Robust Speaker Identification using Denoised Wave Atom and GMM

Mohammed Alhanjouri
Assoc. Prof. in Computer Engineering Dept., Islamic University of Gaza (IUG)
Gaza, Palestine

Mohammed A. H. Lubbad
Computer Engineering Dept., Islamic University of Gaza (IUG)
Gaza, Palestine

Mahmoud Z. Alkurdi
Computer Engineering Dept., Islamic University of Gaza (IUG)
Gaza, Palestine

ABSTRACT

This paper introduces the use of Wave atom transformation as an efficient speech noise filter with Gaussian mixture models (GMM) for robust text-independent speaker identification. The individual Gaussian components of a GMM are shown to represent some general speaker identity. The focus of this work is on applications which require high robustness of noise and high identification rates using short utterance from noisy (Natural Noise) numerical speech and alphabetical words speech. A Full experimental evaluation of the Gaussian mixture speaker model is conducted on a 10 speakers. The experiments examine algorithmic issues (Preprocessing (Denoising by Wave Atom), Feature Extraction (MFCC), Training using GMM, Pattern Matching (Maximum likelihood estimation ML), Decision Rule (Expectation Maximization EM)). The Proposed algorithm attains 95% identification accuracy using 5 seconds noisy speech utterances without Wave atom preprocessing it attains 90% identification accuracy using 5 seconds noisy speech utterances. Proposed denoisy algorithm increases the identification ratio by 5% for noisy speech signals, this ratio is interesting enough.

Keywords

Wave Atom Transformation, MFCC, Gaussian Mixture Model GMM, Wavelet Transformation, Speaker recognition.

1. INTRODUCTION

Speech processing is the study of speech signals and the processing methods of these signals, speech processing can be divided into the many categories like Speech recognition, which deals with analysis of the linguistic content of a speech signal, Speaker recognition, where the aim is to recognize the identity of the speaker, Speech coding, a specialized form of data compression, is important in the telecommunication area, Voice analysis for medical purposes, such as analysis of vocal loading and dysfunction of the vocal cords, and Speech denoise: enhancing the intelligibility and/or perceptual quality of a speech signal, like audio noise reduction for audio signals. In this paper we will deal with two kinds of speech processing which are speaker recognition(Identification) and speech denoise, The speech signal consists of several levels of information, it conveys them to the listener. The primary task of the speech signal is to conveys the words or message being spoken, but on other level, the signal also conveys information about identity of the speaker [1]. While many existing systems for speaker identification achieve good performance in relatively constrained environments, performance invariably deteriorates in noisier environment. Speaker identification system is the process of selecting the best matched speaker among the enrolled speakers, with features extracted from speech signals [2]. Many techniques involving statistical or probabilistic approaches have been applied to speaker specific speech patterns (Leena Mary and Yegnanarayana (2008), Jyoti et al (2011)) [3] [4]. Several methods were employed to separate mixed signals known as 'Blind Source Signals' (BSS) [5]. The term blind refers to the fact that the method of combination and source signal characteristics are unknown, so BSS permits a wide range of signals as input.

Text independent speaker identification system has many potential applications like security control, telephone banking, information retrieval systems, speech and gender recognition systems, etc. Speaker identification system involves two parts: front-end (feature extractions) and back-end (actual recognition). These system use processed form of speech signals instead of using raw speech signals as it is obtained. This is to reduce the time consumed in identifying the speaker and to make the process easy, by reducing the data stream and exploiting its advantage of being redundant. Computation of cepstral coefficients using preprocessing and feature extraction phases plays a major role in text independent speaker identification systems Ning Wang et al (2010) [6].

During transmission reception and recording signals are often corrupted by noise which can cause severe problems for downstream processing and user perception, Speech denoise aims to improve speech quality by using various algorithms, All the speech denoise methods aimed at suppressing the background noise are (naturally) based in one way or the other on the estimation of the background noise. If the background noise is evolving more slowly than the speech, i.e., if the noise is more stationary than the speech, it is easy to estimate the noise during the pauses in speech. Finding the pauses in speech is based on checking how close the estimate of the background noise is to the signal in the current window. Voiced sections can be located by estimating the fundamental frequency. Both methods easily fail on unstressed unvoiced or short phonemes, taking them as background noise. On the other hand, this is not very dangerous because the effect of these faint phonemes on the background noise estimate is not that critical. Therefore an automated means of removing the noise would be an invaluable first stage for many signal processing tasks. Denoising has long been a focus of research

Simple methods originally employed the use of time-domain filtering of the corrupted signal [7]; however, this is only successful when removing high frequency noise from low frequency signals and does not provide satisfactory results under real world conditions. To improve performance, modern algorithms filter signals in some transform domain such as z

or Fourier. Over the past two decades, a flurry of activity has involved the use of the wavelet transform after the community recognized the possibility that this could be used as an superior alternative to Fourier analysis [8]. Numerous signal and image processing techniques have since been developed to leverage the power of wavelets. These techniques include the discrete wavelet transform, wavelet packet analysis, and most recently, wave atom analysis. Wave atoms are a recent addition to the collection of mathematical transforms for harmonic computational analysis. Wave atoms are a variant of 2D wavelet packets that retain an isotropic aspect ratio. Wave atoms have a sharp frequency localization that cannot be achieved using a filter bank based on wavelet packets and offer a significantly sparser expansion for oscillatory functions than wavelets, curvelets and Gabor atoms. Wave atoms capture coherence of pattern across and along oscillations whereas curvelets capture coherence only along oscillations. Wave atoms precisely interpolate between Gabor atoms [9] (constant support) and directional wavelets [10] (wavelength ~ diameter) in the sense that the period of oscillations of each wave packet. (Wavelength) related to the size of essential support by the parabolic scaling i.e. wavelength ~ (diameter)^2 [11]. In this paper we merge between speaker identification and speech denoising since noise in training or testing speech cause bad results and make a big problem so propose as a preprocessing step in speech recognition is to make denoising using wave atom techniques. Section 2 is a brief introduction on GMM and Wave Atom Transformation. Section 3 discusses our Speaker Identification method. Section 4 show the experimental evaluation of the effectiveness and efficiency of our approach using data sets as illustrated. Finally in Section 6, concluding remarks are offered.

2. WAVE ATOM AND GMM

2.1 WAVE ATOM TRANSFORM

Single channel denoising depend on filtering noise by applying single band low filter to speech in order to remove high frequencies which may cause noise, this may be unfair and cause damage to some parts of important data so many studies tries to solve this problem using multi band technique like wavelet and anew part of its family oriented which is called wave atom. , classical wavelet transform, pass from one stage to another, only the approximation will decomposed. In other hand the decomposition in wavelets packets could be pursued into the other sets, which is not optimal. So the optimality is related to the maximum energy of the decomposition. The idea is then to looking for the path yielding to the maximum energy through the different subbands. Wave atom is multiscale transforms for image and numerical analysis. Some fundamentals notions were as following [12].

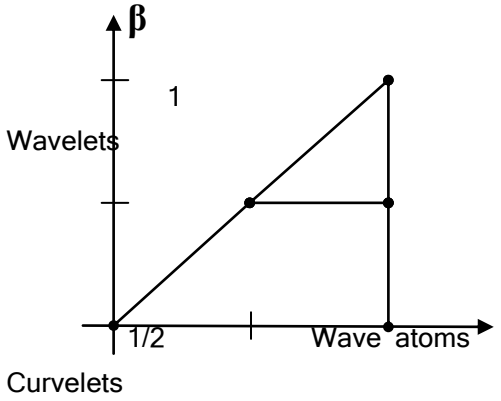
Let us define 2D Fourier transform as:

$$\hat{f}(w) = \int e^{-ixw} f(x)dx \quad [11]$$

$$f(x) = \left(\frac{1}{2\pi}\right)^2 \int e^{ixw} \hat{f}(w)dw \quad [12]$$

Wave atoms are noted as, with subscript. The indexes are integer-valued associated to a point in the phase-space defined as follows:

$x_\mu = 2^{-j}n$,$w_\mu = \pi 2^j m$, $C_1 2_j \le \max_{i=1,2}|m_i| \le C_2 2^j$In [10], they suggest that two parameters are sufficient to index α lot of known wave packet architectures. The index indicates whether the decomposition is multi scale (α = 1) or not (β = 0) ; and β indicates whether basis elements are localized and poorly directional (α = 1) or, on the contrary, extended and fully directional (β = 0).The description in terms of α and β will clarify the connections between various transforms of modern harmonic analysis. Wavelets (including Multi Resolution Analysis, directional and complex) correspond to α = β = 1 , for ridgelets [β] α= 1, β = 0 , Gabor transform α = β = 0 and curvelets correspond to α = β = 1/2. Wave atoms are defined for α = β = 1/2. Figure 1 illustrates classification. In order to introduce the wave atom, let us first consider the 1D case. In practice, wave atoms are constructed from tensor products of adequately chosen 1D wavelet packets. An one-dimensional family of real-valued wave packets $\psi^j_{m,n}(x), j \ge 0, m \ge 0, n \in Z$, centered in frequency around $\pm w_{j,m} = \pm\pi_2 2^j m$ with $C_1 2^j \le m \le C_2 2^j$; and centered in space around $x_{j,n} = 2^{-j}n$, is constructed. The one-dimensional version of the parabolic scaling inform that the support of $\psi^j_{m,n}(w)$ be of length $0(2^{2j})$, while $w_{j,m}(w) = 0(2^{2j})$ [9]. The desired corresponding tiling of frequency is illustrated at the bottom of Figure 2. Filter bank-based wavelet packets is considered as a potential definition of an orthonormal basis satisfying these localization properties. The wavelet packet tree, defining the partitioning of the frequency axis in lD, can be chosen to have depth j when the frequency is 2^{2j}, as illustrated in Figure 2. Figure 2 presents the wavelet packet tree corresponding to wave atoms. More details on wavelet packet trees can be found in [9]. The bottom graph depicts Villemoes wavelet packets on the positive frequency axis. The dot under the axis indicates a frequency where a change of scale occurs. The labels "LH", respectively "RH" indicates a left-handed, respectively right- handed window [10].



In 20 domains, the construction presented above can be modified to suit certain applications in image processing or numerical analysis: The orthobasis variant [9]. In practice, one may want to work with the original orthonormal basis $\varphi^+_\mu(x)$ instead of a tight frame. Since $\varphi^+_\mu(x) = \varphi^1_\mu(x) + \varphi^2_\mu(x)$ each basis function $\varphi^+_\mu(x)$ oscillates in two distinct directions, instead of one. This is called the orthobasis variant.

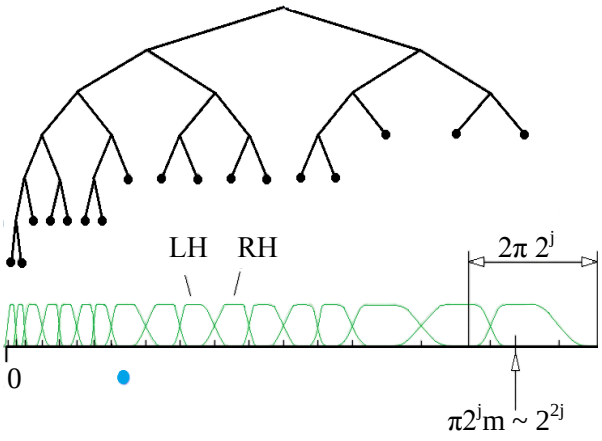


Fig: 2 wavelet packet tree corresponding to wave atoms

2.2 GMM

This section describes the form of the Gaussian mixture model (GMM) and motivates its use as a representation of speaker identity for text-independent speaker identification. The speech analysis for extracting the mel-cepstral feature representation used in this work is presented first. Next, the Gaussian mixture speaker model and its parameterization are described. The use of the Gaussian mixture density for speaker identification is then motivated by two interpretations. First, the individual components Gaussians in a speaker-dependent GMM are interpreted to represent some broad acoustic classes. These acoustic classes reflect some general speaker-dependent vocal tract configurations that are useful for modeling 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. Finally, the maximum-likelihood parameter estimation and speaker identification procedures are described.

A. Speech Analysis

Although there is no exclusively speaker distinguishing speech features, the speech spectrum has been shown to be very effective for speaker identification. This is because the spectrum reflects person's vocal tract structure, the predominant physiological factor which distinguishes one person's Voice from others. LPC spectral representations, such as LPC cepstral and reflection coefficients, have been used extensively for speaker recognition; however, these model-based representations can be severely affected by noise [1]. Recent studies have found directly computed filterbank features to be more robust for noisy speech recognition [1], In this paper we use the cepstral coefficients derived from mel-frequency filterbank to represent the short-time speech spectra.

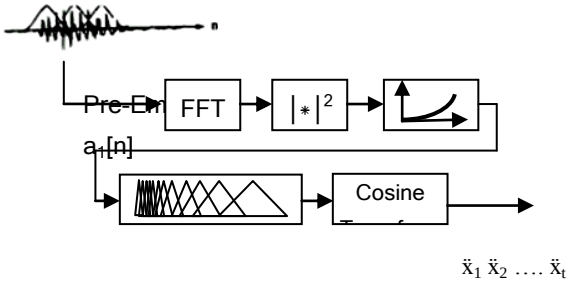


Fig 3. Shows a block diagram of the steps in our frontend feature extraction. The magnitude spectrum from a 20 ms short-time segment of speech is pre-emphasized and processed by a simulated Mel-scale filterbank. The filterbank follows that by described in [1]. The log-energy filter ouputs are then cosine transformed to produce the cepstral coefficients. The zeroth cepstral coefficient is not used in the cepstral feature vector. This processing occurs every 10 ms, producing 100 feature vectors per second.

B. Model Description

A Gaussian mixture density is a weighted sum of M component densities, as depicted in Fig.4 and given by the equation

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \qquad (1)$$

Where $\vec{x}$ is D-dimensional random vector, $b_i(\vec{x})$, i = 1,…,M, are the component densities and p_i, i = 1,…,M, are the mixture weights. Each component density is a D-variate Gaussian function of form

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\sum_i|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2(\vec{x}-\vec{\mu_i})'} \sum_i^{-1} (\vec{x}-\vec{\mu_i})\right\} (2)$$

With mean vector $\vec{\mu_i}$ and covariance matrix $\sum_i$. The mixture weights statisfy the constraint that $\sum_{i=1}^{M} p_i = 1$. The complete Gaussian mixture density is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. Theses paramters are collectively represented by the notation

$$\lambda = \{p_i, \vec{\mu_i}, \varepsilon_i\}\ i = 1, \ldots, M. \qquad (3)$$

For speaker identification, each speaker is represented by a GMM and is referred to by his/her model λ.

The GMM can have several different forms depending on the choice of covariance matrices. The model can have one covariance matrix per Gaussian component as indicated in nodal covariance, one covariance matrix for all Gaussian components in a speaker model (grand covariance), or a single covariance matrix shared by all speaker models (global covariance).

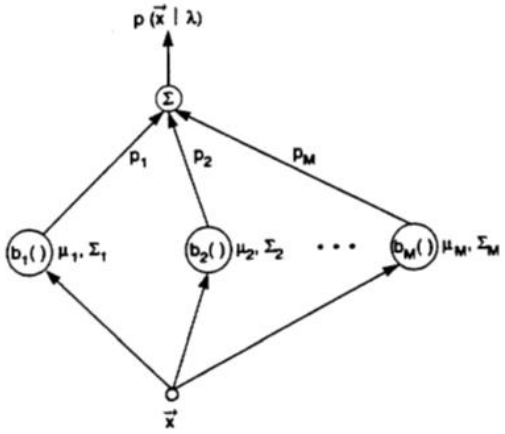


Fig: 4 GMMs for speaker recognition: quoted from [4]

The covariance matrix can also be full or diagonal. In this paper, nodal, diagonal covariance matrices are primarily used for speaker models, except as noted for some experiments. This choice is based on initial experiments are primarily used for speaker models, except as a noted for some excremental results indicating better identification performance using nodal, diagonal variances compared to nodal and grand full covariance matrices.

C. Model Interpretations

There are two principal motivations for using Gaussian mixture densities as representation of speaker identity. The first motivation is the intuitive notion that the individual component densities of a multi-modal density, like the GMM, may model some underlying set of acoustic classes. It is reasonable to assume the acoustic space corresponding to a speaker's voice can be characterized by a set of acoustic classes representing some broad phonetic events, such as vowels, nasal, or fricatives. These acoustic classes reflect some general speaker-dependent vocal tract configurations that are useful for characterizing speaker identity. The spectral shape of the i-th acustic class can in turn be represented by the mean $\vec{\mu_i}$ of the i-th component density, and variations of the average spectral shape can be represented by the covariance matrix ε_i . Because all training or testing speech is unlabeled, the acoustic classes are "hidden" in that the class of an observation is unknown. Assuming independent feature vectors, the observation density of feature vectors drawn from these hidden acoustic classes is a Gaussian mixture.

The second motivation for using Gaussian mixture densities for speaker identification is empirical observation that

A linear combination of Gaussian basis functions is capable of representing a large class of sample distributions. One of the powerful attributes of GMM is its ability to form smooth approximation to arbitrary-shaped densities. The classical unimodal Gaussian speaker model represents a speaker's feature distribution by a position (mean vector) and a elliptic shape (covariance matrix) and VQ model represents a speaker's distribution by a discrete set of characteristic

D. Maximum Likelihood Parameter Estimation

Given training speech from a speaker, the goal of speaker model training is to estimate the parameters of the GMM, λ, which describes the distribution of the training feature vectors. By far the most popular and well-established is Maximum Likelihood (ML) estimation.

These GMMs are trained separately on each speaker's enrollment data using the Expectation Maximization (EM) algorithm [1]. The update equations that guarantee a monotonic increase in the model's likelihood value are:

Mixture Weights:

$$\bar{p}_i = \frac{1}{T}\sum_{t=1}^{T} p(i|\vec{x_t}, \lambda)$$

Means:

$$\vec{\bar{\mu}_i} = \frac{\sum_{t=1}^{T} p(i|\vec{x_t}, \lambda)\vec{x_t}}{\sum_{t=1}^{T} p(i|\vec{x_t}, \lambda)}$$

Variances:

$$\bar{\sigma_i}^2 = \frac{\sum_{t=1}^{T} p(i|\vec{x_t}, \lambda)\vec{x_t}}{\sum_{t=1}^{T} p(i|\vec{x_t}, \lambda)} - \vec{\bar{\mu}_i}^2$$

where σ^2, x_t and μ_i are elements of $\bar{\sigma_i}^2, \vec{x_t}, \vec{\mu_i}$ respectively. The a posteriori probability for acoustic class i is given by,

$$p(i|\vec{x_t}, \lambda) = \frac{p_i b_i(\vec{x})}{\sum_{k=1}^{M} p_k b_k(\vec{x})}$$

In speaker identification, given a group of speakers S = {1, 2....M}, represented by GMMs λ 1, λ 2, λ 3.... λ S, the objective is to find the speaker model which has the maximum a posteriori probability for a given test sequence,

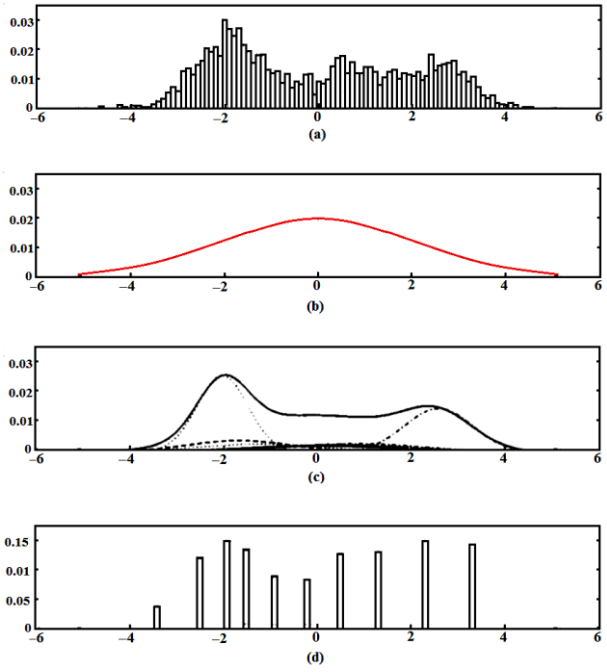


Fig5: Comparison of distribution modeling: (a) Histogram of signal cepstral coefficients from 25 second utterance by Mel speaker. (b) Maximum likelihood unimodal Gaussian model; (c) GMM and 10 underlying component densities;

$$\hat{S} = \arg \max_{1 \le k \le M} p(\lambda_k) = \arg \max_{1 \le k \le M} \frac{P(X'|\lambda_k)P(\lambda_k)}{P(X')}$$

Assuming that all speakers are equally likely and that the observations are independent, and since p(X) is same for all speakers, this simplifies to

$$\hat{S} = \arg \max_{1 \le k \le M} p(\lambda_k) = \arg \max_{1 \le k \le M} [\prod_{t=1}^{T} p(x_t'|\lambda_k)]$$

Each GMM outputs a probability for each frame, which is multiplied across all the frames. The classifier makes a decision based on these product posterior probabilities.

3. PERFORMANCE EVALUATION

This Section will evaluate the performance of proposed speaker identification system to show its affectivity. Tests were done on artificial dataset as shown below, table 1 is a speech audio recorded by natural noise, table 2 is the same but recorder with brown noise, Finally table 3 is data set contains from 30 wav training speech audio signals from different speakers with long duration, 10 were used as training data and 20 of them are used as testing data.

Table1: Audio (natural noise) Signal Propriety

Audio (natural noise) Signal Propriety	
Type	Artificial .wav
Sampling Rate	8000 Hz
Bit depth	16
Channel	Mono
Length	10 sec

Table 2: Audio (Brown noise) Signal Propriety

Audio (Brown noise) Signal Propriety	
Type	Brown .wav
Sampling Rate	8000Hz
Bit depth	16
Channel	Mono
Length	10 sec

Table 3: Training and testing data Propriety

Dataset audio Signal Propriety	
Type	Artificial .wav
Sampling Rate	8000 Hz
Bit depth	16
Channel	Mono
Length	10 sec
Classes	10
Tests	20

3.1 SPEECH DENOISING

Wave atom transformation has been applied to speech audio signal with natural noise shown in table 1.To denoise the audio signal using wave atom transformation, spectrogram of speech signal has been exported before and after applying wave atom transform as shown in figure 6 and figure 7, it is clear that the speech signal was denoised, and high frequencies was filtered, figure 6 is the same but length was taken about 1.2 sec the figure show the effect of using wave atom transformation.

Brown noise has been added to the same signal and apply wave atom transformation to denoise signal figure 8,9 with different coefficient threshold to measure the SNR in each time and to find the best threshold as in figure 10, power SNR was increased from 0.003312 to 0.34 . Choosing best coefficient t threshold changes from signal to signal and there is no equation to find the optimal threshold

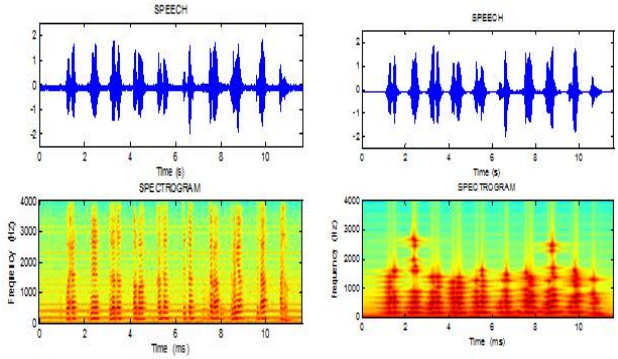


Fig 6: Speech signal after and before applying wave atom transformation.

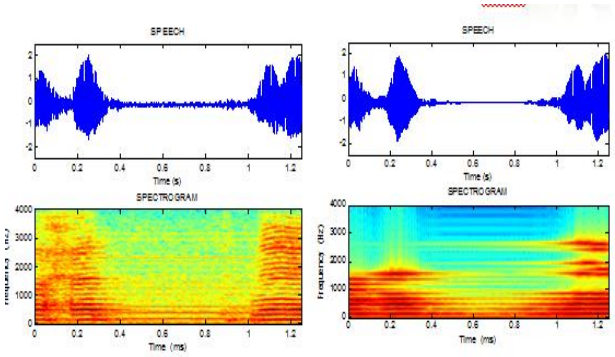


Fig7: Speech signal after and before applying wave atom transformation with 1.2 sec of figure 1

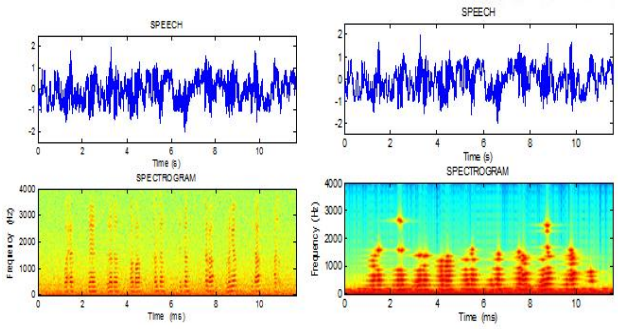


Fig8: Speech noisy signal after and before applying wave atom transformation

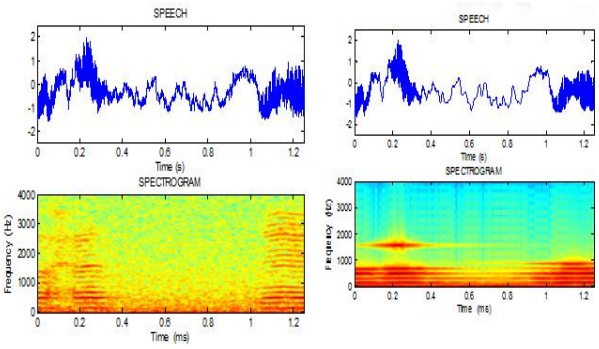


Fig9: Speech noisy signal after and before applying wave atom transformation with 1.2 sec of figure 4

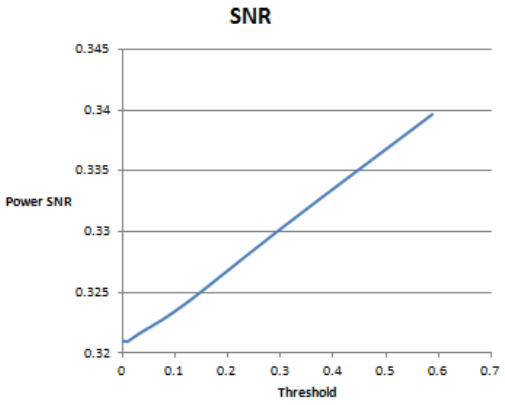


Fig10: SNR with changing in coefficient threshold

3.2 SPEAKER IDENTIFICATION

Dataset in table 3 which contain from 10.wav audio training data each wav file with natural noise is counting form 1 to 10 with identical utterance. And dataset contains 20.wav testing each wav file with natural noise is counting from 1 to 3 and name of utterance. To measure recognition rate of our proposed speaker Identification approach figure 11, the first step was denoising dataset using wave atom transformation then applying speaker identification, Finally measuring recognition rate with and without denoising , the result is recognition rate was 90% before and 95% after using waveatom denosing.

Table 4: Speaker Identification Performance

Speaker Model	% Correct Identification (5 second test length)
GMM-WaveAtom Denoise	95.0±1.5
GMM-nv	94.5±1.8
VQ-100	92.9±2.0
GMM-gv	89.5±2.4
VQ-50	90.7±2.3
RBF	87.2±2.6
TGMM	80.1±3.1
GC	67.1±3.7

Fig11: Our Speaker Identification Approach.

4. CONCLUSION

In this paper, the Wave Atom transformation is presented denoising method for speech audio signals and the method used in speaker identification using GMM. apply wave atom transform on real data set used as training and test the results was good with about 95% recognition rate for our used dataset. Wave atom denoising takes more time but it gives more denoised results,it is insensitive to the order of input data to be processed. Moreover, it offers good result in recognition process.

5. REFERENCES

[1] D. Reynolds, R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models", IEEE Trans. Speech Audio Process., vol. 3, no.1, pp. 72-83, Jan. 1995.

[2] N M Ramaligeswararao, Dr.V Sailaja and Dr.K. Srinivasa Rao," Text Independent Speaker Identification using Integrated Independent Component Analysis with Generalized Gaussian Mixture Model" (IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 2, No. 12, 2011

[3] Leena mary and yegnanaryana(2008), "Extraction and representation of prosodic feature for language and speaker recognition" SPEECH COMMUNICATION 50(10):782-796.Michael Charles (1999), "Orthogonal GMM in Speaker Recognition," Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing", pp. 845-848.

[4] Jyothi et al (2011), "Text independent speaker identification with finite multivariate generalized Gaussian mixture model with distant microphone speech" proceeding of the international journel of computer applications (IJCA)14(4):5-9.

[5] H. Gish et a (1985), "Investigation Of Text-dependent Speaker Identification Over Telephone Channels," in Proc. IEEE ICASSP, pp. 379-382.

[6] Ning Wang , P. C. Ching(2011), "Nengheng Zheng, and Tan Lee, "Robust Speaker Recognition Using Denoised Vocal Source and Vocal Tract Features speaker verification," IEEE Transaction on Audio Speech and Language processing, Vol. 1, No. 2, pp. 25-35.

[7] Nitin Trivedi, Dr. Vikesh Kumar, Saurabh Singh, Sachin Ahuja, Raman Chadha, 2011. Speech Recognition by Wavelet Analysis, International Journal of Computer Applications (0975 – 8887) Volume 15– No.8.

[8] B. Jawerth and W. Sweldens, 1994, "An overview of wavelet based multiresolution analysis," SIAM Review, vol. 36, no. 3, pp. 377–412.

[9] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard, 1995 "Wavelet shrinkage: Asymptotic?," Journal of the Royal Statistics Society, vol. 57, pp. 301–369.

[10] S. Mallat, 1999, A wavelet Tour of Signal Processing, Second Edition, Academic Press, Orlando-SanDiego.

[11] J.P. Antoine and R. Murenzi, 1996, Two-dimensional directional wavelets and the scale-angle representation, Sig. Process., vol. 52, pp. 259-281,

[12] Demanety and L. Ying, 2007. Wave atoms and sparsely of oscillatory patterns, appear in Appl. Comput. Harm. Anal, VoL 23, Issue 3, pp. 368-387.