Optimizing the Objective Measure of Speech Quality in Monaural Speech Separation

M. Dharmalingam, M.C. John Wiselin and R. Rajavel

Abstract Monaural speech separation based on computational auditory scene analysis (CASA) is a challenging problem in the field of signal processing. The Ideal Binary Mask (IBM) proposed by DeLiang Wang and colleague is considered as the benchmark in CASA. However, it introduces objectionable distortions called musical noise and moreover, the perceived speech quality is very poor at low SNR conditions. The main reason for the degradation of speech quality is binary masking, in which some part of speech is discarded during synthesis. In order to address this musical noise problem in IBM and improve the speech quality, this work proposes a new soft mask as the goal of CASA. The performance of the proposed soft mask is evaluated using perceptual evaluation of speech quality (PESQ). The IEEE speech corpus and NOISEX92 noises are used to conduct the experiment. The experimental results indicate the superior performance of the proposed soft mask as compared to the traditional IBM in the context of monaural speech separation.

Keywords Monaural speech separation · Computational auditory scene analysis · Perceptual evaluation of speech quality · Ideal binary mask · Optimum soft mask

M. Dharmalingam (✉)
PRIST University, Thanjavur, Tamilnadu, India
e-mail: dharmalingamrandd@gmail.com

M.C. John Wiselin
Department of EEE, Travancore Engineering College, Kollam, Kerala, India
e-mail: dr.wiselin16@gmail.com

R. Rajavel
Department of ECE, SSN College of Engineering, Chennai, India
e-mail: rajavelr@ssn.edu.in

A. Nagar et al. (eds.), *Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics*, Smart Innovation, Systems and Technologies 43, DOI 10.1007/978-81-322-2538-6_56

1 Introduction

In day to day life, the human auditory system receives number of sounds, in which some sounds may be useful and others are not. The speech separation problem in digital signal processing is to separate out the target speech signal from the other unwanted interferences. Human auditory system handle this complex source separation problem easily, whereas it is very difficult for the machine to perform the same as human beings. However, the acoustic interferences (music, telephone ringing, passing a car and other people speaking/shouting etc.) in a natural environment is often unavoidable. Reducing the impact of acoustic interference on the speech signal may be useful in a number of applications, such as voice communication, speaker identification, digital content management, teleconferencing system and digital hearing aids. Several approaches have been proposed in the last two decades for monaural speech separation, such as speech enhancement, blind source separation (BSS), model-based and feature-based CASA.

Speech enhancement approaches [1] utilize the statistical properties of the signal to separate the speech under stationary noisy conditions. Blind source separation is an another signal processing technique for speech separation. BSS can be done using independent component analysis (ICA) [2], spacial filtering, and nonnegative matrix factorization (NMF) [3]. BSS technique fails to separate the target speech signal effectively in the case where trained basis functions of two speech source overlaps [4]. Computational auditory scene analysis (CASA) is the most successful technique among these approaches to monaural speech separation. It aims to achieve human performance in auditory scene analysis [ASA] [5] by using one or two microphone recording of the acoustic scene generally called acoustic mixture [6]. CASA based speech separation techniques can be divided into model-based and feature-based techniques [7]. The model-based CASA techniques use trained models of the speaker to separate the speech signal [8]. Feature-based technique transform the observed signal into a relevant feature space and then it is segmented into cells, which are grouped into two main streams based on cues [7]. Generally in CASA based speech separation system, the noisy speech signal will be decomposed into various T-F units to decide whether a particular T-F unit should be designated as target or interference. After T-F decomposition, a separation algorithm will used to estimate the binary T-F mask based on the signal and noise energy. This binary mask is used in the synthesis process to convert the T-F representation into target speech and background noise. DeLiang Wang suggested that the IBM can be considered as a computational goal of CASA [9]. It is basically a matrix of binary numbers which is set to one when the target speech energy exceeds the interference energy in the T-F unit and zero otherwise. Even though, IBM is the optimal binary mask [3], it introduces objectionable distortions, called musical noise. It is mainly due to the repeated narrow-frequency-band switching [6] and moreover the perceived quality of binary-masked speech is poor. In order to address this musical noise problems in IBM, this work propose a genetic algorithm based optimal soft mask (GA-OSM) as the goal of CASA. The rest of the paper is organized in the following manner. The next section provides an overview of IBM and its shortcomings. Section 3 presents the proposed GA-optimum soft mask as the computational goal of CASA. Systematic evaluations and experimental results are provided in Sect. 4. Finally, Sect. 5 summarizes the research work and gives the direction for future extension.

2 The Ideal Binary Mask and Its Shortcomings

DeLiang Wang proposed ideal binary mask as a computational goal in CASA algorithms [9]. The IBM is a two-dimensional matrix of binary numbers and it is determined as in [9] by comparing the power of target signal to the power of masker (interfering) signal for each T-F unit obtained using Gammatone filter bank [10].

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

where $s(t,f)$ is the power of the speech signal, $n(t,f)$ is the power of the noise signal, at time t and frequency f respectively and LC is the local SNR criterion [9]. Only the T-F units with local SNR exceeding LC are assigned the binary value 1 in the binary mask and others are assigned zero [10]. In CASA, an LC value of 0 dB is commonly used, since it shows higher speech intelligibility even at low SNR levels (−5 dB, −10 dB) [11]. In IBM based speech separation, T-F units with binary value 1 are retained, and with value 0 are discarded. The region with binary value 0 is generally interpreted as the deep artificial gap and it is being discarded during synthesis and produce musical noise at the output.

3 Proposed Genetic Algorithm Based Optimum Soft Mask

Research results show the musical noise arising from binary mask can be reduced by using soft mask [12]. However, the choice of soft mask should be made carefully such that it does degrade the quality of the speech signal [6]. Cao et al. [13] has proposed a kind of soft mask by filling the artificial gaps with un-modulated broadband noise. The un-modulated broadband noise shallows the areas of artificial gaps in the time-frequency domain of the IBM processed speech mixture and improves the speech quality. In this work, rather than adding additional broadband noise, the T-F units with local SNR less than LC are filled with certain amount of unvoiced speech to enhance the speech quality. Here, a simple question comes, how much amount of unvoiced speech can be added to get better speech quality?. This motivates to use the Genetic algorithm to find the optimum amount of unvoiced speech to be added to improve the speech quality. The schematic of the proposed GA-optimum soft mask based speech separation system is shown in Fig. 1.

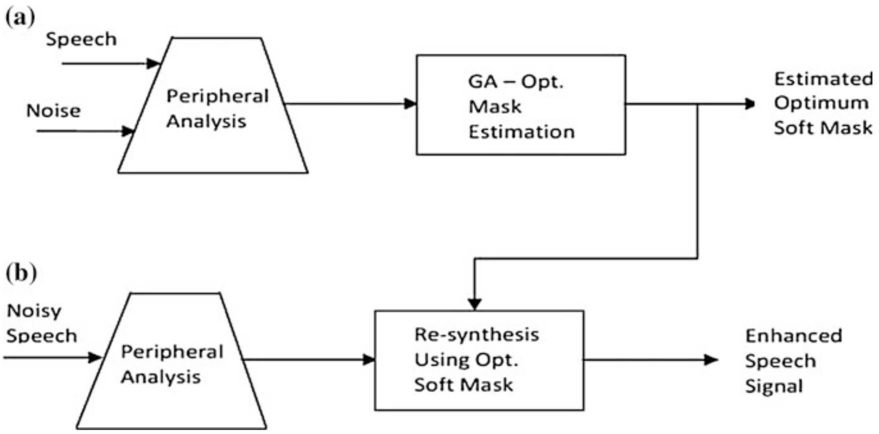

Fig. 1 Proposed GA-optimum soft mask based speech separation system

The input speech signal is first decomposed into various T-F units by a bank of 128 Gammatone filters, with their center frequencies equally distributed on the equivalent rectangular bandwidth (ERB) rate scale from 50 to 4000 Hz. The impulse response of the gammatone filter is given as [14]

$$g_{f_c}(t) = At^{N-1}\exp[-2\pi b(f_c)]\cos(2\pi f_c t + \phi)u(t) \tag{2}$$

where A is the equal loudness based gain, $N = 4$ is the order of the filter, b is the equivalent rectangular bandwidth, f_c is the center frequency of the filter, ϕ is the phase, and $u(t)$ is the step function. In each band, the filtered output is divided into a time frame of 20 ms with 10 ms overlapping between consecutive frames. As a result of this process, the input speech is decomposed into a two-dimensional time-frequency representation $s(t,f)$. Similarly, the noise signal also decomposed into a two-dimensional time-frequency representation $n(t,f)$. The proposed soft mask is defined as

$$GA - OSM(t,f) = \begin{cases} 1, & \text{if } s(t,f) - n(t,f) > LC \\ x & \text{otherwise} \end{cases} \tag{3}$$

where $GA - OSM(t,f)$ is the optimum soft mask. The GA is used here to find the optimum value of x which improves the speech quality. The GA frame work as similar as in [15] to find the optimum value of x is explained as follows:

Step-1 : Initialization: Generate a random initial population N_{POP} of size $[N \times 1]$ for the best value of x in Eq. (3). Where $N = 20$, i.e. N_{POP} = [20 × 1] matrix of chromosomes.

Step-2 : Fitness Evaluation: Fitness of all the solutions $x_1, x_2, x_3, \ldots\ldots x_N$ in the population N_{POP} is evaluated. The steps for evaluating the fitness of a solution are given below:

Step-2a: Determine the signal power at time t and frequency f and denote it as $s(t,f)$.

Step-2b: Determine the noise power at time t and frequency f and denote it as $n(t,f)$.

Step-2c: Compute the optimum soft mask $GA - OSM(t,f)$ as

$$GA - OSM(t,f) = \begin{cases} 1, & \text{if } s(t,f) - n(t,f) > LC \\ x_i & \text{otherwise} \end{cases} \tag{4}$$

where $x_i \in x_1, x_2, x_3, \ldots\ldots x_N$

Step-2d: Synthesize the speech signal using the computed soft mask as defined in step 2c.

Step-2e: The PESQ (fitness value) is calculated.

Step-2f: The steps 2c–2e are repeated for all solutions in the population.

Step-3: Updating Population: The populations are updated via mating and mutation procedure of Genetic algorithm.

Step-4: Convergence: Repeat steps 2–3 until an acceptable solution is reached or number of iteration is exceeded. At this point the algorithm should be stopped.

The final solution of this GA algorithm gives the best value of x in Eq. (3) and hence the optimum soft mask. This estimated optimum soft mask is used in the online speech separation stage to resynthesize the speech signal.

4 Performance Evaluations and Experimental Results

4.1 Experimental Database and Evaluation Criteria

The clean speech signals are taken from the IEEE corpus [16] and noise signals are taken from the Noisex92 database [17]. To generate noisy signals, clean speech signals are mixed with the babble and factory noises at different SNRs. The performance of the proposed optimum soft mask and IBM is assessed by using PESQ value, since PESQ measure is the one recommended by ITU-T for speech quality assessment [1, 18].

4.2 Performance Evaluation of GA-OSM Versus IBM

The clean speech signal and the babble noise are used to estimate the values of x in Eq. (3) and hence the optimum soft mask. Later, the clean speech and noise signals are manually mixed at SNRs in the range of −5 to 10 dB. The IBM computed using Eq. (2) and GA—optimum soft mask using Eq. (3) are applied to the mixture signals after T-F decomposition by the Gammatone analysis filter bank. Finally, the IBM and the GA-OSM weighted responses are processed by the Gammatone synthesis filterbank to yield an enhanced speech signal.

Figures 2, 3 and 4 show the PESQ value obtained by processing mixture signals using the proposed GA-OSM and IBM for the babble and factory noise respectively. As it can be seen from the figure, the GA-OSM processed signal has the best speech quality, compared to the IBM processed noisy signals. The average PESQ improvement for the speech signal "The sky that morning was clear and bright blue" [16] with babble noise is 0.4253 and with factory noise is 0.3871. The PESQ improvement for the speech signal "A large size in stocking is hard to sell" with babble noise is 0.4551 and with factory noise is 0.3673. Similarly, for the speech

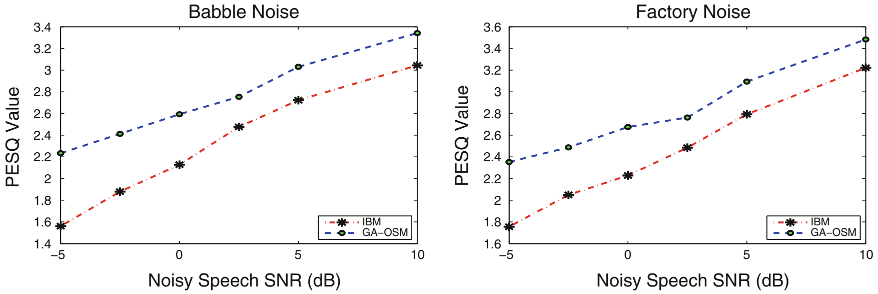

Fig. 2 Average PESQ values obtained by IBM and GA-OSM for the sentence "The sky that morning was clear and bright blue" [16] at different input SNRs and noise types

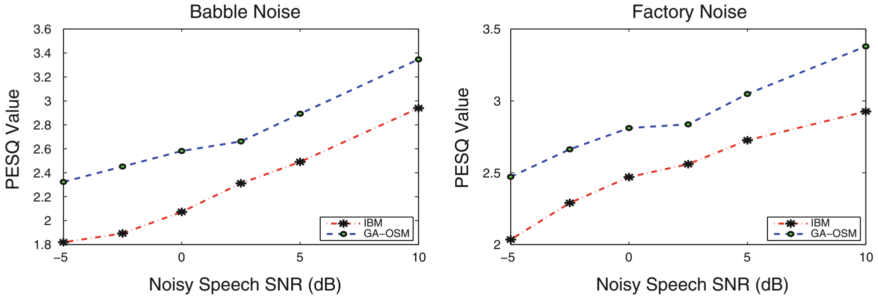

Fig. 3 Average PESQ values obtained by IBM and GA-OSM for the sentence "A large size in stocking is hard to sell" [16] at different input SNRs and noise types

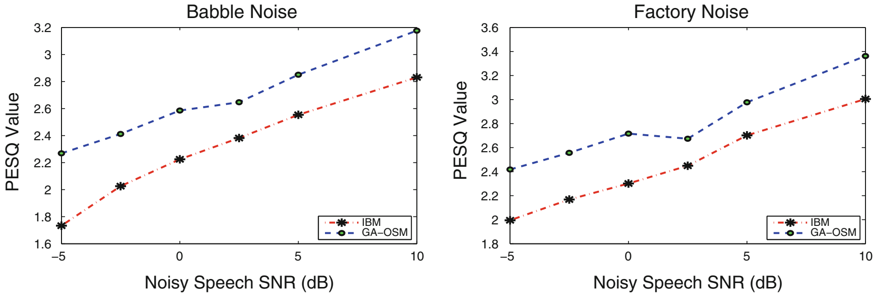

Fig. 4 Average PESQ values obtained by IBM and GA-OSM for the sentence “Sunday is the best part of the week” [16] at different input SNRs and noise types

signal “Sunday is the best part of the week” [16] the PESQ improvement with babble noise is 0.3649 and with factory noise is 0.3476. Moreover, the proposed GA-OSM mask estimation method use the clean speech signal “The sky that morning was clear and bright blue” [16] with babble noise to estimate the optimum soft mask using GA. This estimated soft mask is later used to separate and evaluate the performance of the system with other type of noise knows as factory noise. The results in Figs. 2, 3 and 4 illustrate the generalization ability of the proposed GA-OSM mask estimation approach to unseen speech and noises.

5 Summary and Future Work

Monaural speech separation is one of the challenging problem in the field of signal processing. Various approaches had been proposed including speech enhancement, Wiener filtering, noise tracking, BSS, CASA and so on. CASA based speech separation is the best among these techniques and sets the IBM as the computational goal. However, it introduces objectionable distortions called musical noise and degrades the quality of speech signal. In order to address this musical noise problem in IBM, this work proposed a genetic algorithm based optimal soft mask (GA-OSM) as the goal of CASA. The PESQ measure is used to examine the quality of the speech signal in the framework of monaural speech separation system. The experimental results in Fig. 2, 3 and 4 show the superior performance of the optimal soft mask (GA-OSM) as compared to the traditional IBM based speech separation system. However, the proposed GA-OSM estimation algorithm in its current form requires the prior knowledge of the clean speech and noise signals. This is one of the limitations of the current proposed algorithm. Further investigation of estimating the soft mask without the prior knowledge of speech and noise signal for monaural speech separation is in progress.

References

1. Loizou, P.C.: Speech Enhancement: Theory and Practice, 2nd edn, CRC Press (2013)
2. Naik, G.R., Kumar, D.K.: An over view of independent component analysis and its applications. Informatica **35**, 63–81 (2011)
3. Grais, E., Erdogan, H.: Single channel speech music separation using nonnegative matrix factorization and spectral masks. In: The 17th International Conference on Digital Signal Processing, pp. 1–6. Island of Corfu, Greece (2011)
4. Jang, G.J., Lee, T.W.: A probabilistic approach to single channel source separation. In: Proceedings of Adv. Neural Inf. Process. System, pp. 1173–1180 (2003)
5. Bregman, A.S.: Auditory Scene Analysis. MIT Press, Cambridge (1990)
6. Christopher, H., Toby, S., Tim B.: On the Ideal Ratio Mask as the Goal of Computational Auditory Scene Analysis. In: Naik, G. R., Wang, W. (eds) Blind Source Separation Advances in Theory, Algorithms and Applications. Signals and Communication Technology, pp. 369–393. Springer-Verlag, Heidelberg (2014)
7. Radfar, M.H., Dansereau, R.M., Chan, W.Y.: Monaural speech separation based on gain adapted minimum mean square error estimation. J. Sign. Process Syst. **61**, 21–37 (2010)
8. Mowlaee, P., Saeidi, R., Martin, R.: Model-driven speech enhancement for multisource reverberant environment. In: Theis, F., Cichocki, A., Yeredor, A., Zibulevsky, M. (eds.) Latent Variable Analysis and Signal Separation. Lecture Notes in Computer Science, vol. 7191, pp. 454–461. Springer-Verlag, Heidelberg (2012)
9. Wang, D.: On ideal binary mask as the computational goal of auditory scene analysis. In: Divenyi, P. (ed.) Speech Separation by Human and Machines, pp. 181–197. Kluwer Academic, Norwell (2005)
10. Geravanchizadeh, M., Ahmadnia, R.: Monaural Speech Enhancement Based on Multi-threshold Masking. In: Naik, G. R., Wang, W. (eds) Blind Source Separation Advances in Theory, Algorithms and Applications. Signals and Communication Technology, pp. 369–393. Springer-Verlag, Heidelberg (2014)
11. Li, N., Loizou, P.C.: Factors influencing intelligibility of ideal binary-masked speech: implications for noise reduction. J. Acoust. Soc. Am. **123**(3), 1673–1682 (2008)
12. Araki, S., Sawada, H., Mukai, R. Makino, S.: Blind sparse source separation with spatially smoothed time-frequency masking. In: International Workshop on Acoustic, Echo and Noise Control, Paris (2006)
13. Cao, S., Li, L., Wu, X.: Improvement of intelligibility of ideal binary-masked noisy speech by adding background noise. J. Acoust. Soc. Am. **129**, 2227–2236 (2011)
14. Patterson R.D., Nimmo-Smith, I., Holdsworth J.: Rice P : An Efficient Auditory Filter bank Based on the Gammatone Function. Report No. 2341, MRC Applied Psychology Unit, Cambridge (1985)
15. Rajavel, R., Sathidevi, P.S.: A new GA optimised reliability ratio based integration weight estimation scheme for decision fusion audio-visual speech recognition. Int. J. Sig. Imaging Syst. Eng. **4**(2), 123–131 (2011)
16. Rothauser, E.H., Chapman, W.D., Guttman, N., Hecker, M.H.L., Nordby, K.S., Silbiger, H.R., Urbanek, G.E., Weinstock, M.: Ieee recommended practice for speech quality measurements. IEEE Trans. Audio Electro Acoust. **17**, 225–246 (1969)
17. Noisex-92, http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html
18. ITU-T: Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs, Series P: Telephone Transmission Quality Recommendation P.862, ITU, 1.4. (2001)