gaussian evaluations for baseline system are *O(DGM)*. Our proposed system is SSR independent, so computational complexity is reduced to *O(DM)*. Hence, the proposed system is faster than the baseline system.

Table I

Cochannel Mixture Condition	Baseline system using MFCC Features	Proposed system using KTCC Features
SG	58.59	89.84
DG	70.83	90.83
ST	99.00	100.00

4 Conclusion

We have presented a novel system to detect speaker identities from a cochannel speech signal with unique speaker specific features for speaker identification. KTCC features are the robust acoustic features for speaker identification. Identification performance is checked over various durations of the test signal, as it is required to take

very small segments while detecting speaker identities for cochannel speech. We proved through experiments that KTCC features perform better than MFCC features even when the duration of the test signal is very small. Speaker identification accuracy of about 70 % is achieved for 0.2 second duration of the test signal.

Our proposed system to detect speaker identities from cochannel speech mixtures achieved average accuracy of 93.56% in detecting speakers correctly in cochannel speech signals, utilizing KTCC features. Additionally, this method is text independent and computationally efficient compared to the baseline system. Baseline systems used speech mixed at various SSR levels. That needed a tremendous amount of training data. This is not required in our system, as our system is SSR independent. We can go for speech segregation with these detected speakers and supervised speech segregation can be made easy.

References:

[1] W. Yu, L. Jiajun, C. Ning, and Y. Wenhao, Improved monaural speech segregation based on computational auditory scene analysis, *EURASIP Journal on Audio, Speech, and Music Processing*, Vol. 2013, No.2, 2013, pp. 1-15.

[2] H, Ke, and D. Wang, An iterative model-based approach to cochannel speech separation, *EURASIP Journal on Audio, Speech, and Music Processing*, Vol. 2013, No.1 2013, pp. 1-11.

[3] Y. Wang, and D. Wang, A structure-preserving training target for supervised speech separation, 2014 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 6107-6111.

[4] A. Reddy and B. Raj, Soft mask methods for single-channel speaker separation, *IEEE Transactions on Audio, Speech and Language Processing*, Vol.15, No.6, 2007, pp. 1766-1776.

[5] K, Gibak, Y. Lu, Y. Hu, and P. Loizou, An algorithm that improves speech intelligibility in noise for normal-hearing listeners, *The Journal of the Acoustical Society of America*, Vol. 126, No.3, 2009, pp. 1486-1494.

[6] Y. Shao, S. Srinivasan, Z. Jin, and D. Wang, A computational auditory scene analysis system for speech segregation and robust speech recognition, *Computer Speech & Language*, Vol.24, No.1, 2010, pp. 77-93.

[7] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T.T. Kristjansson, Super-human multi-talker speech recognition: A graphical modeling approach, *Computer Speech & Language*, Vol.24, No.1, 2010, pp. 45-66.

[8] P. Mowlaee, R. Saeidi, M. G. Christensen, Z. H. Tan, T. Kinnunen, P. Franti, and S. H. Jensen, A joint approach for single-channel speaker identification and speech separation, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol.20, No.9, 2012, pp. 2586-2601.

[9] D. A. Reynolds, Speaker identification and verification using Gaussian mixture speaker models, *Speech communication*, Vol.17, No.1, 1995, pp. 91-108.

[10] M. Cooke, J. R. Hershey, and S. J. Rennie, Monaural speech separation and recognition challenge, *Computer Speech & Language*, Vol. 24. No.1, 2010, pp. 1-15.

[11] X. Zhao, Y. Wang, and D. Wang, Deep neural networks for cochannel speaker identification, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4824-4828.

[12] J. M. Naik, Speaker Verification: A Tutorial, *IEEE Communications Magazine*, Vol. 28. No.1, 1990. pp. 42-28.

[13] H. B. Kekre, S. D. Thepade, and A. Maloo, Performance Comparison of Image Retrieval Using Fractional Coefficients of Transformed Image Using DCT, Walsh, Haar and Kekre's Transform, *CSC-International Journal of Image processing (IJIP)*, Vol.4, No.2, 2010, pp. 142-155.

[14] D. A. Reynolds, and R. C. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models, *IEEE transactions on speech and audio processing*, Vol.3, No.1, 1995, pp. 72-83.

[15] T. Giannakopoulos, A. Pikrakis, *Introduction to Audio Analysis: A MATLAB® Approach*. Academic Press, 2014.