I J C T A, 9(34) 2016, pp. 129-139
© International Science Press

Optimum Soft Mask for Monaural Speech Separation System

M. Dharmalingam[1] and M. C. John Wiselin[2]

ABSTRACT

Monaural speech separation aims to separate the target speech from speech mixture recorded by single mike. The ideal binary mask (IBM) has been projected as a procedure goal in computational auditory scene analysis (CASA) based monaural speech separation. The IBM is essentially a matrix of binary numbers, the binary value 1 is allotted to the mask if the native signal-to-noise ratio (SNR) of a specific time-frequency (T-F) units exceeds the native criterion (LC) otherwise the value 0 is allotted to the mask. The IBM based speech synthesis could discard some components of the speech and leads to associate unnatural sound known as musical noise. This analysis work proposes the optimum soft mask (OSM) as an alternate to IBM to cut back the musical noise, by commutation the arduous limiting weights (i.e. 1 or 0) with the variable weights between 0 and 1. The IEEE speech corpus and NOISEX92 noises are wont to appraise the performance of projected optimum soft mask in terms of signal-to-noise ratio (SNR) and also the sensory activity analysis of speech quality (PESQ). The experimental results indicate the superior performance of the projected optimum soft mask as compared to the IBM and ideal multi-threshold mask (IMM) within the context of monaural speech separation.

Keywords: Monaural Speech Separation; Optimum Soft Mask; Ideal Binary Mask; Computational Auditory Scene Analysis; Ideal Multi-threshold Mask.

1. INTRODUCTION

Naturally human auditory system receives the speech signal along with some surrounding noise. The noise can be of any form, for example, sound by a passing car, other people speaking and shouting etc. Several applications require a speech separation system that separates the intended speaker's speech from the noisy signal. For example, voice trans-mission over cellular phone will get affected by the surrounding noise present at the trans-mitting end. Speech separation system can be used at receiving end to improve the quality of the speech. In an air-ground communication, the pilot speech will be affected by the high level of cockpit noise. A speech separation system can be used to remove the cockpit noise considerably from the pilot's speech to improve the speech intelligibility. In hearing aid design, the persons having hearing loss will feel difficult to understand the speech in noisy conditions. In such a situation, a speech separation system can be used before amplification to remove the noise from the speech signal. In a teleconferencing system, the noise in one location will be broadcasted to all other locations. Hence there is a need for a speech sepa-ration system to block the noise from being broadcasted to all other locations. Because of its importance in many applications, speech separation is broadly studied in signal process-ing field. There are several methods used for speech separation. The most popular methods are, blind source separation (BSS)[2] and spatial filtering[12]. But both these methods are in need of more than one sensor for separating the speech from noise. i.e., it works well in binaural conditions. But in many practical applications multiple sensors are not possible. For example, telecommunication and audio retrieval uses only one sensor and requires a monaural solution[13]. However, monaural speech separation is a challenging problem, but the human auditory

[1] PRIST University. Thanjavur, Tamilnadu, India, Associate Professor/ ECE, Kongunadu College of Engineering and Technology, Trichy, Tamilnadu, India, *E-mail: dlingam6@gmail.com*

[2] Department of EEE, Vidya Academy of Science & Technology, Thrissur, Kerala- 680501, *E-mail: dr.wiselin16@gmail.com*

system shows remarkable capacity to the monaural separation. The concept of auditory scene analysis (ASA) has been first proposed by Bregman[3], according to him, the human auditory system is similar to that of human visual system. When a visual system views an object, the edges, textures and colors of the objects are evaluated and interpreted as perceptual wholes (e.g., a fan or a table). Similarly, the sound reaching our ears is subjected to an auditory scene analysis. Computational auditory scene analysis (CASA) performs the ASA in machines as same as human beings. In which, the input sound mixture is divided into different time-frequency (T-F) segments and segments are grouped based on cues to extract the target signal. The ideal binary mask (IBM) has been one of the most successful techniques in CASA systems[20,21]. The CASA based speech separation system employing IBM shows large benefits in intelligibility even at low SNR level (-5 dB, -10 dB)[14] and it consistently resolves the ASA constraints in terms of audibility and segregation capacity. 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 process. This introduces an unnatural sound called musical noise.

Some other notable works has been exhausted in the realm of soft mask based source separation. The method proposed in[19] is a two stage frequency-domain procedure for blind separation of convolutive mixed sources. In the first stage, the frequency bin-wise mixtures along the time axis are classified based on the Gaussian mixture model fitting. In the second stage, the permutation ambiguities of classified signals are aligned by clustering the posterior probability sequences which have been calculated in the first stage. Then T-F masking is performed to separate source signals. In another study, a source separation method that uses probabilistic models of sources and expectation-maximization parameter estimation is presented[15]. In which the model-based expectation-maximization source separation and localization (MESSL) is used based on inter-aural phase difference (IPD) and inter-aural level difference (ILD) to cluster the individual spectrogram points. Then the probabilistic mask is created by MESSL to separate the sound sources from the reverberant mixture. In [16] a minimum mean-square error (MMSE) based technique is planned to estimate the ideal multi-threshold mask (IMM) that has been utilized in the realm of monaural speech improvement. It contains 2 stages, specifically training stage and enhancement stage. Within the training stage, a man-made neural network is trained by exploitation the SNR of every T-F unit of training information. Second stage uses the calculated SNR to estimate IMM and to separate the target speech from clamant signal.

This analysis work proposes a method to scale back the impact of musical noise pro-duced by IBM, by planning a soft mask which might be utilized in speech separation applications. Genetic algorithmic rule (GA) is employed during this work to search out the optimum soft mask weights between 0 and 1. The objective measures like S/N improvement and perceptual evaluation of speech quality (PESQ) are used to measure the performance of planned optimum soft mask with the prevailing IBM [22] and IMM[16] based speech separation systems. Rest of the paper is organized in the following manner. Section two provides an outline of computational auditory scene analysis (CASA). Section three presents the proposed optimum soft mask based speech separation system. Section four provides the experimental results of IBM, IMM and also the planned soft mask. Section five describes the conclusion and future work.

2. COMPUTATIONAL AUDITORY SCENE ANALYSIS (CASA)

CASA can be defined as the study of ASA in computational means[20]. The CASA based speech separation system does not require strong assumptions of acoustic properties of interferences. The typical structure of CASA system is shown in Fig. 1.

The input sound mixture having both speech and noise signal has been processed to extract the features. Some system directly performs grouping based on the features but many systems forms an intermediate stage, in which discrete symbols are formed by the significant components in time-frequency. Grouping rules are then used to identify signals of same source. Several CASA systems use an approach of time frequency mask in the intermediate stage to separate the speech and noise signals from the input mixture.

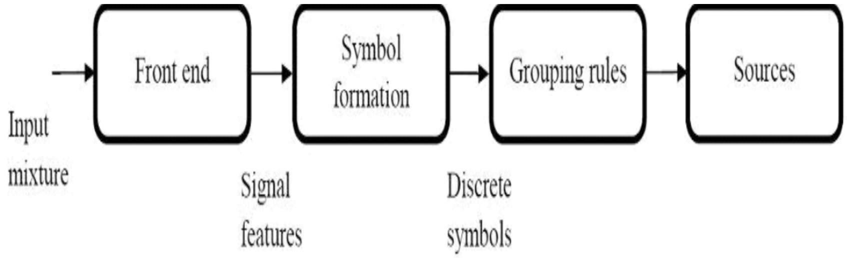

Figure 1: Typical data driven CASA system[4]

2.1. Ideal Binary Mask (IBM)

In time-frequency mask based CASA system, one of the successful approaches is to employ ideal binary mask (IBM)[4]. Employing IBM shows large benefits in intelligibility even at low SNR level (-5 dB, -10 dB) [14]. The value of ideal binary mask is either 1 or 0 and these values are obtained based on the energies of corresponding T-F units of speech and noise signals. The ideal binary mask is given as follows,

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

where $s(t\,,f\,)$ denotes the speech energy and $n(t\,,f\,)$ denotes the noise energy. These energies are calculated in terms of decibels at particular time t and frequency f , and compared against a local criterion (LC). Since it is shown in[1] that the optimum value for LC is 0 dB in terms of SNR, this work adopts 0 dB for LC. As LC is 0 dB, if speech energy is greater than noisy energy the binary value 1 is assigned to the mask, otherwise value 0 is assigned [11]. The generated IBM is applied to the mixture signal to segregate the speech from noise signal. However, a problem with IBM in speech separation applications is employing bi-nary weights (i.e., 0 or 1). This binary weighting may cause some parts of the speech to be discarded during synthesis process [5]. For example, during the unvoiced parts of speech, there is more possibility for noise energy to dominate the speech energy. Due to which speech parts will be discarded and produces an unnatural sound called musical noise.

2.2. Ideal Multi-threshold Mask (IMM)

Masoud et.al [16] states that segregating small amount of noise energy along with the segregated speech does not have destructive role, but it reduces the impact of musical noise. The region between -12 dB to 0 dB is divided into several intervals and weights are employed to segregate the speech. The ideal multi-threshold mask is given by [16],

$$IMM\,(t,\, f) = \begin{cases} 0 & SNR\,(t,\, f) < -12\,dB \\ 0.2 & -12\,dB \le SNR\,(t,\, f) \le -7\,dB \\ 0.4 & -7\,dB \le SNR\,(t,\, f) \le -3\,Db \\ 0.6 & -3\,dB \le SNR\,(t,\, f) \le -1\,dB \\ 0.8 & -1\,dB \le SNR\,(t,\, f) < 0\,dB \\ 1 & SNR\,(t,\, f) \ge 0\,dB \end{cases} \tag{2}$$

where SNR_{TF} denotes the signal-to-noise ratio of a T-F unit. The IMM solves the issue of musical noise to some extent but still there are several problems associated with IMM. First, the weights assigned to the particular regions are not optimum values and has the probability of decreasing the performance. Second, the region between -12 dB to 0 dB is taken into account. Since the amount of noise added depends on the

SNR level [16] it may increase the probability of introducing more noise at the output. Third it requires a training phase, so it takes more time and also in need of more samples to train the system. To reduce the impact of musical noise along with the problems associated with IMM an optimum soft mask based speech separation system is proposed. In which the binary weights (i.e., 1 or 0) are replaced by the variable weights called soft mask in the region between -5 dB to 0 dB. Then genetic algorithm (GA) is used to find the optimum values for the soft mask in the region between -5 dB to 0 dB.

3. PROPOSED OPTIMUM SOFT MASK (OSM)

As mentioned previously, musical noise is the main problem associated with IBM based speech separation. To solve this issue, this work proposes an optimum soft mask based speech separation system, in which the optimum values are obtained using GA. The pro-posed system contains two stages. First stage is the estimation stage, in which the genetic algorithm is used to estimate the mask as shown in Fig. 2. Second stage is the synthesis stage in which the estimated mask is applied to extract the speech from input mixture as shown in Fig. 4.

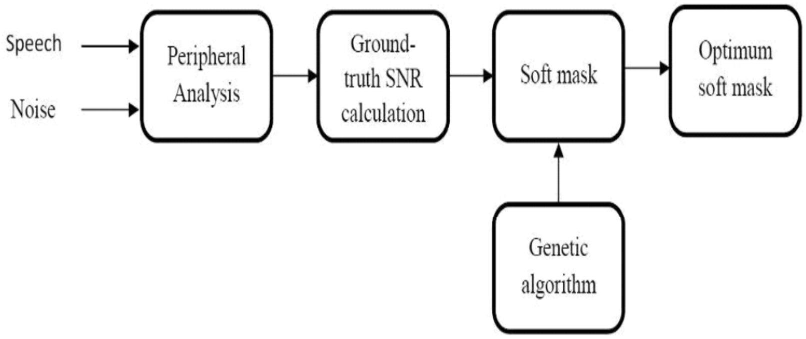

Figure 2: Estimation stage of optimum soft mask

In mask estimation stage, both speech and noise signals are given as an input to the peripheral analysis stage. In which the time-frequency analysis of speech and noise signals are performed by using Gammatone filter bank [6], which are similar to frequency selectivity of human ear. To represent the input signals in two dimensional T-F representations the normal frequency of the signal in Hz is converted into equivalent rectangular band-width (ERB) rate scale which gives an approximation to the bandwidth of filters in human hearing. The upper and lower bound of ERB is calculated by,

$$erb = 21.4\ log_{10}(0.00437\ f + 1) \tag{3}$$

ERB rate scale is divided into several segments within the upper and lower bound, where the number of segments is same as the number of channels. After dividing into segments the ERB rate scale is converted to normal frequency scale by,

$$c\,f = 10^{(erb/21.4-1)}/0.00437 \tag{4}$$

where $c\,f$ denotes the center frequency array indexed by channel. Then the center frequencies are distributed over frequency in proportion to their bandwidths by using,

$$b(c\,f\,) = 1.019 * 24.7 * (4.37 * c\,f\,/1000 + 1) \tag{5}$$

Then impulse response of the filter bank is a product of gamma function and a tone found by using,

$$g_{\,fc}(t\,) = At^{n-1}\,exp[-2\pi\ b(\,f_c)]cos(2\pi\,f_c t\ + \varphi) \tag{6}$$

Here n is the filter order, $c\,f$ is the center frequency in Hz, φ is the phase, A is the loudness-based gain adjustments and $b(c\,f\,)$ determines the bandwidth for a given center frequency. Then the impulse response of each channel is divided into frames, where each frame segment represents the T-F unit. Energy for each

T-F unit in each channel is calculated in dB. Based on the energies, Ground Truth SNR (GTSNR) can be found by the ratio of energy corresponding to T-F units of speech and noise signal,

$$GT\,SNR_{T\,F} = 10\log\frac{\Sigma_n\,(S_{T\,F}(n))^2}{\Sigma_n\,(n_{T\,F}(n))^2} \tag{7}$$

Here s_{TF} and n_{TF} represents the T-F unit energy of speech and noise signals respectively and n is the time index. The Studies made on the application of IBM show that the local SNR value of a particular T-F unit near -5 dB to 0 dB has a greater impact on the speech intelligibility[16]. Hence in this work the interval between -5 dB to 0 dB has been divided into several intervals. The threshold values are assigned for each interval with respect to GTSNR as calculated above. Soft mask can be mathematically described as follows:

$$IMM(t,f) = \begin{cases} 0 & GT\,SNR_{TF} < -5\,dB \\ x1 & -5\,dB \le GT\,SNR_{TF} \le -3dB \\ x2 & -3\,dB \le GT\,SNR_{TF} \le -1dB \\ x3 & -1\,dB \le GT\,SNR_{TF} \le 0\ dB \\ 1 & GT\,SNR_{TF} \ge 0\,dB \end{cases} \tag{8}$$

Where $GT\,SNR_{T\,F}$ denotes the Ground Truth SNR values for particular time-frequency units, and mask (t, f) denotes the weight assigned to the particular T-F unit. The optimum values for variables $x1$, $x2$, and $x3$ are found by using genetic algorithm (GA). GA is an optimization algorithm which will give the best value to substitute in the mask in order to increase the overall performance. The flow diagram of GA is shown in Fig. 3. Initial parameters such as cost function, cost variables, population size M, mutation rate R, selection rate S and number of bits N are defined to progress the objective function. Later, an initial population of random numbers is generated based on the population size M, in which each row of initial population is called chromosomes. A single chromosome C contains,

$$\text{C = Number of variables * N bits} \tag{9}$$

The initial population I_p is given by,

$$I_p = M * C \tag{10}$$

Where I_p contains only binary numbers. Decode chromosomes stage performs the conversion of binary numbers into continuous numbers. The cost of each chromosome is calculated by employing each chromosome values in the mask. Speech is extracted based on the generated mask and SNR improvement is calculated, where SNR improvement denotes the cost value. Chromosomes are ranked from highest to lowest based on the cost value. Chromosomes which yields low cost are discarded based on the selection rate S, which is arbitrary.

$$P_{pre} = S * M \tag{11}$$

Where P_{pre} denotes the highest of the chromosomes preserved for conjugation and thus the lowest M “ P_{pre} chromosomes are discarded to make space for the new set of chromosomes. Pair of chromosomes are chosen from P_{pre} by rank weight methodology to produce a pair of new off-spring. Among that the additive chance of each chromosomes are calculated by victimization,

$$P_n = \frac{P_{nre} - n + 1}{\Sigma_n^P\ pre_1 n} \tag{12}$$

A random selection is generated and thus the chromosome whose chance larger than the random selection is chosen for conjugation. Same weight technique is perennial to choose another chromosome for conjugation.

Conjugation is that the tactic of creating one or further offspring from the parents (selected chromosomes). A crossover point is chosen randomly in between first and last bits of parents' chromosomes. New offspring are created by pairing half the first selected chromosome with the half the second selected chromosome and therefore the different means around. The new offspring are placed into the population. The randomness of the chromosomes is additional increased by the tactic of mutations, among that it alters a definite share of bits among the chromosomes. After mutation methodology, all over again the value associated with chromosomes is calculated. At last, convergence check is made by checking whether or not an acceptable resolution is reached or the amount of iteration exceeds the outlined value. If every condition is not glad, it will repeat the tactic as shown in Fig. 3. If anyone condition is glad, chromosomes at high of the population results the optimum values for for $x1$, $x2$ and $x3$. Supported these values the optimum soft mask is calculated as in equation (8).

The input speech mixture is decomposed into T-F units by using a bank of gammatone filters in peripheral analysis stage. Then the estimated optimum soft mask is employed in the synthesis stage to suppress the noise and enhance the speech signal. Weintraub [23] describes the steps in synthesis stage to separate the speech from the input mixture as follows, first step is to time reverse the response of the filter and passing

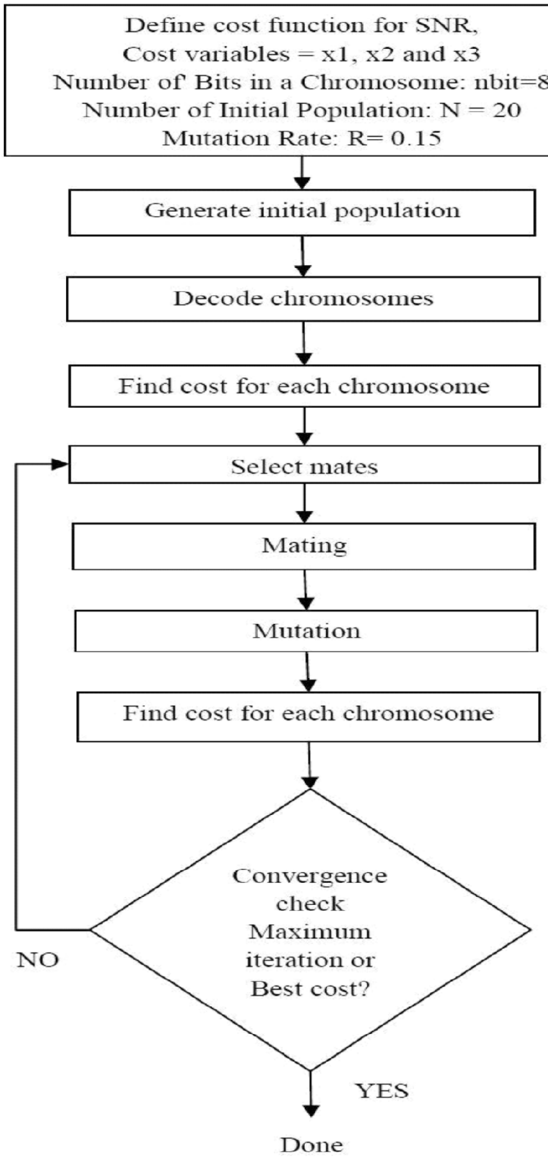
Define cost function for SNR,
Cost variables = x1, x2 and x3
Number of Bits in a Chromosome: nbit=8
Number of Initial Population: N = 20
Mutation Rate: R= 0.15
Generate initial population
Decode chromosomes
Find cost for each chromosome
Select mates
Mating
Mutation
Find cost for each chromosome
Convergence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

Figure 3: Flowchart of a binary GA optimization of variables in soft mask

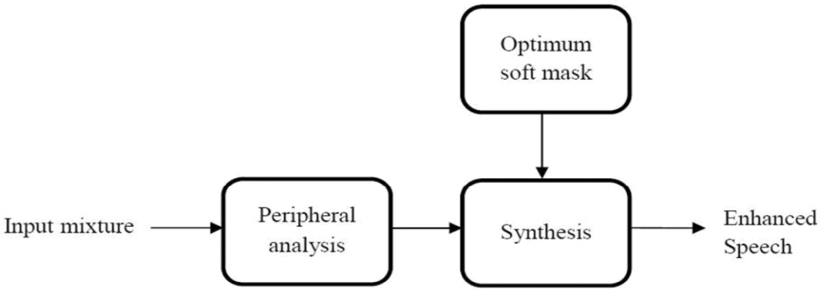

Figure 4: Synthesis stage of the optimum soft mask

it back to the filter. Again the response is time reversed to remove the phase shifts across channels in the filter bank output. Next step is to divide the phase corrected output from each channels into frames by windowing with a raised cosine. Frame size should be same as used in decomposing the input mixture into T-F units. Then the constructed optimum soft mask is applied to T-F units of the input mixture. Energies in each T-F units of the input mixture are weighted based on the corresponding mask values. After weighting the T-F units, the weighted responses across all frequency channels are summed to reconstruct the speech from the input speech mixture. The quality of the segregated speech signal is assessed by SNR improvement and PESQ measures.

4. ANALYSIS AND EXPERIMENTAL RESULTS

The performance and capability of projected optimum soft mask (OSM) within the mono speech separation is evaluated by mistreatment objective measures like SNR improvement and PESQ measurement. The SNR improvement is calculated by,

$$SNR = 10\log_{10}\left(\frac{\Sigma_n S_{Allonemask}(n)^2}{\Sigma_n (S_{Allonemask}(n) - S_{out}(n))^2}\right) \tag{13}$$

Where $S_{Allonemask}$ (n) denotes the speech obtained from synthesis method by keeping all the mask value as one and S_{out} (n) denotes the enhanced speech obtained by applying the optimum soft mask in the synthesis method Perceptual evaluation of speech quality (PESQ) is the ITU-T P.862 recommendation [9,10] used to evaluate quality of speech signal. The scores range from 4.5 to -0.5, where 4.5 represents highest quality and -0.5 represents the low quality of speech. To conduct the experiment 25 clean speech samples are selected randomly from the IEEE speech corpus [18] and two noises such as, babble noise and factory noise are taken from the noisex-92 database [17]. The clean speech samples are shown in Table 1.

The performance of optimum soft mask (OSM) is compared against the ideal binary mask (IBM) and ideal multi-threshold mask (IMM) based speech separation system. The process is as follows, first, the clamant speech mixture is created by manually mixing the clean speech and noise signals at totally different SNRs (-5 dB to +5 dB). Then IBM, IMM and the planned OSM are applied to the clamant mixture. The weighted responses are finally processed by synthesis module to yield the enhanced speech. Table 2, 3 and 4 shows the the typical SNR improvement and PESQ scores obtained by processing the mixture signals with IBM, IMM and OSM, at totally different input SNRs, for the 25 clean speech samples with babble and factory noise respectively. As seen from the table, the typical SNR improvement of planned OSM is slightly greater than IBM and considerably greater than IMM. The PESQ score of the planned OSM is greater than IBM but however smaller than IMM. The IMM based speech separation gives higher PESQ scores but low SNR improvement as compared to IBM. However the IBM provides higher SNR improvement and low PESQ scores. The planned OSM provides sensible SNR improvement than IBM and IMM and produce con-siderable improvement in PESQ score.

Table 1
IEEE Speech corpus - clean speech signal

Sl. No.	*Speech Ref. No* [18]	*Clean speech signal*
1.	S-01-06	The juice of lemons makes fine punch
2.	S-01-10	A large size in stockings is hard to sell
3.	S-02-06	A pot of tea helps to pass the evening
4.	S-03-09	It snowed, rained, and hailed the same morning
5.	S-04-02	Take the winding path to reach the lake
6.	S-04-08	The young girl gave no clear response
7.	S-04-09	The meal was cooked before the bell rang
8.	S-05-01	A king ruled the state in the early days
9.	S-05-09	The friendly gang left the drug store
10.	S-06-03	Adding fast leads to wrong sums
11.	S-07-03	He ran half way to the hardware store
12.	S-09-03	There are more than two factors here
13.	S-10-06	Bail the boat, to stop it from sinking
14.	S-18-04	The sky that morning was clear and bright blue
15.	S-18-06	Sunday is the best part of the week
16.	S-21-07	After the dance, they went straight home
17.	S-24-10	Madam, this is the best brand of corn
18.	S-25-08	To make pure ice, you freeze water
19.	S-31-02	The plant grew large and green in the window
20.	S-38-07	Go now and come here later
21.	S-38-10	That move means the game is over
22.	S-43-02	Draw the chart with heavy black lines
23.	S-47-01	The music played on while they talked
24.	S-47-09	Birth and death mark the limits of life
25.	S-48-07	We don’t get much money but we have fun

Table 2
Performance results of Ideal Binary Mask (IBM) based speech separation

Input SNR (dB)		*-5*	*-2.5*	*0*	*2.5*	*5*
Babble Noise	SNR Improvement	5.96556	7.25673	8.66472	10.13486	11.66749
	PESQ Score	1.65498	1.89659	2.12066	2.339624	2.565556
Factory Noise	SNR improvement	5.45233	6.97722	8.69293	10.41973	12.24714
	PESQ Score	1.92777	2.13836	2.33584	2.53116	2.727864

Table 3
Performance results of Ideal Multi-threshold Mask (IMM) based speech separation

Input SNR (dB)		*-5*	*-2.5*	*0*	*2.5*	*5*
Babble Noise	SNR Improvement	5.953652	7.194888	8.521268	9.927964	11.44784
	PESQ Score	2.032544	2.227384	2.421092	2.611448	2.808444
Factory Noise	SNR improvement	5.239368	6.794612	8.467148	10.21994	12.08419
	PESQ Score	2.261324	2.42176	2.59776	2.776636	2.957288

Table 4
Performance results of Optimum Soft Mask (OSM) based speech separation

Input SNR (dB)		*-5*	*-2.5*	*0*	*2.5*	*5*
Babble Noise	SNR Improvement	6.033648	7.312076	8.692428	10.14156	11.66068
	PESQ Score	1.81782	2.026532	2.242892	2.449748	2.654328
Factory Noise	SNR improvement	5.4751	6.987576	8.681976	10.406996	12.24458
	PESQ Score	2.062332	2.258864	2.44238	2.623508	2.826252

Alternatively, ideal binary mask (IBM) and therefore the planned optimum soft mask (OSM) are compared supported cochleagram of the signals. Fig. 5 shows the cochleagram of the clean speech signal "The sky that morning was clear and bright blue" taken from the IEEE database[18] and also the mixture signal in which the clean speech signal is mixed with babble noise taken from Noisex-92 database [17] at SNR= -5 db. The mixture signal is then masked with the perfect binary mask and optimum soft mask severally, and it may be seen that gap due to loss of some speech components within the IBM are ûlled significantly with OSM and improves the standard.

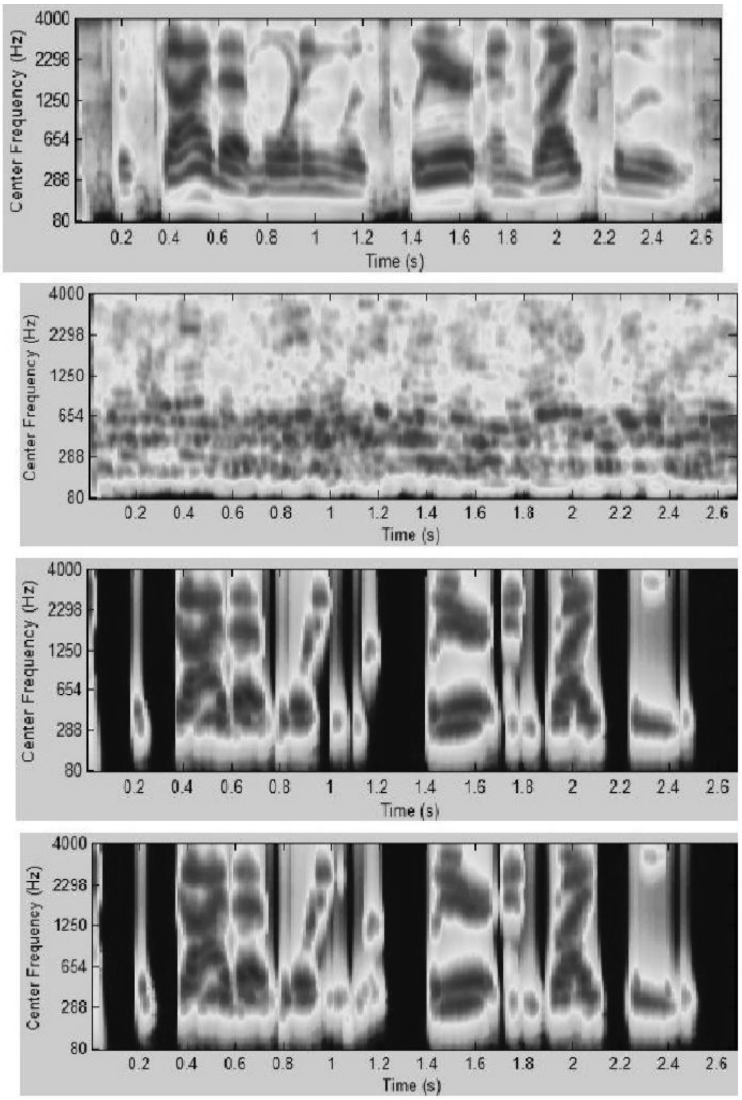

Figure 5: Comparison of performance of ideal binary mask (IBM) and optimum soft mask (OSM). (a) Represents the cochleagram of the clean speech signal. (b) Represents the cochleagram of noise signal. (c) Shows the cochlea-gram of the enhanced speech signal using IBM. (d) Shows the cochleagram of enhanced speech signal using OSM

5. CONCLUSION AND FUTURE WORK

The ideal binary mask has been one among the foremost victorious techniques in CASA algorithm. However it introduces musical noise since some components or regions of the improved speech is discarded throughout synthesis. particularly the unvoiced components of the speech are lost if the noise has bigger energy and ends up in reduction of speech quality and comprehensibility. In this research work, associate optimum soft mask is planned to scale back the musical noise in speech separation systems. The experimental results are shown in Table 2-4 and ascertained that, the quality of the speech made by optimum soft mask is healthier than the normal IBM and IMM. However, IBM, IMM and OSM in its current type need the previous information of the clean speech and noise signals. This is often one in all the restrictions of this system. Further investigation of estimating the optimum soft mask without the previous information of speech and noise signal for mono speech separation is ongoing. Additionally the implementation of proposed optimum soft mask in digital signal processor for real speech separation is in progress.

REFERENCES

[1] (M.C. Anzalone, L. Calandruccio, K. A. Doherty and L. H Carney.,2006) Determination of the potential benefit of time-frequency gain manipulation, Ear Hear- PubMed article, pp. 480-492.

[2] (A. K. Barros, T. Rutkowski, F. Itakura and N. Ohnishi.,2002) Estimation of speech embedded in a re-verberant and noisy environment by independent component analysis and wavelets, *IEEE Trans-action on Neural Network* **13(4)** pp. 888-893.

[3] (G. J. Brown and D. L. wang.,2005) Speech Enhancement, *Separation of Speech by Computational Auditory Scene Analysis*, eds. J. Benesty, S. Makino and J. Chen, (Springer, New York), pp. 371-402.

[4] (S. Cao, L. Li and X. Wu.,2011) Improvement of intelligibility of ideal binary-masked noisy speech by adding background noise, *Journal of Acoustic Society of America*.

[5] (V. Hohmann.,2002) Frequency analysis and synthesis using a Gammatone filter bank, *Acta Acoustica united with Acustica, pp*. 433-442.

[6] (G. Hu, D. L.,2001) Wang, Speech segregation based on pitch tracking and amplitude modulation, *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, 21-24., pp. 79-82.

[7] (G. Hu and D. L. Wang., 2004) Monaural speech segregation based on pitch tracking and amplitude modulation, *IEEE Transaction on Neural Networks* p.p. 1135-1150.

[8] (Y. Hu, P. C. Loizou., 2008)Evaluation of objective quality measures for speech enhancement, *IEEE Transaction on Audio Speech Language Processing pp*. 229-238.

[9] (ITU-T.,2001) Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codec Series P: Telephone Transmission Quality Recommendation Pp.862.

[10] (Yi Jiang, Hong Zhou and Zhenming Feng, 2011) Performance analysis of ideal binary masks in speech enhancement, *Image and Signal Processing (CISP)*.

[11] (H. Krim and M. Viberg., 1996) Two decades of array signal processing research: The parametric ap-proach, *IEEE Signal Processing Magazine*.

[12] (P. Li, Y. Guan, B. Xu and W. Liu., 2006) Monaural Speech Separation Based on Computational Au-ditory Scene Analysis and Objective Quality Assessment of Speech, *IEEE Transaction on Audio Speech Language Processing*.

[13] (N. Li and C. Loizou., 2008) Factors influencing intelligibility of ideal binary-masked speech: implica-tions for noise reduction, *Journal of Acoustic Society of America*, pp. 1673-1682.

[14] (M. I. Mandel, R. J. Weiss and D. P. W. Ellis,2010) Model-based expectation-maximization source sep-aration and localization, *IEEE Transaction on Audio Speech Language Processing* pp. 382-394.

[15] (Masoud Geravanchizadeh and Reza Ahmadnia,2014) Blind Source Separation, *Monaural Speech En-hancement Based On Multi-threshold Masking*, eds. G.R. Naik, W. Wang (Springer-Verlag Berlin Heidelberg, pp. 369-393.

[16] (E. H. Rothauser, W. D. Chapman, N. Guttman, M. H. L Hecker, K. S. Nordby, H. R. Silbiger, G. E. Urbanek and M. Weinstock,1969) IEEE recommended practice for speech quality measurements, *IEEE Transaction on Audio Electro acoustics*, **17** pp. 225-246.

[17] (H. Sawada, S. Araki and S. Makino, 2007) A two stage frequency-domain blind source separation method for under determined convolutive mixtures, *IEEE workshop on applicants of signal pro-cessing to audio and acoustics*, New Yark, USA, 21-24, pp. 139-142.

[18] (D. L. Wang and G. J. Brown,2006) *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Hoboken: Wiley-IEEE Press).

[19] (D. L. Wang,2005) Speech Separation by Humans and Machines, *On ideal binary mask as the compu-tational goal of auditory scene analysis* eds. Divenyi and Pierre (Springer US), 20, pp.181-197.

[20] (D. L. Wang,2008) Time-Frequency Masking for Speech Separation and Its Potential for Hearing Aid Design, *Trends in Amplification*, **12** 332-353.

[21] (M. Weintraub,1985) A Theory and Computational Model of Auditory Monaural Sound Separation, Ph.D. Thesis, Stanford University.