Adaptive energy threshold for monaural speech separation

S. Shoba and R. Rajavel

***Abstract*—Speech separation is a process of segregating the target speech from the noisy mixture. Human auditory system naturally has the capability to separate the speech from background noise; however the machine implementation of the same has been failed. Computational Auditory Scene Analysis (CASA) is a recent approach which models the human auditory system by computational means to separate the target speech from the noisy mixture. Most of the speech separation system based on CASA uses energy as one of the reliable feature to segregate the target speech from the noisy mixture. This research work proposes a new approach using adaptive energy selection threshold to segregate the target speech. The experimental result shows significant improvement in the signal to noise ratio as compared to the conventional energy threshold method.**

***Index Terms*—ASA, CASA, Conventional energy selection threshold, Adaptive energy selection threshold.**

I. Introduction

Speech separation process remains a challenging task for machines; however human being has the inherent capability to separate the intended speech from the speech mixture. Researchers have put lot of effort to develop computational systems to automatically separate target speech from the monaural noisy mixtures. Such computational system has potential for many speech processing applications such as automatic speech and speaker recognition, hearing aids, automatic music transcription, audio information retrieval etc. In the last few decades many researchers have developed monaural speech separation system using speech enhancement approaches [1], subspace analysis [2], model based approaches [3], and CASA [4][5][6]. In speech enhancement approaches, the statistical property of the speech and noise is used to enhance the speech that is degraded by noise. In sub space analysis, eigen decomposition of acoustic mixture has been done and then subspace analysis such as Independent component analysis (ICA), Principal component analysis (PCA) are used to remove the interference from the noisy mixture. In model based approaches, trained models of speech and noise are used for separation. All these methods require some form of prior knowledge about speech or interference and failed to show improvement in speech quality and intelligibility.

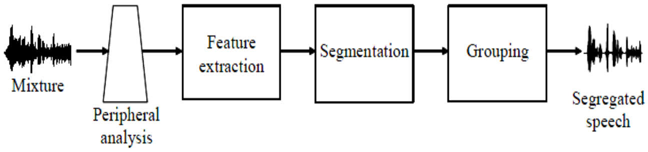

Fig. 1. Schematic representation of a CASA System [7]

Recently CASA has been introduced in the literature to separate target speech from the monaural noisy mixture using perceptual cues such as proximity in time and frequency, periodicity, onset/offset, amplitude and frequency modulation etc. Generally, typical CASA system consists of four stages as shown in Fig. 1 as (1) Analysis filter bank, (2) Feature extraction, (3) Segmentation and (4) grouping. Analysis filter bank decomposes the speech signal into time-frequency (T-F) units using filtering and windowing techniques. The auditory cues such as cross-channel correlation, correlogram, energy, etc., are extracted from the T-F units which will be used for segmentation and grouping. Based on the extracted features, the system generates segments corresponding to speech and noise and then grouped as speech stream and noise stream.

The Wang and Brown [8] CASA model uses cross channel correlation and temporal continuity as major cues to segregate the voiced speech. However, the system failed to segregate the unvoiced speech. Later, Hu and Wang proposed a system [9] to segregate the unvoiced speech using onset and offset. In another research work, the authors combined CASA with spectral subtraction [10] to segregate the unvoiced speech [11]. The CASA model proposed by Hu and Wang uses energy of a T-F unit as a primary cue to segregate the voiced speech [12]. In their work, the T-F unit energy is computed and compared with a constant threshold (θ^2_H = 2500) and determined whether that particular T-F unit belongs to speech source or noise source. This approach eliminates some T-F units which belong to speech where it is least dominant and in turn reduces the speech quality. This research work proposes an adaptive energy threshold to determine the T-F units belong to speech or noise. The performance of the proposed method is measured in terms of signal to noise ratio (SNR) and the experimental results shows significant improvement in SNR and in turn improves the speech quality and intelligibility.

The rest of the paper is organized as follows. The next section introduces an overview of the proposed speech separation system. The experimental results are discussed in

S. Shoba (Research Scholar) is with the Dept. of Electronics and Communication Engineering, SSN College of Engineering, Chennai, India. (E-mail: shobansb@gmail.com).
R. Rajavel is with the Dept. of Electronics and Communication Engineering, SSN College of Engineering, Chennai, India. (E-mail: rajavelr@ssn.edu.in).

978-1-5090-3799-5/17/$31.00 ©2017 IEEE

Section III. Finally, the conclusion and future direction of this research work is presented in Section IV.

II. Speech Separation System with Adaptive Energy Threshold

The conventional speech separation system using CASA is shown in Fig.1. In which, the noisy speech signal is decomposed into time-frequency (T-F) unit using Gammatone filtering [13] and windowing. The energy, temporal continuity and cross-channel correlation are extracted as a feature from the T-F units. These extracted features are used to segregate each T-F unit into speech dominant and noise dominant. Speech dominant T-F units are represented by one and noise dominant T-F units are represented by zero. This constitutes a binary mask and used in the synthesis filter bank to separate the noisy mixture into speech and noise. Each and every blocks of the typical CASA system is described in the following sub sections.

A. *Analysis Filter Bank*

The first stage in typical CASA system is the frequency analysis of a cochlea, i.e. decompose the speech into various T-F units. This has been done by a bank of 128 Gammatone filters whose impulse response with center frequency f is given by [12]

$$g(f,t) = b^N t^{N-1} e^{-2\pi bt}\cos(2\pi ft), \quad t \geq 0 \qquad \text{(1)}$$
$$0 \qquad\qquad\qquad\qquad \textit{Otherwise}$$

where N = 4 is the order of the filter and b is its equivalent rectangular bandwidth, f is the centre frequency which increases from 80 to 5000 Hz. Let $s(t)$ be the input noisy speech signal and the response from channel c is given as [12]

$$y(c,t) = s(t) * g(f,t) \qquad \text{(2)}$$

where * denotes linear convolution and the response of each filter is passed via the Meddis model which represents the firing rate of an auditory nerve fiber. The output of the Meddis modal denoted as $h(c,t)$ is divided into 20ms time frames with 10ms overlapping to generate a T-F unit which is denoted as $h(c,m)$, where c represents channel and m represents frame.

B. *Feature Extraction*

The auditory features such as energy, correlogram and cross channel correlation are extracted from the T-F units obtained from the Meddis hair cell model. The following subsection explains how an energy, correlogram and cross channel correlation features are extracted.

1) Correlogram: Correlogram is a technique for pitch extraction which is a running autocorrelation of filter responses across all filter channels [12]. For each T–F unit, its autocorrelation function of the hair cell response is given by [12]

$$A_H(c,m,\tau) = \frac{1}{N_c}\sum_{n=0}^{N_c-1} h(c,mT-n) * h(c,mT-n-\tau) \qquad \text{(3)}$$

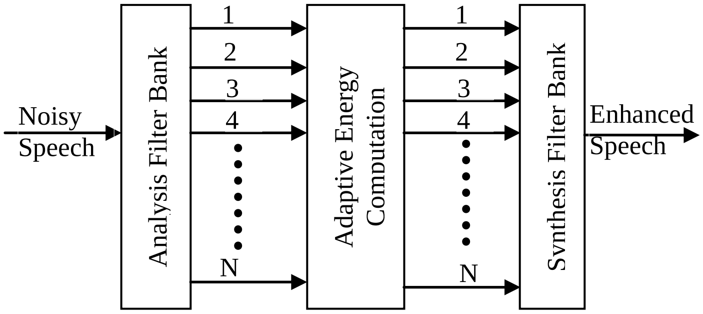

Fig. 2. Speech separation system with adaptive energy threshold

where τ is the delay ϵ [0 12.5] ms, N_c is number of frames in the channel c and n is the sample index.

2) Response Energy: The Energy feature $E(c,m)$ is obtained from the autocorrelation of hair cell output $h(c,m)$ at zero lag, which is given as [12]

$$E(c,m) = \frac{1}{N_c}\sum_{n=0}^{N_c-1} h(c,mT-n) * h(c,mT-n) \qquad \text{(4)}$$

3) Cross-channel correlation: The cross channel correlation between two adjacent filter channels indicates whether the filters mainly respond to the same source or not [12]. For a T-F unit with an autocorrelation denoted as $A_H(c,m,\tau)$, its cross-channel correlation is given by [12]

$$C_H(c,m) = \sum_{\tau=0}^{L-1} A_H(c,m,\tau) * A_H(c+1,m,\tau) \qquad \text{(5)}$$

C. *Segmentation*

The third stage in the segregation process is the segmentation stage in which each T-F unit is segregated into speech dominant and noise dominant T-F units using energy and cross channel correlation. The CASA model proposed by Hu-Wang computes the energy $E(c,m)$ of a T-F unit and compared with a constant threshold θ^2_H which is a spontaneous firing rate of the auditory nerve [12]. If the energy $E(c,m)$ of a particular T-F unit is greater than θ^2_H the corresponding T-F unit is denoted by 1 to represent the speech dominant, otherwise it is denoted as 0 to represent the T-F unit is noise dominant.

Similarly the cross channel correlation is compared with a threshold θc = 0.985 (chosen to be same as in [15]) and is 1 if it is greater than θc otherwise 0. This approach sometimes misses some of the T-F units in which speech is a least dominant and reduces the speech quality. In order to solve this issue, this research work proposes an adaptive energy threshold scheme for each channel to determine particular T-F units belong to speech dominant or noise dominant. The

TABLE I

AVERAGE SNR RESULTS FOR CONVENTIONAL THRESHOLD METHOD AND ADAPTIVE ENERGY SELECTION THRESHOLD

Noise	N_0	N_1	N_2	N_3	N_4	N_5	N_6	N_7	N_8	N_9	Average
Input mixture	-3.26	-4.07	10.19	4.34	3.99	-5.82	1.90	6.62	10.37	0.73	2.50
Speech with constant threshold	-0.42	6.02	9.91	6.79	5.49	-2.04	3.87	6.47	10.08	0.74	4.69
Speech with adaptive energy threshold	0.18	6.28	10.54	7.02	5.61	-1.31	3.61	6.18	9.08	0.90	4.81

adaptive energy threshold $\theta_{AT}(c)$ for each channel is computed by taking the scaled average energy of all T-F units in a particular channel as follows.

$$\theta_{AT}(c)=\frac{1}{N_c*\beta}\sum_{m=1}^{N_c}E(c,m) \tag{6}$$

where N_c is the total number of time frame in each channel c and β is scaling constant which is determined via experimentation as β = 1.7. Finally, the binary mask $M(c,m)$ is constituted as follows

$$M(c,m)=\begin{cases}1 \;\; if \;\; E(c,m)\geq\theta_{AT}(c)\; and \; C(c,m)\geq\theta_c \\ 0 \qquad\qquad otherwise\end{cases} \tag{7}$$

In the above binary mask, 1 represents speech dominant T-F units having sufficient speech energy and 0 represents noise dominant T-F unit.

D. Resynthesis

The final stage in the speech separation process is to re-synthesis the speech from the noisy mixture. In the resynthesis process, first, the response of each filter is reversed in time and passed through the same filter and time reversed again. Second, each channel output is divided into time frames by windowing with a raised cosine window, with a frame size as equal to the one used in the decomposition of input mixture into T-F units [4]. The energy in each T-F unit is then weighted by the corresponding T-F mask value obtained from the previous stage. Finally, the weighted responses are summed across all frequency channels to obtain the target speech from the noisy mixture.

III. EXPERIMENTAL RESULTS

The proposed speech separation system is evaluated with a set of 100 mixtures which contains 10 voiced utterances and 10 noises collected by Cooke [16]. The noise samples are N_0, 1-kHz pure tone; N_1, white noise; N_2, noise bursts; N_3, “cocktail party” noise; N_4, rock music; N_5, siren; N_6, trill telephone; N_7, female speech; N_8, male speech; and N_9, female speech. The noise samples are mixed with the speech signal at the desired SNR as used in Hu and Wang system [12]. The performance of the proposed system is evaluated by measuring the SNR as follows

$$SNR=10\log\left(\frac{\sum_n S(n)^2}{\sum_n (S(n)-S_{out}(n))^2}\right) \tag{8}$$

where $S(n)$ denotes the speech and $S_{out}(n)$ denotes the segregated speech by the proposed algorithm. Ten speech samples denoted as $V_0 - V_9$ are taken from Cooke [16] speech data base and it is mixed with each noise according to the input SNR value as used in the Hu-Wang [12] system. Noisy mixtures V_0N_0, V_1N_0,…..V_9N_0 are given as an input to both conventional energy selection threshold based system and the adaptive energy selection threshold based system and the results are averaged and is given in Table I. This process is repeated for all other noises and reported in Table I and its equivalent graphical representation is given in Fig. 3. The average SNR improvement across all noises is given in the last column of Table I. From the Table I, it is observed that average SNR value for the proposed adaptive energy selection threshold method is high for certain noise mixtures for example N_0, N_1, N_2, N_3, N_4 and N_5. The conventional energy selection threshold method improves the SNR value for the remaining noises.

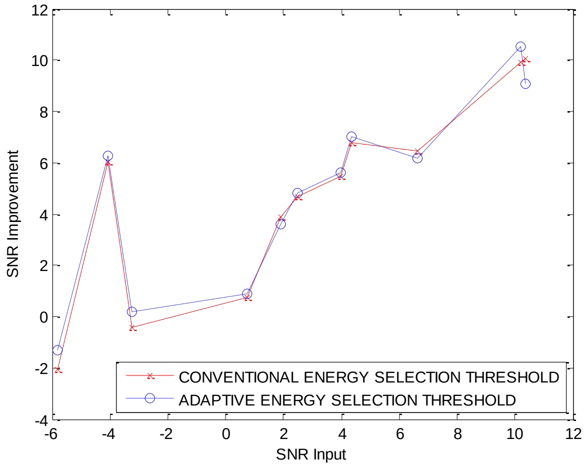

Fig. 3. Average SNR input versus SNR improvement for conventional energy selection threshold and adaptive energy selection threshold.

Finally, by comparing the SNR improvement for both the methods, the proposed adaptive energy selection threshold method shows a slight improvement as compared with the conventional energy selection threshold method

IV. Conclusion and Future work

This research work proposes a speech separation system using adaptive energy threshold selection to improve the quality of segregated speech. In the conventional energy feature based T-F unit selection method miss some of the T-F units where speech component is present but not as a dominant one. This approach degrades the speech quality for certain noises and is reported in Table I. In order to solve this issue, this research work proposed an adaptive energy based T-F unit selection for each filter channel. The proposed method improves the quality of the segregated speech in terms of SNR improvement and is reported in Table I. However, the improvement in SNR is not as expected. The future work of this research concentrates to further improve the SNR and in turn speech quality.

References

[1] J. Jensen, J.H.L. Hansen, “Speech enhancement using a constrained iterative sinusoidal model,” *IEEE Trans. Speech Audio Process.*, vol.9, 2001, pp. 731-740.

[2] Y. Ephraim and H. L. Trees, “A signal subspace approach for speech enhancement,” *IEEE Trans. Speech Audio Processing*, vol. 3, 1995, pp. 251–266.

[3] H. Sameti, H. Sheikhzadeh, L. Deng, R.L. Brennan, “HMM-based strategies for enhancement of speech signals embedded in nonstationary noise,” *IEEE Trans. Speech Audio Process.*, vol. 6, 1998, pp. 445-455.

[4] M. Weintraub, “A theory and computational model of auditory monaural sound separation,” Ph.D. dissertation, Dept. Elect. Eng., Stanford Univ., Stanford, CA, 1985.

[5] N.Harish Kumar, R.Rajavel, “Monaural speech separation system based on optimum soft mask,” *IEEE Int. Conf. on Computational Intelligence and Computing Research*, 18-20 Dec 2014.

[6] G. J. Brown and M. P. Cooke, “Computational auditory scene analysis,” *Comput. Speech Language*, vol. 8, 1994, pp. 297–336.

[7] G Hu, D Wang, “An auditory scene analysis approach to monaural speech segregation,” *Topics in Acoustic Echo and Noise Control*. (E Hansler, G Schmidt, eds.) (Springer, New York), 2006, pp. 485–515.

[8] G.J. Brown, D.L. Wang, “Separation of speech by computational auditory scene analysis,” *J. Benesty, S. Makino, J. Chen (eds.), Speech Enhancement, Berlin, Germany: Springer*, 2005, pp. 371-402.

[9] G Hu, D Wang, “Auditory segmentation based on onset and offset analysis,” *IEEE Trans. Audio Speech Lang. Process*. **15**(2), 2007, pp. 396–405.

[10] S.F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoust. Speech Signal Process*., vol. 27, 1979, pp. 113-120.

[11] K Hu, D Wang, “Unvoiced speech segregation from nonspeech interference via CASA and spectral subtraction,” *IEEE Trans. Audio Speech Lang. Process.* **19**(6), 2011, pp. 1600–1609.

[12] G Hu, D Wang, “Monaural speech segregation based on pitch tracking and amplitude modulation,” *IEEE Trans. Neural Networks.* **15**(5), 2004, pp. 1135–1150.

[13] R.D. Patterson, I. Nimmo-Smith, J. Holdsworth, P. Rice, “An efficient auditory filterbank based on the gammatone function,” MRC Applied Psych. Unit. 2341, 1988.

[14] J. Holdsworth, I. Nimmo-Smith, R.D. Patterson, P. Rice, “Implementing a gammatone filter bank,” *MRC Applied Psych. Unit*, 1988.

[15] D.L. Wang, G.J. Brown, “Separation of speech from interfering sounds based on oscillatory correlation,” *IEEE Trans. Neural Networks*, vol.10, 1999, pp. 684-697.

[16] E.H Rothauser et al., “IEEE recommended practice for speech quality measurements,” *IEEE Trans. on Audio and Electroacoustics*, vol.17, 1969, pp.225–246.