ICA BASED SINGLE MICROPHONE BLIND SPEECH SEPARATION TECHNIQUE USING NON-LINEAR ESTIMATION OF SPEECH

Chandan K A Reddy, Anshuman Ganguly, Student members, IEEE, Issa Panahi, Senior member, IEEE

Dept. of Electrical Engineering, The University of Texas at Dallas, Richardson TX

ABSTRACT

In this paper, a Blind Speech Separation (BSS) technique is introduced based on Independent Component Analysis (ICA) for underdetermined single microphone case. In general, ICA uses noisy speech from at least two microphones to separate speech and noise. But ICA fails to separate when only one stream of noisy speech is available. We use Log Spectral Magnitude Estimator based on Minimum Mean Square Error (LogMMSE) as a non-linear estimation technique to estimate the speech spectrum, which is used as the other input to ICA, with the noisy speech. The proposed method was tested for machinery, babble and traffic noise types mixed with speech at Signal to Noise Ratios (SNRs) of -5 dB, 0 dB and 5 dB. Objective and subjective results show high quality and intelligibility in the separated speech using the developed method.

Index Terms— ICA, BSS, Single Microphone, Non-linear estimation.

1. INTRODUCTION

Extensive research has been done in the past decade on Blind Speech Separation (BSS) algorithms for underdetermined case [1, 2]. Majority of the work involves popular BSS techniques like Independent Component Analysis (ICA) and Independent Vector Analysis (IVA) [3, 4]. Theoretically, to satisfactorily separate speech and noise sources (two sources), we need noisy speech data from at least two microphones with less correlation with each other. Most of the existing BSS techniques use a microphone array with multiple microphones placed in different orientation or at least two microphones with linear placement to collect the data. The reason being, underdetermined BSS techniques cannot exploit the "spatial diversity" of the sources [5].

In [6, 7] researchers have proposed underdetermined BSS techniques using ICA and sparse coding. Their approach is to project the sources onto a set of basis functions whose coefficients are as sparse as possible. However, [8] shows that these techniques do not work well when the trained basis functions of the sources (speech and noise) overlap. Single channel speech enhancement using ICA is proposed in [9]. They employ a training phase on large ensemble of clean speech to reveal their underlying statistical independent basis and estimate the distribution of the ICA transformed data. But this algorithm does not suit well for changing noise environments as the estimation of demixing matrix is performed offline and requires large data set.

In this paper, we propose a BSS technique which uses the data stream (noisy speech) collected from a single microphone. This is the worst case scenario of a system which is underdetermined with two unknowns and one equation. In our method, we estimate the speech magnitude spectrum obtained using the Single Channel Speech Enhancement (SCSE) non-linear estimator by minimizing the mean square error of log magnitude spectrum (LogMMSE). We analyzed popular SCSE techniques based on Spectral subtraction, Statistical model based method and Subspace method to choose the best method for preprocessing the data. The estimated speech and the original noisy speech are used as the two inputs to ICA which then separates speech and noise. It is shown through analysis that the estimate of speech by LogMMSE is less dependent on the noise components than that of the speech in unprocessed noisy data. This aids in enhancing the performance of ICA in separating the speech from noise. Although LogMMSE enhances the speech, it also distorts the speech at low Signal to Noise Ratios (SNRs) due to inaccuracies in the estimate of noise power spectrum. The results show that ICA recovers distorted components from the original data. The separated speech signal from ICA is finally filtered using LogMMSE to suppress the residual noise. LogMMSE is computationally inexpensive; hence we use it to suppress the residual noise at the end.

The performance of the proposed method is evaluated by considering noisy speech at SNR levels of -5 dB, 0 dB and 5 dB for machinery, babble and traffic noise signals mixed with speech. The method was evaluated objectively using four different metrics which measured the quality, intelligibility and amount of noise suppression. Subjective tests were performed for the aforementioned noise types and SNR levels. The enhanced speech using the proposed method shows drastic improvement in quality and intelligibility.

2. CHOICE OF SCSE TECHNIQUE

In the past two decades, many SCSE algorithms have been proposed. Among them Spectral Subtraction, Statistical model based methods and Subspace methods are well known to give good quality and intelligibility when tested both objectively and subjectively. In a 2-stage processing approach, the challenge is to pick the right method in the first stage which estimates the speech well and improves the performance of ICA in the second stage.

We know [10] that the performance of ICA improves when the sources to be separated are less dependent on each other. Also when the dependency between the two microphone signals is less, the performance in separating the signals improves. Therefore, we consider two factors while choosing the SCSE method. First, the enhanced signal using SCSE should be of high quality and intelligibility. Second, the output of SCSE (enhanced speech) should be less dependent with the second input to ICA, which is the original noisy speech as shown in Figure 4. The dependency between the signals can be quantified by measuring Mutual Information using the method of estimation [12]. Figures 1, 2 and 3 show the plots of SNR versus Mutual Information for machinery, traffic and babble noise types. The mutual information was calculated by taking an average over 20 noisy speech signals and their corresponding enhanced speech signals. The three overlapped curves are for three different SCSE techniques each chosen from Spectral Subtraction, Statistical model based method and Subspace methods respectively.

From Figures 1, 2 and 3, Spectral Subtraction seems to give least mutual information across all SNR values for machinery and babble noise types. Subspace methods do well in making the signals independent for traffic noise. Although Spectral Subtraction and Subspace methods give lower mutual information, they introduce significant musical noise and speech distortion compared to Statistical model based method, especially in comparison with LogMMSE. These comparisons are well illustrated quantitatively in [13, 14]. Therefore with little compromise on mutual information, while achieving better speech quality and intelligibility, we chose LogMMSE as our SCSE technique. Let $x(n)$ be the noisy speech given by $x(n) = s(n) + d(n)$, where $s(n)$ is the clean speech and $d(n)$ is the noise. The mean-square error of the log-magnitude spectra is given by,

$$E\{(log\, S_k - log\, \hat{S}_k)^2\} \quad (1)$$

where S_k is the k^{th} bin of magnitude spectrum of $s(n)$ and $\hat{S}_k$ is k^{th} bin of estimated clean speech magnitude spectrum.

The optimal log-MMSE estimator can be obtained by evaluating the conditional mean of the $log\, S_k$, that is,

$$log\, \hat{S}_k = E\{log\, S_k | X(\omega_k)\} \quad (2)$$

Hence the estimate of the speech magnitude spectrum is given by, $\hat{S}_k = exp(E\{log\, S_k | X(\omega_k)\})$ (3)

The expectation in (3) can be simplified using the moment-generating function of S_k conditioned on $X(\omega_k)$.

Letting $Z_k = log\, S_k$, then the moment-generating function of Z_k conditioned on $X(\omega_k)$ is given by,

$$\Phi_{Z_k|X(\omega_k)}(\mu) = E\{exp[\mu Z_k]|X(\omega_k)\} = E\{S_k{}^{\mu}|X(\omega_k)\} \quad (4)$$

By taking the derivative of $\Phi_{Z_k|X(\omega_k)}(\mu)$ with respect to μ and evaluating at $\mu = 0$, we get the conditional mean of the $log\, S_k$:

$$E\{log\, S_k|X(\omega_k)\} = \frac{1}{2} log\, \lambda_k + \frac{1}{2} log\, v_k + \frac{1}{2} \int_{v_k}^{\infty} \frac{e^{-t}}{t} dt \quad (5)$$

where $\lambda_k = \frac{\lambda_s(k)}{1+\xi_k}$ in which $\lambda_s(k)$- variance of the k^{th} spectral component of the clean speech. ξ_k is referred as a priori SNR and is defined as, $\xi_k = \frac{\lambda_s(k)}{\lambda_d(k)}$ where $\lambda_d(k)$ is the variance of the k^{th} spectral component of the noise.

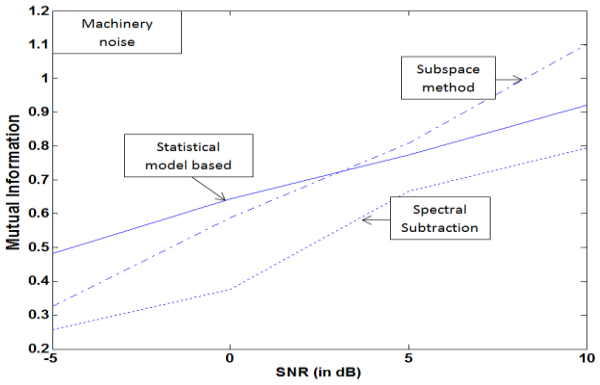

Fig. 1 Mutual Information measures for machinery noise for 3 different SCSE techniques

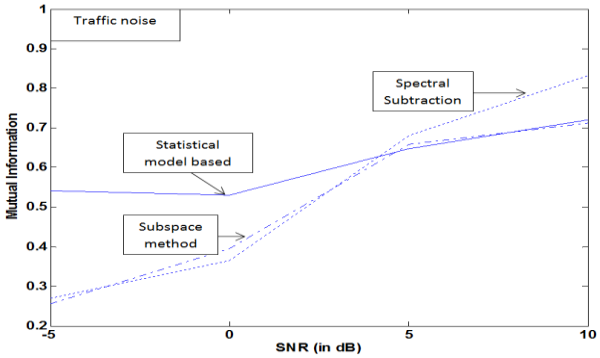

Fig. 2 Mutual Information measure for Traffic noise for 3 different SCSE techniques

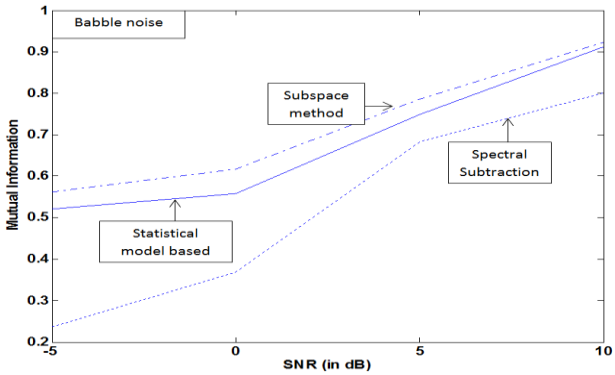

Fig. 3 Mutual Information measures for babble noise for 3 different SCSE techniques

$v_k = \frac{\xi_k}{\xi_k+1}\, \gamma_k$ in which $\gamma_k = \frac{S_k{}^2}{\lambda_d(k)}$ is the a posteriori SNR.

The final estimate of speech magnitude spectrum is given by,

$$\hat{S}_k = \frac{\xi_k}{\xi_k+1} exp\{\frac{1}{2} \int_{v_k}^{\infty} \frac{e^{-t}}{t} dt\}\, X_k \quad (6)$$
$$\triangleq G_{LSA}(\xi_k, v_k)\, X_k$$

The exponential integral in equation (6) is evaluated numerically.

3. BSS USING FASTICA

The BSS techniques aim at finding a set of N unknown source signals $s_i(n)$ from a set of P observed signals $x_j(n)$ [10]. The P observed signals can be modeled as linear instantaneous

mixtures of N unknown source signals under the general frame work given by,

$$\mathbf{x} = \boldsymbol{A}\,\boldsymbol{s} \tag{7}$$

where $\mathbf{x} = [x_1(n), \ldots\ldots, x_P(n)]^{\mathrm{T}}$ is the observation vector, $\boldsymbol{s} = [s_1(n), \ldots\ldots, s_N(n)]^{\mathrm{T}}$ is the source vector and $\boldsymbol{A}$ is the scalar mixing matrix. We assume that the signals and the mixing matrix are real–valued and the sources are mutually statistically independent. Now the goal of the ICA algorithm is to recover $\boldsymbol{s}$ from the observation vector $\mathbf{x}$ by finding a transformation in which the transformed signals are less dependent on each other. In our case, we have the number of observations equal to the number of sources ($P = N = 2$). Hence we assume that the mixing matrix $\boldsymbol{A}$ to be invertible. The problem now is to find a weighting matrix $\boldsymbol{W}$ so that the linear transformation of the observed variables is given by,

$$\mathbf{y} = \boldsymbol{W}\,\mathbf{x} \tag{8}$$

where $\mathbf{y} = [y_1(n), \ldots\ldots y_N(n)]$ is the estimate of the original clean speech. The matrix $\boldsymbol{W}$ in (2) is obtained as the (pseudo) inverse of the estimate of matrix $\boldsymbol{A}$.

FastICA is one of the most popular iterative methods for ICA because of its high convergence speed and satisfactory performance in a wide range of applications [11]. The matrix $\boldsymbol{W}$ is determined by finding the directions in which the negentropy is maximized. The approximation for the negentropy is of the form,

$$J(y_i) \approx c[E\{G(y_i)\} - E\{G(v)\}]^2 \tag{9}$$

where $G(.)$ is any non-quadratic function, c is any constant and v is a Gaussian variable of zero mean and unit variance. y_i is a random variable with zero mean and unit variance. We use cumulant based approximation for $G(.)$ given by $G(\mathrm{y_i}) = -\,log\,cosh y_i$. In order to find the individual independent component, or projection pursuit direction as $y_i = \mathbf{w}^{\mathbf{T}}\mathbf{x}$, we maximize the function J_G given by,

$$J_G(\mathbf{w}) = [E\{G(\mathbf{w}^{\mathbf{T}}\mathbf{x})\} - E\{G(v)\}]^2 \tag{10}$$

where $\mathbf{w}$ is an P-dimensional weight vector constrained so that $\mathrm{E}\{(\mathbf{w}^{\mathbf{T}}\mathbf{x})^2) = 1$.

4. PROPOSED METHOD

Figure 4 shows the block diagram of the proposed method. The previous two sections gave the general frame work of LogMMSE and FastICA algorithms. In the proposed method, $x(n)$ is passed to SCSE block which is essentially LogMMSE algorithm. The output of SCSE $\hat{s}_1(n)$ is used as the first input to ICA and noisy speech $x(n)$ as the second input. The output $\hat{s}_2(n)$ is the estimate of speech and $\hat{d}(n)$ is the estimate of noise which is ignored. It is shown through our results that the quality and intelligibility of speech is much better in $\hat{s}_2(n)$ than in $\hat{s}_1(n)$. Though speech is slightly distorted in $\hat{s}_1(n)$ due to non-linear LogMMSE operation, ICA uses $x(n)$ which contains the original speech $s(n)$ mixed in noise. Hence during the demixing process in ICA, majority of the components in $\hat{s}_2(n)$ are recovered from $x(n)$ which is composed of clean speech. Hence $\hat{s}_2(n)$ sounds better than $\hat{s}_1(n)$ and has higher quality and intelligibility. The residual noise in $\hat{s}_2(n)$ is further suppressed using LogMMSE at the end, as LogMMSE enhances speech with minimal distortion when operated at higher SNR values. Finally, $\hat{s}_3(n)$ is of high perceptual quality and intelligibility as we show through results in the next section.

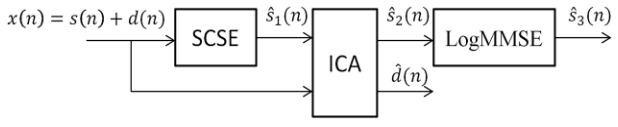

Fig. 4 Block diagram of the proposed method with improved Quality and Intelligibility

5. EXPERIMENTAL RESULTS

The proposed algorithm is evaluated using the noisy speech at SNR levels of -5 dB, 0 dB and 5 dB. The clean speech was selected from IEEE corpus. The noisy speech recorded by us using the microphones on the smartphone sampled at 16 kHz was used. We recorded three different noise types, machinery, traffic and babble for analyzing the performance of our method. The outcome of the proposed method is now compared with the standard single channel speech enhancement methods evaluated in [13, 14] which gave best subjective and objective results for different noise types. Statistical model based methods (in short we call SMM to refer in figures) were shown to perform the best across noise types [13, 14]. Figure 5 shows the effectiveness of the proposed method, spectrograms of noisy speech at SNR -5 dB for machinery noise, enhanced speech using SMM and enhanced speech using the proposed method. The proposed method suppresses the noise significantly by eliminating dark red frequency components that last across time in noisy speech spectrogram.

The proposed method is evaluated objectively using Perceptual Evaluation of Speech Quality (PESQ) as quality measure and Coherence Speech Intelligibility Index (CSII) as intelligibility measure [15]. PESQ ranges between 0.5 and 4, with 4 being high perceptual quality. CSII ranges from 0 and 1, with 1 being highly intelligible. Figure 6(a) shows PESQ measures for Noisy Speech, SMM and Proposed method for considered noise types and SNR levels. We can see that the proposed method gives statistically significant improvement compared to SMM. We also performed subjective tests with 11 normal hearing people. The results are shown in Figure 7. The subjects were presented with noisy speech, enhanced speech using SMM and enhanced speech using proposed method for 3 different noise types and corresponding SNR levels considered in this paper. Subjects were asked to score between 1 and 5 for each audio file presented based on how pleasant is the background noise (Quality) and how many words they can identify (Intelligibility). They were also given the flexibility to go back and change their scores after listening to other audio files. This gave good subjective comparison of our method and SMM to enhance the speech. The subjects preferred our method and rated high across all noise types and SNR values, except for babble noise at SNR of -5 dB. One reason why the proposed method fails for babble noise at low SNR levels is, both speech and babble noise have same statistical properties. They both can be modeled using Laplacian distribution. Also the Voice Activity Detector (VAD) present in LogMMSE fails to accurately detect the speech frames [16], which affects the

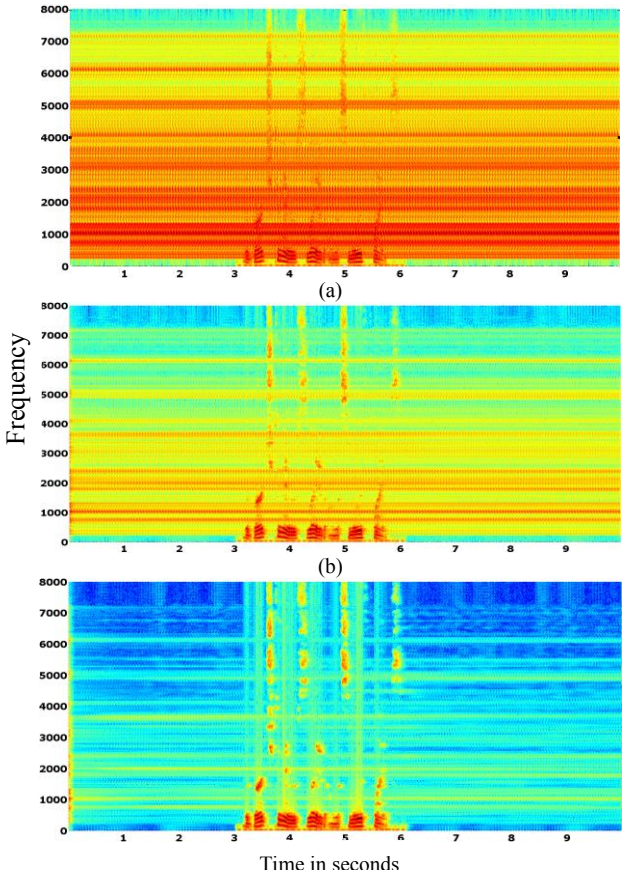

Fig. 5. (a) Spectrograms of noisy speech at -5 dB SNR for machinery noise. (b) Enhanced speech using SMM (c) Enhanced speech using proposed method.

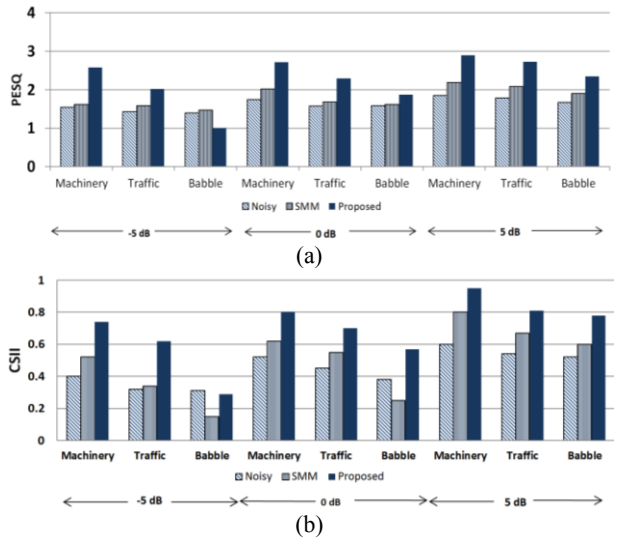

Fig. 6. Objective comparison between noisy speech, enhanced speech using SMM and enhanced speech using proposed method using (a) PESQ measures (b) CSII measures

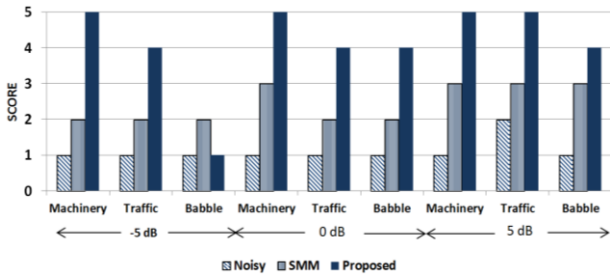

Fig.7 Comparison of Subjective results for noisy speech, enhanced speech using SMM and enhanced speech using proposed method.

precise estimation of noise power spectrum. This will result in wrong estimation of *a-priori* SNR estimation, which introduces background musical noise. Other than this noise condition, the proposed method performs well and is robust across noise types and different SNR levels.

The computational complexity of proposed method depends on two blocks, LogMMSE and FastICA. LogMMSE is computationally very fast and is of $\mathcal{O}(K) + const.$, where $K = 512$ is the number of frequency bins for 20 ms frame sampled at 16 kHz. On the other hand, convergence speed of FastICA depends on the number of samples of the signal and the type of signal we deal with [17]. Also the computational time of FastICA is dynamic in nature. There are several variations of FastICA which are faster. But the main contribution of this paper is to show that a single microphone signal can be used to separate two sources using ICA by prior non-linear estimation of speech signal.

Though most of the existing smartphones and other audio application devices like hearing aids come with 2 or more microphones, there is a challenge of accessing all the microphones at the same time (synchronization). Also, when both the microphones are very close to each other like in hearing aids, the performance of BSS algorithms deteriorate and the proposed method comes in handy to synthetically generate 2 input signals which are more suitable for BSS algorithms.

6. CONCLUSION

We proposed a single microphone Speech Enhancement technique using ICA by estimating the speech magnitude spectrum using non-linear LogMMSE estimator. The obtained enhanced speech is compared with SMM Speech Enhancement technique both objectively and subjectively. The proposed method shows promising results in improving quality and intelligibility of speech.

7. ACKNOWLEDGEMENT

This work was supported by the National Institute of the Deafness and Other Communication Disorders (NIDCD) of the National Institutes of Health (NIH) under award number 1R01DC015430-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

8. REFERENCES

[1] S. Araki, S. Makino, H. Sawada, R. Mukai, "Underdetermined blind speech separation with directivity pattern based continuous mask and ICA," *Signal Proc. Conf. 12th European*, IEEE, pp. 1991-1994, Sept 2004.

[2] T. Guo, Q. Lin, X. Gong, "An improved BLUES with adaptive threshold of condition number for separating underdetermined speech mixtures," *Intelligent Control and Inf. Process. (ICICIP), 2012 3rd Int. Conf. on,* pp. 694-698, 2012.

[3] H. Saruwatari, S. Ukai, T. Takatani, T. Nishikawa, K. Shikano, "Two-stage blind source separation combining SIMO-model-based ICA and adaptive beamforming," *Signal Process. Conf., 13th European,* IEEE, pp. 1-4, 2005.

[4] J. Harris, B. Rivet, S. M. Naqvi, J. A. Chambers, C. Jutten, "Real-time independent vector analysis with Student's t source prior for convolutive speech mixtures," *ICASSP,* IEEE, pp. 1856-1860, 2015.

[5] J.F. Cardoso, "Blind signal separation: Statistical principles," Proc. IEEE, vol. 86, no. 10, pp. 2009-2025, Oct.1998.

[6] M.S. Lewicki and T.J. Sejnowski, "Learning nonlinear overcomplete representations for efficient coding," in Advances in Neural Information Processing Systems, M. I. Jordan, M. J. Kearns, and S. A. Solla, Eds. Cambridge, MA: MIT Press, 1998.

[7] T. Virtanen, "Sound source separation using sparse coding with temporal continuity objective," in Proc. Int. Comput. Music Conf., pp. 231-234, 2003.

[8] G.J. Jang and T.W. Lee, "A probabilistic approach to single channel source separation," in Proc. Adv. Neural Inf. Process. Syst., 2003, pp. 1173-1180.

[9] Liang Hong, Rosca, J., Balan, R., "Independent component analysis based single channel speech enhancement," IEEE Proc. Signal Proc and Infor. Tech (ISSPIT 2003). pp 522-525. 14-17 Dec 2003.

[10] A. Hyvarinen, J. Karhunen and E. Oja, Independent Component Analysis. Newyork: Wiley, 2001.

[11] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," IEEE Trans. Neural Networks. Vol. 10, Issue 3, pp 626-634, May 1999.

[12] A. Kraskov, H. Stogbauer and P. Grassberger, "Estimating mutual information," Phys. Rev. E 69, 066138, vol. 69, Iss. 6-June 2004.

[13] Hu, Y. and Loizou, P., "Subjective comparison of speech enhancement algorithms," IEEE Int. Conf. Acoustic, Speech, Signal Process. (ICASSP), Toulouse, France, vol.1, Pp. 153-156, May 2006.

[14] Hu, Y. and Loizou, P., "A Comparative Intelligibility Study of Speech Enhancement Algorithms," *IEEE Int. Conf. Acoustic, Speech, Signal Process. (ICASSP),* Honolulu, HI, vol.4, Pp. 561-564, Apr 2007.

[15] P. Loizou, "Speech Enhancement: Theory and Practice", Boca Raton, FL: CRC Press, 2007.

[16] A. Ganguly, Y. Hao, C. Reddy, "Robust unsupervised voice activity detection using spectral magnitude estimators and Teager Kaiser energy operators", *IEEE Int Conf. Acoustic, Speech, Signal Process. (ICASSP)* 2017 (Submitted).

[17] V. Zarzosa, P. Comon, M. Kallel, "How fast is FastICA?", *EUSIPCO,* Florence, Italy, Sept. 4-6, 2006.