Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 13, no. 3, May 2017, pp. 259-263
DOI: 10.1555/10.23940/ijpe.17.03.p2.259263

CASA FOR IMPROVING SPEECH INTELLIGIBILITY IN MONAURAL SPEECH SEPARATION

M. Dharmalingam[a] and M. C. John Wiselin[b]*

[a]*PRIST University Thanjavur, Tamilnadu, India / Kongunadu College of Engineering and Technology, Trichy, Tamilnadu, India*

[b]*Department of EEE, Vidya Academy of Science &. Technology, Thrissur, Kerala, India*

Abstract

Speech separation is the process of separating the target speech and noise from the noisy speech mixture. Speech separation algorithms are useful in improving the quality and intelligibility of the speech. The various traditional speech separation algorithms such as spectral-subtractive algorithms, Wiener filtering, statistical model-based methods and subspace algorithms are mainly focus on improving the speech quality. But there are applications such as mobile communication, air ground communication and hearing aids, needs speech intelligibility than speech quality. In order to satisfy the requirements of intelligibility, this work proposes an algorithm using Computational Auditory Scene Analysis (CASA) and Support Vector Machine (SVM) to separate the noisy speech into target speech and noise and at the same time improves the speech intelligibility. The proposed algorithm decomposes the clean speech and noise into time-frequency units (T-F) and computes the energy from each frame of clean speech and noise to train the SVM. Latter in the testing phase, the trained SVM is used to estimate the binary mask from the energy of the noisy speech based on whether each T-F unit is dominated by speech or noise. The estimated mask by SVM is used to synthesize the speech signal and is presented to normal-hearing listeners with different age groups to measure the performance of the proposed algorithm. The experimental results show substantial improvements in recognition score because the separated speech has reasonable speech intelligibility.

Keywords: Computational Auditory Scene Analysis (CASA); Support Vector Machine (SVM); Time-frequency units (T-F)

(Submitted on December 14, 2016; Revised on March 8, 2017; Accepted on March 16, 2017)

1. Introduction

In a real scenario, when a person is speaking, his speech is getting affected by various types of noise such as car noise, train noise, restaurant noise and sometimes speech from other person speaking. For an effective speech communication, these unwanted noises have to be separated. Speech separation is a process of separating the speech and noises from the noisy speech mixture. The traditional speech separation algorithms mainly focus on improving the speech quality rather than speech intelligibility [1]. Speech intelligibility is defined as the clear understanding of the content of spoken words. This plays a vital role in variety of applications such as mobile phone, air ground communication and hearing aids. As the traditional speech separation algorithms fails to separate the noise from noisy speech and improve speech intelligibility, this work propose a new model based CASA for speech separation. Since CASA mimic human auditory scene analysis by computational means [2],[3], this paper adopts model based CASA to discard noise from the noisy mixture and improve intelligibility.

* Corresponding author. Tel.: +91-9486583937; fax: +91-4326-277572.
E-mail address: dlingam6@gmail.com or .

Fig. 1, shows the typical CASA based speech separation system. In model based CASA, the noisy speech signal undergoes auditory peripheral analysis, to extract the reliable features such as energy, amplitude modulation, frequency modulation etc [4]. These features are used in the classifier to classify the speech and noise from the noisy speech.

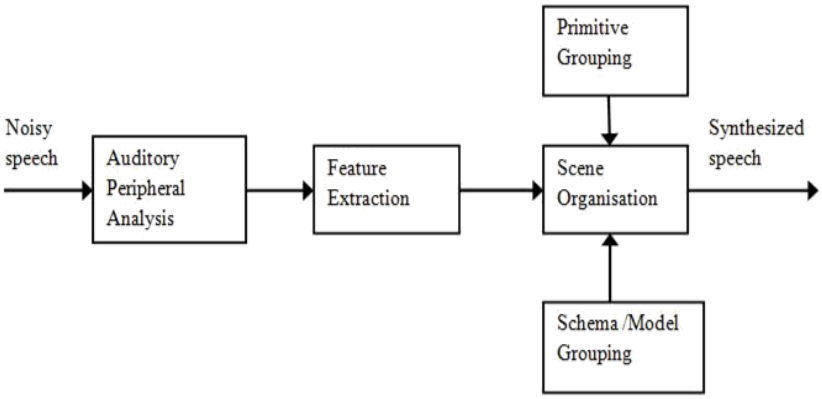

Figure 1: A typical CASA based speech separation system

The traditional algorithms used in speech separation are spectral-subtractive algorithms, subspace, Wiener filtering, statistical model based algorithm. The spectral-subtractive algorithms follow a simple principle, which estimate the spectrum of clean signal by subtracting the spectrum of noise signal from the spectrum of noisy signal. Boll [5] has proposed a spectral subtractive algorithm and performed the intelligibility and quality measurements using diagnostic rhyme test (DRT). The results indicated that speech quality was improved using their proposed spectral subtractive algorithms. Lim [6] has proposed another spectral subtractive algorithm and compared the performance. It is observed that there was no significant difference in the speech intelligibility scores and moreover the speech intelligibility is decreased somewhat by the spectral–subtractive algorithms. The main reason for the degradation of speech intelligibility using spectral-subtractive algorithms is that they are largely intrusive and heuristic. The drawbacks of spectral-subtractive algorithms are overcome by Wiener filtering algorithm, which has several applications in speech enhancement. Arehart et al [7] proposed an audible speech separation algorithm and found that small improvement in speech intelligibility with normal and hearing- impaired listeners. The Wiener filter approach yields linear estimators but another technique exists for deriving a non-linear estimator statistical– model based methods. Soon [8] proposed a self-adaptive MMSE based estimator for speech separation and the results shows an improvement in SNR. The above mentioned speech separation algorithms are based on the theory from signal processing and statistical estimation. Another algorithm for speech separation is subspace algorithm based on linear algebra theory. Kim [9] described a sub space algorithm based on masking threshold for speech separation and the results show substantial improvement in speech quality. Rezayee [10] proposed an adaptive KLT approach with non-stationary colored noise for speech separation and this approach shows a small improvement in speech intelligibility.

This work combines CASA and Support Vector Machine to classify noisy speech T-F units into noise T-F unit and speech T-F unit. The proposed model is simpler to construct and once the training is over, the testing requires minimum time to separate noisy T-F units into speech and noise dominant T-F units. Hence the computational time is less when compared to the feature based CASA which involves segmentation and grouping principles such as pitch tracking, amplitude and frequency modulation etc. In this proposed algorithm, input speech signal is subdivided into various T-F units using gamma tone filter bank. Then Energy features are extracted from the speech signal. The extracted features are given to the SVM classifier. The SVM is the machine learning algorithm which is widely used for classification problems. After classification, it is followed by waveform synthesis to get the synthesized speech. The obtained speech is highly intelligible and is presented to normal-hearing listeners with different age groups to measure the performance of the proposed algorithm. The experimental results show substantial improvements in recognition score since the separated speech has reasonable speech intelligibility. The rest of the paper is organized in the following manner. Section II describes about the proposed speech separation system. Section III shows the experimental results and Section IV includes the conclusion and future work.

2. Proposed Speech Separation Systems

The block diagram of proposed CASA and SVM based speech separation system is shown in Fig. 2. The input speech clean and noise is given to the Gamma tone filter bank [11] separately, which decomposes these signals into time-frequency units

(T-F) based on the frequency response of human cochlea. The gamma tone is a band pass filter, whose impulse response is the product of gamma function and tone as follows:

$$g(t) = At^{n-1}e^{-2\pi bt}\cos(2\pi ft + \phi) \quad (1)$$

where, 'f' in Hertz is the centre frequency, 'ϕ' in radians is the phase, 'A' is the amplitude, 'n' is the filter's order, 'b' is the filter bandwidth and 't' is time in seconds. In this work 128 band Gamma tone filter bank is used to decompose the speech and noise. The energy from each frame in each sub band is extracted as a feature using the following equation:

$$E(c,i) = g_{out}(c,i) * g_{out}(c,i) \quad (2)$$

where, 'c' represents the number of sub band which is 128 in this work, 'i' represents the number of frame which will vary from signal to signal. Therefore the energy feature would be a matrix of size c and ii. Next, the energy matrix E(c, i) is given to the SVM classifier for classification purposes. In the training stage, the energy matrix extracted from various clean speeches and noises are used to train the SVM classifier. In the SVM classifier, '1' is assigned as the target for the energy of the clean speech signal and '0' is assigned as a target for the energy of the noise. After training, a noisy speech at various SNR is given as input for testing. In testing phase, the energy of the noisy speech at various SNR are extracted and given to the SVM for classification. The classifier output is a mask of '1's and '0's of size (c, i). This mask is used in the synthesis filter bank [12] to convert the (T-F) units of noisy speech into speech and noise. This procedure is repeated for various types of speech and noise. Finally, the synthesised speech is played to normal-hearing listeners of various age groups and asked to identify the words correctly. The proposed system is computationally efficient as compared to the similar system proposed in [9]. It uses clean speech, noise and noisy speech for training and using rethresholding technique to obtain the mask for synthesizes the speech. The proposed research work use clean speech and noise for training and noisy speech at different SNRs for testing. The performance of the system is evaluated in terms of improvement in speech intelligibility.

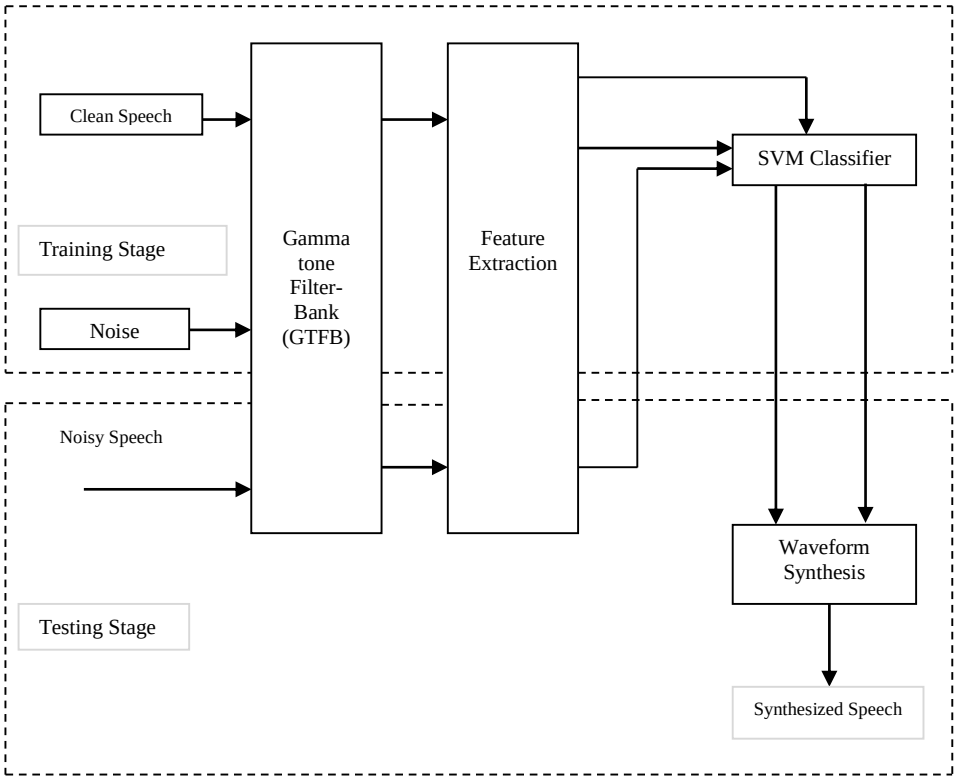

Figure 2: Block Diagram of Schema Based CASA Speech Separation System

3. Experimental Results

In this research work, the input speech is taken from the IEEE speech corpus [13] and noises from [14] collected by Cooke are used to test the performance of the proposed model based CASA system for speech separation to improve speech intelligibility. About 30 clean speech sentences and noises such as babble noise, noise bursts, white noise, rock music, male voice, female voice, siren, telephone noise, and train noise are given to the classifier for training. In the testing phase, speech

sentences with the babble noise, noise bursts, white noise and rock music at different SNRs in the range [-2.5db to 15db] are used to estimate the binary mask for speech synthesis. Finally, the synthesized speech signal is given to the normal hearing listeners at different age groups which include 15 male and 10 female for evaluation. The performance of the proposed system is evaluated via word test and intern recognition scores are obtained. The recognition score for the speech sentence "the sky that morning was clear and bright blue" with babble noise at different SNRs are shown in Table1. The recognition score for the speech sentence "how willing you marry marline" with rock music at different SNRs are shown in Table2.Similarly; the recognition score for the same sentence with noise bursts and white noise at different SNRs are shown in Table 3 and Table 4.

Table 1: Recognition scores for the sentence "the sky that morning was clear and bright blue" with babble noise

Listener	Recognition Scores (%)				
	-2.5db	0db	5db	10db	15db
Male(15listener)	68	70	85	80	85
Female(10 Listener)	72	73	79	76	90

Table 2: Recognition Scores for the sentence "how willing you marry marline" with rock music.

Listener	Recognition Scores (%)				
	-2.5db	0db	5db	10db	15db
Male(15listener)	55	63	75	80	88
Female(10 Listener)	57	68	78	85	90

Table 3: Recognition Scores for the sentence "how willing you marry marline" with noise bursts.

Listener	Recognition Scores (%)				
	-2.5db	0db	5db	10db	15db
Male(15listener)	75	77	80	85	85
Female(10 Listener)	80	82	85	88	90

Table 4: Recognition Scores for the sentence "how willing you marry marline" with white noise.

Listener	Recognition Scores (%)				
	-2.5db	0db	5db	10db	15db
Male(15listener)	75	80	88	90	90
Female(10 Listener)	80	82	85	88	90

4. Conclusion and Future work

The traditional speech separation algorithm proposed in the literature mainly focus on speech quality rather than speech intelligibility. The present work focus on improving speech intelligibility using schema based CASA speech separation system using SVM classifier. The mask estimated from the classifier is used to enhance the speech. The performance of the system is evaluated by conducting word test and scores are obtained from the listeners of various age groups are shown in Table 1-4. Results show that proposed system improves speech intelligibility under various noises. The real time implementation of this proposed system is in progress.

References

1. Brown, G. J. and Wang, D. L. "Speech Enhancement", In Benesty, J., Makino, S. and Chen, J. (Eds), „Separation of Speech by Computational Auditory Scene Analysis, Springer, New York, 2005; pp. 371-402.
2. Bregman, A. S. Auditory Scene Analysis, MIT Cambridge.,1955
3. Wang, D. L. and Brown, G. J. Computational Auditory Scene Analysis: Principles, Algorithms, and Applications, Wiley-IEEE Press, Hoboken.2006.
4. Hu, G. and Wang, D. L. Monaural Speech Segregation based on Pitch Tracking and Amplitude Modulation", IEEE Transaction on Neural Networks, 2004;Vol. 15, No. 5, pp. 1135-1150.
5. Boll. S.F. Suppression of Acoustic Noise in Speech using Spectral Subtraction, IEEE Trans .Acoust.Speech Signal Processing., 1979; 27(2),113-120.

6. Lim,J. Evaluation of a Correlation Subtraction Method for Enhancing Speech Degraded by Additive Noise , IEEE Trans.Aoust.Speech Signal Process.,1978;37(6),471-472.
7. Arehart, K., Hansen, J., Gallant, S., and Kalstein, L., Evaluation of an Auditory Masked Threshold Noise Suppression Algorithm in Normal Hearing and Hearing Impaired Listeners ,Speech Coommun.,2003; 40, 575-592.
8. Soon,I.,Koh,S.,andYeo,C., Improved Noise Suppression Filter using Self Adaptive Estimator of Probability of Speech Absence, Signal Process.,1999;75,151-159.
9. Kim,J.,Kim,S., and Yoo,C.,The Incorporation of Masking Threshold to Sub Space Speech Enhancement , proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing ,Hong Kong ,2003;Vol.I,pp.76-79.
10. Rezayee, A. And Gazor, S. An Adaptive KLT Approach for Speech Enhancement, IEEE Trans.Speech Audio Process, 2001; 9(2), 87-95.
11. Hohmann, V. „Frequency Analysis and Synthesis using a Gammatone Filter Bank", Acta Acoustica United with Acustica, 2002; Vol. 88, pp. 433 - 442.
12. Weintraub, M. A Theory and Computational Model of Auditory Monaural Sound Separation, Ph.D. Thesis, Stanford University.1985.
13. Rothauser, E. H., Chapman, W. D., Guttman, N., Hecker, M. H. L, Nordby, K. S., Silbiger, H.R.,Urbanek,G. E. and Weinstock,M. IEEE Recommended Practice for Speech Quality Measurements", IEEE Transaction on Audio Electro Acoustics,1969; Vol. 17, pp. 225-246.
14. Noisex-92, www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html.2014

Mr. M. Dharmalingam doing Ph.D. in Department of Electronics and Communication Engineering from PRIST University - Thanjavur, and Completed his ME applied Electronics from RVS College of Engineering and Technology, Dindigal. He is presently working as HOD & Associate Professor in Electronics and Communication Engineering**,** Kongunadu College of Engineering and Technology, Thottiam - 621215, Tamilnadu, India. He has 17 years of teaching experience and he has guided more than 10 PG students and more than 20 students in UG. His research interest includes signal processing, speech processing and image processing.

Dr. M. C. John Wiselin has done Ph. D. in Faculty of Electrical and Electronics Engineering (Power Systems) from Bharath University – Chennai, Tamilnadu. He is presently working as HOD & Professor in Electrical and Electronics Engineering, Vidya Academy of Science & Technology**,** Thrissur - 680501, Kerala, India. He has 13 years of teaching experience .His research interest includes signal processing, speech processing and image processing. Power system Analysis, Electric Circuit Analysis, Digital Electronics, Measurement and Instrumentation and signal processing. He has guided more than 20 M. Tech theses and one Ph.D. thesis. He is currently guiding 4 Ph.D. students for their doctoral degree.