MONAURAL SPEECH SEPARATION SYSTEM BASED ON OPTIMUM SOFT MASK

N. Harishkumar
Post Graduate Scholar
Electronics & Communication Engineering
SSN College of Engineering, Kalavakkam
Chennai, India
harishkumarsrc@gmail.com

R. Rajavel
Associate Professor
Electronics & Communication Engineering
SSN College of Engineering, Kalavakkam
Chennai, India
rajavelr@ssn.edu.in

Abstract— **The ideal binary mask (IBM) has been one of the most successful techniques in computational auditory scene analysis (CASA) algorithms. The binary value 1 is assigned to the mask if the local signal-to-noise ratio (SNR) of a particular time-frequency (T-F) units exceeds the local criterion (LC), otherwise the value 0 is assigned to the mask. This binary weighting may discard some parts of the speech during synthesis, which leads to an unnatural sound called musical noise. This paper proposes the optimum soft mask (OSM) to reduce the musical noise, by replacing the hard limiting weights (i.e., 1 or 0) with the variable weights between 0 and 1. The Signal-to-Noise ratio is used as a performance measures to compare the performance of the proposed soft mask with IBM in the context of monaural speech separation. The experimental results show the superior performance of the proposed soft mask as compared with IBM.**

Keywords—computational auditory scene analysis (CASA), ideal binary mask (IBM), signal-to-noise ratio (SNR), optimum soft mask (OSM), genetic algorithm (GA).

I. INTRODUCTION

In real environment, if a person is speaking, the speech will get affected by the surrounding noise for example, other people speaking/shouting, passing a car, and many natural sounds. Several applications require a system that separates the intended speaker's speech from the noisy speech signal. For example, voice communication over cellular phone will get affected by the surrounding noise present at the transmitting end. In an air-ground communication, the pilot speech will be affected by the high level of cockpit noise. In a teleconferencing system, the noise in one location will be broadcasted to all other locations. In the literature, two main approaches are proposed for speech separation. They are, speech enhancement, and blind source separation (BSS). The speech enhancement approach is well suited when the noise is stationary and is not when the noise is non-stationary. However, BSS is the successful speech separation approach for both stationary and non-stationary noisy conditions. Computational auditory scene analysis (CASA) is the best technique in BSS separation. CASA aims to build sound separation systems that adhere to the known principles of human hearing [1]. The CASA based speech separation system employing IBM shows large benefits in speech intelligibility even at low SNR level (-5 dB, -10 dB) [1]. However, a problem with IBM in speech separation applications is employing binary weights (i.e., 0 or 1). This binary weighting may cause some parts of the speech to be discarded during synthesis. This introduces an unnatural sound called musical noise. This paper, proposes a technique to reduce the impact of musical noise, by replacing the binary weights (i.e., 1 or 0) by the variable weights called soft mask. The binary Genetic Algorithm (GA) is used in this work to find the optimum soft mask having values between 0 and 1. The Signal-to-Noise ratio is used as a measure to evaluate the performance of the proposed soft mask with the existing IBM [2] in the context of monaural speech separation. The rest of the paper is organized in the following manner. Section II gives an overview of the CASA with IBM and its short comings. Section III presents the proposed optimum soft mask based speech separation system. Section IV shows the experimental results and finally section V concludes the paper with possible future extension.

II. COMUTATIONAL AUDITORY SCENE ANALYSIS AND IBM

CASA is defined as the study of ASA in computational means [3]. The typical structure of CASA based speech separation system is shown in *Fig. 1*. The input mixture having both speech and noise signal has been processed to extract the features by the front end speech separation system. Some CASA system directly performs grouping based on the features extracted [5] but many systems [6] use the intermediate stage to separate the speech and noise signals from the mixture. The ideal binary mask (IBM) is one of the successful approaches in CASA based speech separation [3]. The IBM is framed based on the intermediate T-F representation. Employing IBM shows large benefits in speech intelligibility even at low SNR level. The value of ideal binary mask is either 0 or 1 and these values are obtained based on the energy in the corresponding T-F units of speech and noise. The IBM is defined as in (1) [2, 3],

$$\mathrm{IBM}(t,f) = \begin{cases} 1, & \text{if } s(t,f) - n(t,f) > LC \\ 0, & \text{Otherwise} \end{cases} \qquad (1)$$

978-1-4799-3975-6/14/$31.00 ©2014 IEEE

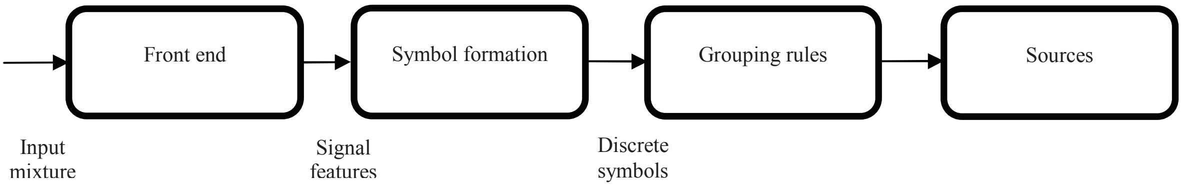

Fig. 1. Typical data driven CASA based speech separation system [4].

Where $s(t, f)$ denotes the speech energy in dB and $n(t, f)$ denotes the noise energy in dB. If the difference between signal energy and noise energy is greater than LC, the binary value 1 is assigned to IBM, otherwise 0 is assigned. The generated IBM is used in the synthesis stage to separate input mixture into speech and noise. However, the problem with IBM in speech separation applications is the binary weighting which causes some parts of the speech to be discarded during synthesis [7]. This introduces an unnatural sound called musical noise. To reduce the impact of this musical noise, the binary weights are replaced by the variable weights (soft mask) between 0 and 1. The binary Genetic Algorithm (GA) is used in this proposed work to find the optimum values for the soft mask between 0 and 1.

III. PROPOSED OPTIMUM SOFT MASK (OSM)

The musical noise is one of the unavoidable problems in IBM based speech separation, which degrades the quality of the synthesized speech. To solve this issue, this paper proposes a method using soft mask. The soft mask is obtained using genetic algorithm based optimization approach; hence it is named as optimum soft mask. The proposed optimum soft mask (OSM) based speech separation system contains two stages. The first stage is the optimum mask estimation stage, in which the mask is estimated using binary GA is shown in *Fig. 2*. Second stage is the synthesis stage, in which the estimated mask is applied to extract the speech from the input mixture as shown in *Fig. 4*. The speech and noise signals are given as an input to the Gamma tone filter bank (peripheral analysis) to decompose the signals in two dimensional T-F representations [8]. After this decomposition, first, the energy of the speech and noise is measured in dB and then the ground truth SNR (GTSNR) is calculated as in (2) [9].

$$\mathrm{GTSNR}_{\mathrm{T-F}}=10\log\left(\frac{\sum_{n}(\mathrm{S}_{\mathrm{T-F}}(n))^{2}}{\sum_{n}(\mathrm{N}_{\mathrm{T-F}}(n))^{2}}\right) \qquad (2)$$

Here, $\mathrm{S}_{\text{T-F}}$ and $\mathrm{N}_{\text{T-F}}$ represents the T-F unit energy of speech and noise signals respectively and n is the time index. The SNR value of a particular T-F unit near 0 dB has a great impact on the speech intelligibility [9], hence in this paper, SNRs in the range between -5 dB to 0 dB has been divided into several intervals. The next stage is the estimation of optimum soft mask. The optimum soft mask is formulated as in (3), where x1, x2 and x3 are the variables to be optimized. The flowchart representation [12] using Binary GA [10] to optimize the variables is shown in *Fig. 3.*

$$\mathrm{OSM}(t,f)=\begin{cases}0, & \text{if } \mathrm{GTSNR}_{\text{T-F}}<-5\,\mathrm{dB},\\ x1, & \text{if } -5\,\mathrm{dB}\le\mathrm{GTSNR}_{\text{T-F}}<-3\,\mathrm{dB},\\ x2, & \text{if } -3\,\mathrm{dB}\le\mathrm{GTSNR}_{\text{T-F}}<-1\,\mathrm{dB},\\ x3, & \text{if } -1\,\mathrm{dB}\le\mathrm{GTSNR}_{\text{T-F}}<0\,\mathrm{dB},\\ 1, & \text{if } 0\,\mathrm{dB}\le\mathrm{GTSNR}_{\text{T-F}}.\end{cases} \qquad (3)$$

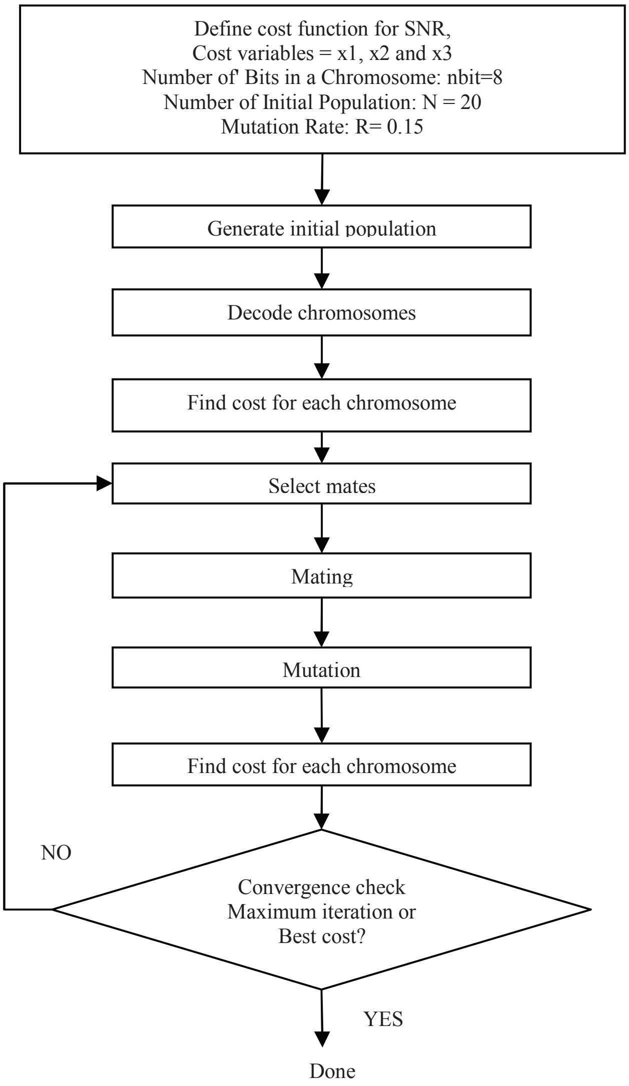
Define cost function for SNR,
Cost variables = x1, x2 and x3
Number of' Bits in a Chromosome: nbit=8
Number of Initial Population: N = 20
Mutation Rate: R= 0.15
Generate initial population
Decode chromosomes
Find cost for each chromosome
Select mates
Mating
Mutation
Find cost for each chromosome
NO
Convergence check
Maximum iteration or
Best cost?
YES
Done

Fig. 3. Flowchart of a binary GA optimization of variables in soft mask

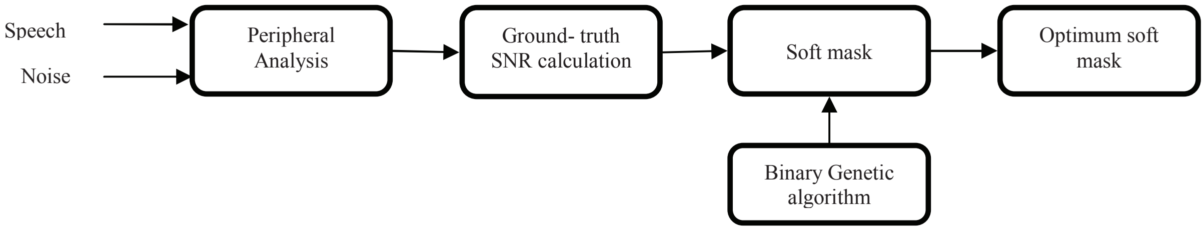

Fig. 2. Estimation of optimum soft mask using binary GA

At the end of the convergence, the GA returns optimum value for x1, x2 and x3. Based on these values the optimum soft mask is calculated as in (3). In the synthesis stage, first, the input mixture is decomosed into T-F units by gammatone filterbank (Peripheral analysis). And then, the calculated optimum soft mask using binary GA is applied in the synthesis process to suppress the noise and enhance the speech signal.

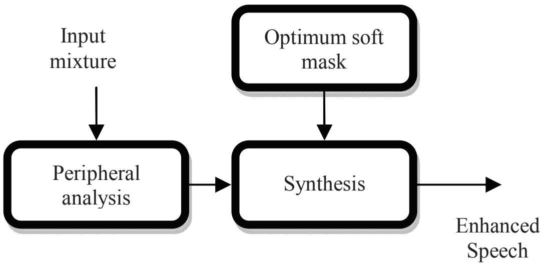

Fig. 4. Synthesis of noisy speech using optimum soft mask

The performance of the proposed optimum soft mask (OSM) in the context of monaural speech separation is evaluated using SNR improvement. The SNR improvement [9, 13] is defined as in (4),

$$\mathrm{SNR}=10\log\left(\frac{\sum_{n} S_{\mathrm{oneall}}(n)^{2}}{\sum_{n}\left(S_{\mathrm{oneall}}(n)-S_{\mathrm{out}}(n)\right)^{2}}\right) \quad (4)$$

Where $S_{\mathrm{Oneall}}(n)$ denotes the synthesized speech with all one mask and $S_{\mathrm{out}}(n)$ denotes the synthesized speech with optimum soft mask. The speech signals to conduct the experiment are taken from the IEEE corpus [11], and the noise signals are taken from Noisex-92 [14]. The noisy speech signals are obtained by mixing the clean speech signal with the noise at different SNRs (-5 dB to +5 dB).

The IBM and the proposed OSM are applied to the noisy mixture signals at the synthesis module to yield the enhanced speech signal. Table 1 and 2 shows the SNR improvement of IBM and OSM for the speech "The sky that morning was clear and bright blue" with the babble noise and factory noise respectively. Table 3 and 4 shows the SNR improvement for the speech "A large size in stockings is hard to sell" with the babble noise and factory noise respectively. The last column in all the tables shows the average SNR improvement at all SNRs. The results revels that, the SNR improvements from OSM and IBM is close together, and the amount of average SNR improvement from OSM is slightly greater than that obtained from IBM. Moreover, the OSM processed speech signal has the best quality (human listening test conducted locally), compared with those speech signals obtained using IBM as the processing mask.

Table 1. SNR improvement of ideal binary mask (IBM) and optimum soft mask (OSM) for the clean speech "The sky that morning was clear and bright blue" with babble noise.

INPUT SNR (dB)	-5	-2.5	0	2.5	5	Average SNR improvement (dB)
IBM	5.8788	6.8294	8.3496	9.8052	11.4470	8.462
OSM	5.9153	6.9983	8.3893	9.8311	11.4500	8.5168

Table 2. SNR improvement of ideal binary mask (IBM) and optimum soft mask (OSM) for the clean speech "The sky that morning was clear and bright blue" with factory noise.

INPUT SNR (dB)	-5	-2.5	0	2.5	5	Average SNR improvement (dB)
IBM	7.0591	8.3628	9.7248	11.2300	12.8103	9.8374
OSM	7.0890	8.3554	9.7428	11.2660	12.9910	9.88764

Table 3. SNR improvement of ideal binary mask (IBM) and optimum soft mask (OSM) for the clean speech "A large size in stockings is hard to sell" with babble noise.

INPUT SNR (dB)	-5	-2.5	0	2.5	5	Average SNR improvement (dB)
IBM	6.0810	7.4487	8.8684	10.3500	11.9454	8.9387
OSM	6.1275	7.4823	8.9070	10.3960	11.9650	8.97556

Table 4. SNR improvement of ideal binary mask (IBM) and optimum soft mask (OSM) for the clean speech "A large size in stockings is hard to sell" with factory noise.

INPUT SNR (dB)	-5	-2.5	0	2.5	5	Average SNR improvement (dB)
IBM	6.5004	7.4292	8.9088	10.6987	12.6542	9.23826
OSM	6.6086	7.3598	8.9084	10.7200	12.6610	9.25156

V. SUMMARY

The ideal binary mask has been one of the most successful techniques in CASA algorithm. However, a problem with IBM is musical noise due to the binary weighting. This paper proposed an optimum soft mask for monaural speech separation. The binary GA is used to obtain the optimum soft mask. This optimum soft mask is used to enhance the speech signal at synthesis stage. The performance of the proposed system is evaluated in terms of the SNR improvement. The experimental results are tabulated for various clean speech and noise signals. The results show that the proposed mask improves the SNR slightly as compared to the standard IBM. At some SNRs the proposed mask performs poorer than IBM. But still the overall SNR improvement is better than IBM. This paper could be extended in near future to implement the optimum soft mask based speech separation system in digital signal processor for real – time demonstration of the proposed soft mask.

REFERENCES

[1] N. Li and P.C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: implications for noise reduction," The Journal of the Acoustical Society of America, vol. 123, pp.1673-1682, March 2008.

[2] D.L. Wang, "Time–Frequency Masking for Speech Separation and Its Potential for Hearing Aid Design," Trends in Amplification, Vol. 12, pp.332-353, December 2008.

[3] D.L. Wang, G.J. Brown, "Fundamentals of Computational Auditory Scene Analysis," in Computational Auditory Scene Analysis, D.L Wang and G.J Brown, Wiley-IEEE Press, 2006, pp. 1-38.

[4] G.J. Brown and D.L. Wang, "Separation of Speech by Computational Auditory Scene Analysis," in Speech Enhancement, J. Benesty, S. Makino and J. Chen, New York: Springer, 2005, pp. 371–402.

[5] G. Hu and D. L. Wang, "Auditory Segmentation Based on Onset and Offset Analysis," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, pp. 396-405, February 2007.

[6] D.L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in Speech Separation by Humans and Machines, P. Divenyi. Norwell MA: Kluwer Academic, 2005, pp. 181-197.

[7] S. Cao, L. Li and X. Wu, "Improvement of intelligibility of ideal binary-masked noisy speech by adding background noise," The Journal of the Acoustical Society of America, vol. 129, pp. 2227-2236, April 2011.

[8] V. Hohmann, "Frequency analysis and synthesis using a Gammatone filterbank," Acta Acustica united with Acustica, Vol. 88, pp. 433 – 442, January 2002.

[9] Masoud Geravanchizadeh and Reza Ahmadnia, "Monaural Speech Enhancement Based On Multi-threshold Masking," In Blind Source Separation, G.R. Naik, W. Wang, Springer Berlin Heidelberg, 2014, pp.369-393.

[10] Randy L. Haupt and Sue Ellen Haupt, Practical Genetic Algorithms, 2nd ed., A John Wiley & Sons, 2004, pp. 27-50.

[11] E.H Rothauser et al. "Ieee recommended practice for speech quality measurements," IEEE Transactions on Audio and Electroacoustics, vol.17, pp.225–246, September 1969.

[12] M.S. Arifianto, A. Chekima, L. Barukang, M.Y. Hamid, "Binary genetic algorithm assisted multiuser detector for STBC MC-CDMA," wireless and optical communication networks, 2007, pp. 1-5.

[13] K. Hu and D.L. Wang, "An Unsupervised Approach to Cochannel Speech Separation," IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, pp. 122-131, January 2013.

[14] Noisex-92. http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html. 2014.