International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August, 2019

Improving the Performance of Automatic Speech Recognition Using Blind Source Separation

Santosh Kumar S, Avinash J L, Nataraja N

***Abstract**: In real world applications, Speech recognition system have grown due its significance in various online and offline applications such as security, robotic application, speech translator etc. These systems are widely used now-a-days where acquisition of signal is performed using various instruments which causes noise, source mixing and other impurities which affects the performance of speech recognition system. In this work, issue of source mixing in original speech signal is addressed which causes performance degradation. In order to overcome this we propose a new approach which utilizes non-negative matrix factorization modelling. This method utilizes scattering transform by applying wavelet filter bank and pyramid scattering to estimate the source and minimization of unwanted signals. After estimation the signals or sources, source separation algorithm is implemented using optimization process based on the training and testing method. Proposed approach is compared with other existing method by computing performance measurement matrices which shows the better performance*

***Index Terms**: About four key words or phrases in alphabetical order, separated by commas.*

I. INTRODUCTION

A process of speech signal translation into a useful message is known as speech recognition; if this process is stimulated automatically then it is known as Automatic Speech Recognition [1]. Speech is the most common way for exchanging information among people. Translation of speech of signal into text or morphing of speech signal is carried out by processing the signal through software tools. Recently, robotic applications also have been developed based on machine oriented applications. Key component of these applications is based on speaker recognition system and speech recognition system.

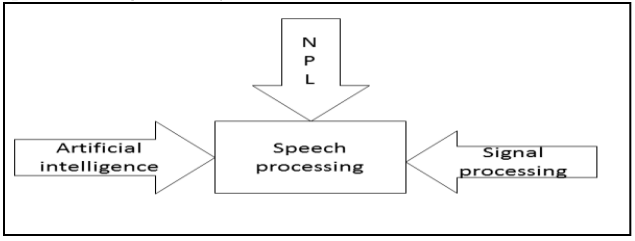


Fig 1: Speech processing area.

Recently, use of Automatic Speech Recognition Systems has experienced tremendous growth in various real time applications such as translators, authentication or security systems, voice commanders etc.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author

Santosh Kumar S, Department of E&CE, Sri Venkateshwara College of Engineering, Bengaluru, India.
Avinash J L, Department of E&CE, Sri Venkateshwara College of Engineering, Bengaluru, India.
Nataraja N, Department of E&CE, Sri Venkateshwara College of Engineering, Bengaluru, India.

Speech Recognition Systems includes concepts of Artificial Intelligence Systems [3], Natural Language Processing etc. [4].Figure 1 show, in which areas speech processing is considered as dominating and effective for applications.
Speech processing techniques covers various technologies which include Speech Encoding, Speech Recognition, and synthesis of speech, Speaker Recognition and translation of language mentioned in fig 2.

1.Speech encoding
2.Speech synthesis
3.Speech recognition
4.Speaker recognition
5. Language translation

Fig.2. Application of speech processing

Speech encoding process is performed when speech signal is transmitted using communication channel. These days voice media is also used widely which required encoding during transmission which transforms the speech signal from one form to another. This process includes signal compression and error correction.
Real time application for Automatic Speech Recognition system demands for more robustness and more accuracy. By taking this issue into account, several techniques have been proposed by researchers [1]. Y. Shao et al. [5] discussed that performance of Automatic Speech Recognition system can be improved by removing the noise in the speech signal. In order to address this, wavelet based schemes has been proposed to remove the noise in signal by applying pre-processing feature extraction method and error in training and testing has been minimized using Bayesian classifier.
ASR state-of-art schemes provide better results for recognition if these approaches are implemented on a clean speech signal. In real time application of speech recognition systems, it is considered that noise is always present in original speech signal which causes performance degradation and robustness of ASR system. To overcome this issue of noisy speech signal, various methods have been discussed for speech signal enhancement. Spectral Subtraction method is discussed in [6] is used for signal enhancement. For noise removal from original speech signal, noise estimation is a challenging task when dealing with real time applications. For noise removal and signal enhancement, Wiener filtering method, subspace method, minimum mean square error (MMSE) has been discussed [7]. These techniques compute the second order statistical structure to compute the statistical difference between noise and speech signal. It has been proved that if noise is removed then the performance of speech recognition system can be improved.

Recently, S. Khoubrouy have proposed a new technique for distant speech recognition with the help of combination of single –level beam-forming and combination of word-hypothesis level. In [12], wavelet based scheme is presented for automatic speech recognition system. Wang Y et al [13] proposed a new technique using speech and facial expression recognition.

For automatic speech recognition system, machine learning techniques have been considered as a promising technique which follows training and testing phases for recognition. Based on the machine learning technique, recurrent neural network have been proposed recently for ASR application system.

There are various challenges present in the field of speech recognition system which are mentioned below:

For speech recognition system, parameters which affect the performance can be categorized in three main classes.

- First class considers the affecting parameters as quality of voice which is affected due to behavioural factors or physiological factors of speaker. Physical parameters such as smoking habit, environmental effects, disease etc.
- Second class describes that during signal transmission for high level, long-term modulation is implemented which affects the signal quality and performance of speech recognition system.
- Third class considers the alteration in pronunciation such as substitutions or suppression of speech.

These challenges in the field of Automatic Speech Recognition System can be considered as variable channel, caused during signal acquisition where noise always varies for a given time. Speech continuity is also anissue whichaffects the performance of ASR. During signal acquisition, movement in body, gestures also cause variation in the original signal which induces miss interpretation in speech signal recognition. Similarly, mixing of various sources also degrades the performance of ASR system.

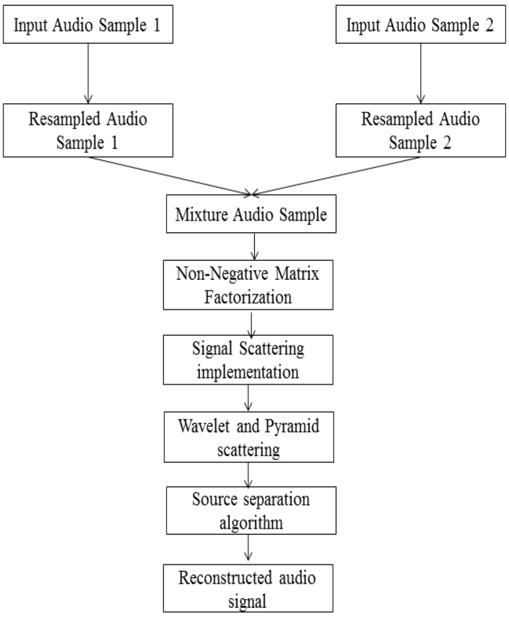
To address these issues, here we propose a new approach for source separation from the noisy input samples using Nonnegative Matrix Factorization based scheme which performs audio de-mixing using matrix decomposition methods.

II. PROPOSED MODEL

This section deals with the source separation using proposed approach for automatic speech recognition system. Proposed approach is divided into three main sections to carry out source separation as mentioned below:

1. Modelling of Nonnegative matrix factorization
2. Implementation of scattering approach using wavelet transform and Joint Time-Frequency Pyramid Scattering
3. Implementation of source separation method.

Figure 3 shows overall system architecture of proposed source separation method. In following section we describe proposed model.



Input Audio Sample 1
Input Audio Sample 2
Resampled Audio Sample 1
Resampled Audio Sample 2
Mixture Audio Sample
Non-Negative Matrix Factorization
Signal Scattering implementation
Wavelet and Pyramid scattering
Source separation algorithm
Reconstructed audio signal

Fig.3. Overall system architecture of source separation approach

2.1 Modelling of Nonnegative matrix factorization

In order to do the modeling of non-negative matrix factorization, a configuration is considered where a temporal speech signal $s(t)$, is considered a combination of two signals such as

$$s(t) = s_1(t) + s_2(t) \tag{1}$$

In this work we estimate the noise parameters present in the original signal. Source separation schemes which are based on the NMF techniques require non-negative time frequency representation of signal $s(t)$. This can be estimated using power spectrum of the signal which is denoted as

$$\Theta(s) \in \mathbb{R}^{p\times q} \tag{2}$$

Where p frequency bins and temporal frames are denoted by q . According to non-negative matrix factorization method, activations functions are computed which represents speech components in dictionaries. Activations functions are given as

$$\mathrm{B}_i \in \mathbb{R}^{r\times q} (3)$$

This can be achieved by solving equation (4)

$$\min_{\mathrm{B}_i} \mathbb{D}(\Theta(s) | \sum_{i=1,2} \mathbb{D}_i \mathcal{A}_i) + \lambda \sum_{i=1,2} \mathcal{R}(\mathcal{A}_i) (4)$$

$\mathcal{A}$denotes non-negative activation function which is given as $\mathcal{A}_i \in \mathbb{R}^{r\times q}$, dictionary is denoted by$\mathbb{D}$ i.e. $\mathbb{D}_i \in \mathbb{R}^{p\times r}$

According to the above expression, first term performs optimization by measuring thedissimilarity between input speech signal and the channels. In this work squared Euclidean distance, the Kullback- Leibler divergence has been used for optimization purpose. Second term of the expression provides the desired structure of activation functions.

2.2 Formulation of Scattering Transform

Speech signals contains discriminative feature which contain longer temporal context which can be constructed by applying scattering transform. These features can be used for classification purpose, in this work we aim on source separation or unmixing of sources.

In order to achieve this, a representation of features requires which provides the temporal structure and results in discriminative analysis. For this objective, at each level of signal, multi-level representation is constructed which consists scattering coefficients. Scattering transform is implemented using two stage process: (a) wavelet filter bank and (b) Joint Time-Frequency Pyramid Scattering.

a) Wavelet filter bank

In this section we describe the modelling of wavelet filter bank, applied on a speech signal. Let $\gamma(t)$ be a band-pass filter which havs the capability to represent frequency components of speech signal and spatial information. Here, a complex wavelet transform is considered along with its quadrature phase. Relationship between Fourier transform and quadrature phase is given as

$$\mathcal{F}_{\gamma}(w) \approx 0 \, for \, w < 0 \qquad (5)$$

It is assumed that centre frequency of Fourier transform is 1 and Q^{-1} is the order of bandwidth.Frequencies where wavelet filters are computed, denoted by $\eta = 2^{j/Q}$, which are able to cover the positive frequencies of the signal as given below:

$$\forall_w \geq 0, 1 - \varepsilon \leq |\mathcal{F}_{\Theta}(w)|^2 + \frac{1}{2}\sum_{\eta \in \Lambda} \left|\mathcal{F}_{\gamma_\eta}(w)\right|^2 \leq 1 \; (6)$$

$\mathcal{F}_{\gamma_\eta}(w) = \hat{\gamma}(\eta^{-1}w)$ and $\forall_w$ is a index set of frequency computed $\eta = 2^{j/Q}$

Resultant filter bank consist constant number of bands and the transform of a signal can be given as

$$\mathrm{W}_s = \left\{s * \Theta(t), x * \gamma_\eta(t)\right\}_{\eta \in \Lambda} \qquad (7)$$

b) Joint Time-Frequency Pyramid Scattering

In this section we describe the computation of scattering coefficients which are computed by an iterating process over a wavelet transform. These scattering coefficients provide non-linear representation of the speech signal. During wavelet transform computation, complex phase are generated with nonlinearities in signal which are removed here.

These coefficients are arranges as follows

$$|W^1|_s = \left\{s_i^1\right\}_{i=1 \dots 1+|\Lambda|} = \left\{s * \Theta_1(\Delta_1 n), \left|s * \gamma_{1,\eta_1}(\Delta_1 n)\right.\right\}_{\eta \epsilon \Lambda} (8)$$

Δ_1 denotes critical sampling rate, using this layer coefficients, localization information can be achieved in time and frequency domain by adjusting the frequency resolution of wavelets. Robustness of the system is increased by down sampling the signal with filter bank and taking the modulus of oscillatory components. For down-sampling same temporal resolution is used for each signal using low pass anti-aliasing filter.

$$|W^2|_s = \left\{s_i^1\right\}_{i=1 \dots 1+|\Lambda|} = \left\{s * \Theta_2(\Delta_2 n), \left|s * \gamma_{1,\eta_2}(\Delta_2 n)\right.\right\}_{\eta \epsilon \Lambda} (9)$$

2.3 Implementation of source separation algorithm

As discussed before that our main is to perform source separation from the original speech signal. In this section we present a solution to resolve inverse problem of source separation by applying sparse non-negative matrix factorization in scattering domain. This problem is considered from equation 1 where we have various components of source and training data denoted as $S_i = \{s_{ij}\}$ We consider a feature vector for simplification of features $\Theta_j(s_i)$ which are obtained with the help of localization of scattering method at differnet resolution and different sampling rate.we consider two feature set such as Θ_1, Θ_2 where Θ_1 contains more information and Θ_2 representes stable temporal context. In this work independent models are trained by following below mentioned expression

$$\min_{\mathbb{D}_i} \sum_{s \in S_i} \min_{z \geq 0} \frac{1}{2} \left|\left|\psi_j(s) - \mathbb{D}_i^j z\right|\right|^2 + \gamma_j ||z|| \qquad (10)$$

Non-negative dictionary is denoted as $\mathbb{D}_i^j$, which is computed of i source , considering j resolution. During testing phase, given input and actual inputs are estimated by minimizing below mentioned problem

$$\min_{\mathbb{D}_i} \sum_{j=1,2} \sum_{i=1,2} \frac{1}{2} \left|\left|\psi_j(s_i') - \mathbb{D}_i^j z_i^j\right|\right|_2^2 + \gamma \left|\left|z_i^j\right|\right| \qquad (11)$$

According to proposed mode, decomposition of both level should maintain coherency at each level of temporal resolution. First temporal level denotes the temporal data localization and results in higher resolution representation of speech signal and second level coherence provide temporal coherence of signal. Proposed approach helps to measure the coherence between two signals or levels which helps to resolve the issue of source separation.

III. Results and Analysis

In this section we describe the performance analysis of proposed model of source separation approach using non-negative matrix factorization method. Here we have considered two scenarios for performance evaluation:

- speaker-specific evaluation
- Multi-speaker performance evaluation.

Here we have made two type of configuration for evaluation according to first configuration evaluation speaker-specific model is trained using different samples and tested with the same user's speech signal which is not present in the training set. In another configuration evaluation, a mixture of speech samples is trained by taking the samples speech of male and female, during testing of multi-speaker evaluation also, test sample is not present in the trained dataset. Mixing of signal is performed at 0 dB and resampling is performed at 16 KHz Simulation parameters are given in table 1.

Table I:Simulation parameters

Parameter	Value
Sampling Frequency	16 kHz
Signal Mixing	0 dB
Resolution Adjustment	32
Sampling time	2048
Normalization Constant	1e-3

The experimental study using parameters which are mentioned in table 1 is carried out suign GRID dataset [10].we take dataset samples of 3 male and 3 female users and training is performed by using 500 clips. In this work training and testing dataset is created using 12 different users where 6 are female and 6 are males users using various combinations of samples. In order to measure the performance we use measurement matrices which are: source to distortion ratio, source to interference ratio and source to artifact ratio which can be computed as mentioned below:

$$\mathrm{SourceToInterfernce\ Ratio(SRT)} = 10 \log \frac{\sum_{\mathrm{t}} \mathrm{y}_{\mathrm{is}}^2(\mathrm{t})}{\sum_{\mathrm{i \neq j}} \sum_{\mathrm{t}} \sum_{\mathrm{t}} \mathrm{y}_{\mathrm{is}}^2(\mathrm{t})} \qquad (12)$$

Retrieval Number F8112088619/2019©BEIESP
DOI: 10.35940/ijeat.F8112.088619
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

$$\mathrm{SDR} + 10\log\frac{\sum_t s_{is}^2(t)}{\sum_t (s_k(t) - \mathbb{C}y_{is}(t-\beta))^2} \qquad (13)$$

$y_{is}^2$is estimated signal of $s(t)$, $\mathbb{C}$ is constant to compensate the difference of amplitudes and phase difference between input and output is given by β and SDR is Source ToI Distortion

$$\mathrm{SDR} = 10\log_{10}\frac{||target+interference+noise||^2}{||artifact||^2} (14)$$

Where SDR is Source ToI Artificial Distortion and proposed approach a standard NMF model is used which contains 1024 samples with 50% overlapping. For dictionary selection 200 atoms are used for speaker specific model and 400 atoms are used for multi-speaker model.

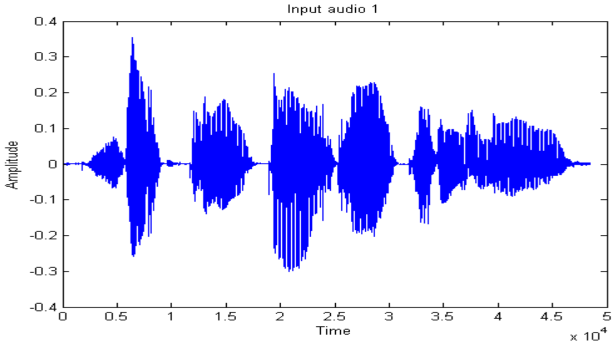


Fig 4(a): *Input audio sample 1*

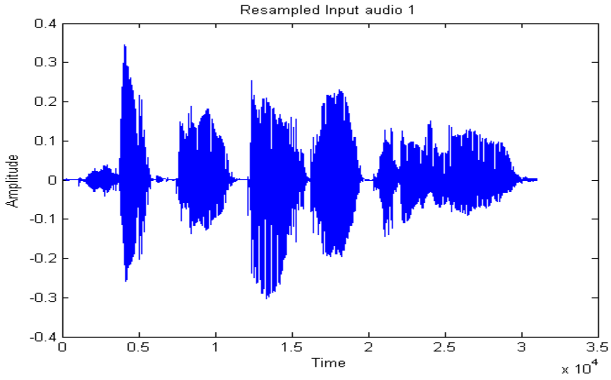


Fig 4(b): *Resampled Input audio sample 1*

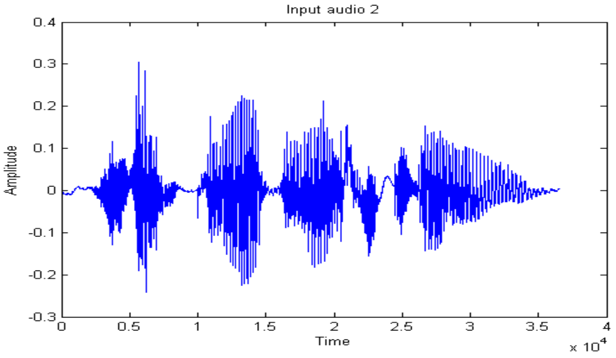


Fig 4(c): Input audio sample 2

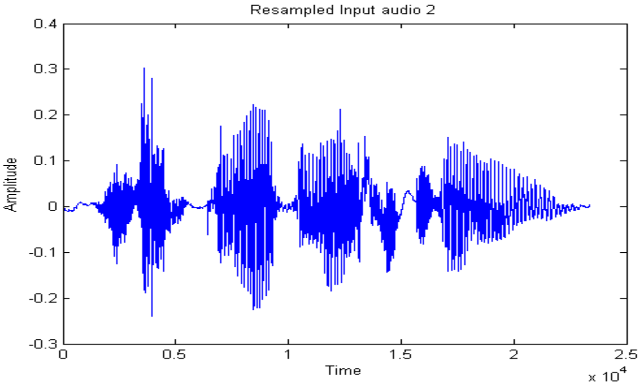


Fig 4(d) : Resampled input audio sample 2

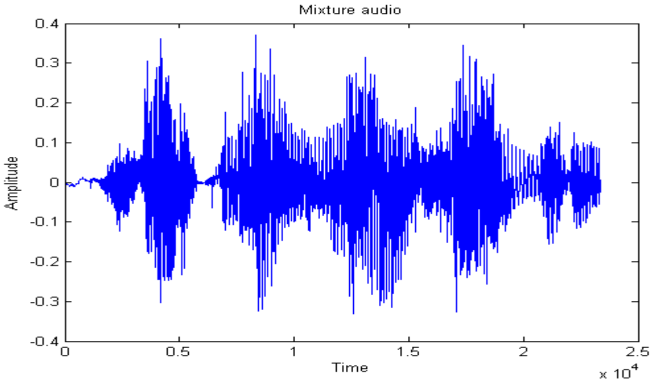


Fig 4(e) : Mixture audio

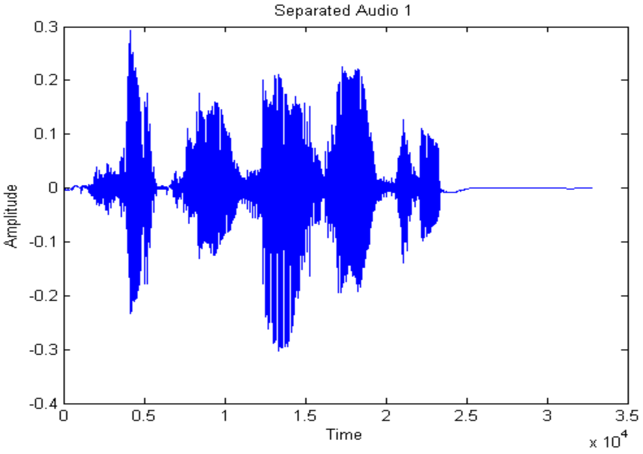


Fig 4(f):Separated audio sample 1

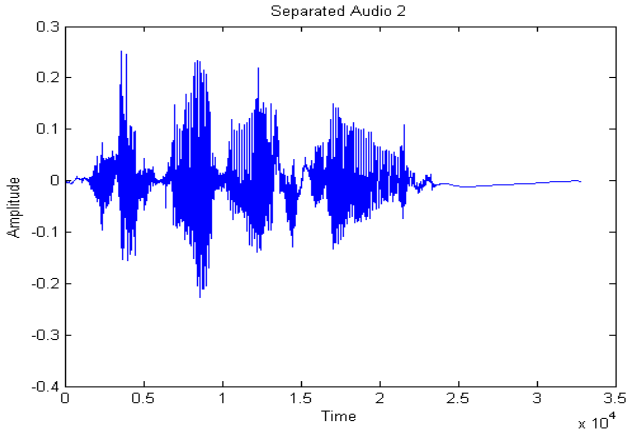


Fig 4(g): Separated audio sample 2

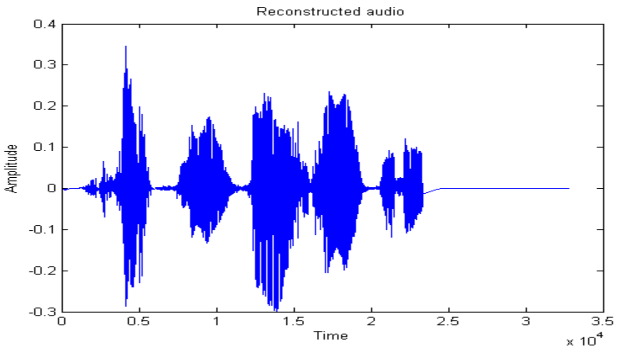


Fig 4(h): Reconstructed original signal

In figure 4, overall processing of proposed approach is presented. As per the proposed methodology, initially audio samples are taken and resampled. Figure 4(a), and 4 (c) are input audio samples where resampled data is presented in figure 4(b) and 4(d). later these signals are combined together to create a mixture audio. By considering the mixture audio, proposed approach is implemented to separate the mixture signal and original speech signals as shown in figure 4(f) and 4(g) for separated audio and 4(h) shows the reconstructed original signal.

In order to show the robustness of proposed model we perform a comparative study by considering speaker –specific and multi-speaker dataset. For speaker-specific performance is presented in table 2 and similarly for multi-speaker evaluation, comparative study is presented in table 3.

Table 2Speaker-specific performance comparison

	Existing Model [11]	Proposed Model
SDR	8.93	10.48
SIR	18.8	19.74
SAR	9.46	11.08

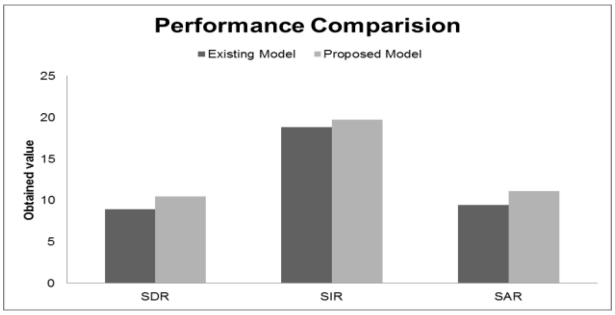


Fig5: Comparison analysis

By analyzing table I and II, it can be concluded that proposed model is able to perform better when compared to state of art technique for source separation or audio demixing purpose.

CONCLUSION

Here in this work a new approach is proposed for source separation or audio demixing using non-negative matrix factorization. This approach is implemented using signal processing techniques where, input data is resampled, mixture of two audio samples is created. This mixture is passed through wavelet filter bank and pyramid scattering where signal is decomposed and finally a source separation algorithm is implemented to perform unmixing of sources.

REFERENCES

1. D. O'Shaughnessy, "Acoustic Analysis for Automatic Speech Recognition," in Proceedings of the IEEE, vol. 101, no. 5, pp. 1038-1053, May 2013.
2. K. Junjea, "A dynamic segment based statistical derived PNN model for noise robust Speech Recognition," 2015 Third International Conference on Image Information Processing (ICIIP), Waknaghat, 2015, pp. 320-325.
3. G. Badr and B. J. Oommen, "On optimizing syntactic pattern recognition using tries and AI-based heuristic-search strategies," in IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 36, no. 3, pp. 611-622, June 2005.
4. E. Zavarehei, S. Vaseghi and Q. Yan, "Noisy Speech Enhancement Using Harmonic-Noise Model and Codebook-Based Post-Processing," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 4, pp. 1194-1203, May 2007.
5. Y. Shao and C. H. Chang, "Bayesian Separation With Sparsity Promotion in Perceptual Wavelet Domain for Speech Enhancement and Hybrid Speech Recognition," in IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 41, no. 2, pp. 284-293, March 2011.
6. Yang Lu and Philipos C. Loizou, "A geometric approach to spectral subtraction," Speech Commun., vol. 50, no. 6, pp. 453– 466, June 2008
7. Loizou, P. C., Speech Enhancement: Theory and Practice, CRC Press, 2007
8. S. Khoubrouy; J. Hansen, "Microphone Array Processing Strategies for Distant based Automatic Speech Recognition," in IEEE Signal Processing Letters , vol.PP, no.99, pp.1-1, 18 July 2016
9. X. Chen; X. Liu; Y. Wang; M. J. F. Gales; P. C. Woodland, "Efficient Training and Evaluation of Recurrent Neural Network Language Models for Automatic Speech Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing , vol.PP, no.99, pp.1-1
10. M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audiovisual corpus for speech perception and automatic speech recognition," J. of the Acoustical Society of America, vol. 120, pp. 2421, 2006
11. P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in ICASSP, 2014, pp. 1562–1566.
12. Adam TB, Salam MS, Gunawan TS, Wavelet Cepstral Coefficients for Isolated Speech.TELKOMNIKA Telecommunication Computing Electronics and Control.2013; 11(5): 2731–2738.
13. Wang Y, Yang X, Zou J. Research of Emotion Recognition Based on Speech and Facial Expression. TELKOMNIKA Telecommunication Computing Electronics and Control.2013; 11(1): 83–90.

AUTHORS PROFILE

Santosh Kumar S Assistant Professor, Department of Electronics & Communication Engineering, Sri Venkateshwara College of Engineering, Bengaluru. - 562157. He has more than 11 years teaching experience. His area of interest are Digital Signal Processing, image processing, adaptive signal signal processing, modern digital signal processing, Digital Communication.

Email ID: reachsun@gmail.com

Avinash J L, Assistant Professor, Department of Electronics & Communication Engineering, Sri Venkateshwara College of Engineering, Bengaluru. - 562157. . He has more than 06 years teaching experience His area of interest are image processing, adaptive signal signal processing, modern digital signal processing

Email ID: avinash.jlb@gmail.com

Nataraja N, Assistant Professor, Department of Electronics & Communication Engineering, Sri Venkateshwara College of Engineering, Bengaluru. - 562157. He has more than 6 years teaching experience His area of interest are Digital signal Processing, MIMO, OFDM and Antennas.

Email ID:nataraja.sp85@gmail.com