



A new Genetic Algorithm based fusion scheme in monaural CASA system to improve the performance of the speech

S. Shoba¹ · R. Rajavel¹

Received: 30 May 2018 / Accepted: 28 April 2019 / Published online: 6 May 2019
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract

This research work proposes a new Genetic Algorithm (GA) based fusion scheme to effectively fuse the Time–Frequency (T–F) binary mask of voiced and unvoiced speech. The perceptual cues such as correlogram, cross-correlogram and pitch are commonly used to obtain the T–F binary mask of voiced speech. Recently, researchers use speech onset and offset to segment the unvoiced speech from the noisy speech mixture. Most of the research work which uses speech onset and offset to represent the unvoiced speech, combine the segments of unvoiced speech with the segments of voiced speech to obtain the T–F binary mask. This research work effectively fuses the T–F binary mask of voiced and unvoiced speech, instead of combining the segments of voiced and unvoiced speech using a Genetic Algorithm (GA). Moreover, a new method is proposed in this research work to obtain a T–F binary mask from the segments of unvoiced speech. The performance of the proposed GA based fusion scheme is evaluated using measures such as quality and intelligibility. The experimental results show that the proposed system enhances the speech quality by increasing the SNR with an average value of 10.74 dB and decreases the noise residue with an average value of 26.15% when compared with noisy speech mixture and enhances the speech intelligibility by increasing the CSII, NCM and STOI with an average value of 0.22, 0.20 and 0.17 as compared with the conventional speech segregation systems.

Keywords Speech segregation · GA fusion scheme · T–F binary mask · Voiced Mask · Unvoiced Mask

1 Introduction

In everyday life, target speech reaching the human auditory system is always accompanied by many other sound sources such as musical sounds, voices of other people speaking and shouting, environmental sounds and man-made sounds. In the last few decades, many researchers developed speech segregation system to segregate the target speech effectively from the other sound sources, unfortunately, there is no such system exist till now. The speech segregation systems find a lot of potential applications including noise robust automatic speech and speaker recognition, music transcription, digital content management, and mobile communication. There are applications which use only one microphone to record

the target speech, for example, voice communication, audio retrieval and hearing aids. Segregating the target speech from one microphone recording of the speech is named as monaural speech segregation and it is one of the most challenging problems in speech signal processing. Although monaural speech segregation is a challenging problem, humans have the inherent capability to focus and understand the target speech recorded using a single microphone in spite of disturbances caused by other sound sources. In the last few decades, research work has been carried on more than hundreds of articles on monaural approaches, and some of the techniques in monaural speech segregation are spectral subtraction (Boll 1979; Hu and Wang 2011; Therese and Lingam 2017), subspace analysis (Ephraim and Trees 1995), model-based approaches (Ellis and Weiss 2006), and feature-based approach (Hu and Wang 2006; Phapatanaburi et al. 2017). Spectral subtraction is the conventional speech enhancement method for reducing the additive noise. This method reduces the noise by subtracting the estimated target speech spectrum from the noisy speech spectrum in the frequency domain (Boll 1979). The main drawback of this

✉ S. Shoba
shobansb@gmail.com

R. Rajavel
rajavelr@ssn.edu.in

¹ SSN College of Engineering, Old Mahabalipuram Road, Chennai 603 110, India

method is that it contains some spectral artifacts called remnant noise which irritates the human ear (Boll 1979; Hu and Wang 2011). In subspace analysis, the vector space of noisy speech will be decomposed into speech subspace and noise subspace using orthogonal matrix vectorization techniques and later the noise subspace will be suppressed (Hyvarinen et al. 2001; Jensen and Hansen 2001). In model-based methods, trained models by GMM/HMM/CNN/DNN of clean speech and interference are used to segregate the speech. Model-based methods show satisfactory performance only when predefined models are available and the model of the speech is matched exactly with the predefined models (Dharmalingam et al. 2017; Gibak and Loizou 2010; Yi-nan et al. 2014; Naik et al. 2012; Naik and Ganesh 2012; Nilesh et al. 2013; Qazi et al. 2018; Sameti et al. 1998; Singhal et al. 2018; Trowitzsch 2017; Wang and Kun 2013; Yu et al. 2017; Zhang and Wang 2017). The remarkable single channel model-based system proposed by Nilesh et al. (2013) uses IBM (Ideal Binary Mask) and IWF (Ideal Wiener Filter) to segregate the target speech. The performance of the single channel model-based system is improved even at low SNR conditions. Another model-based system proposed by Gibak and Loizou (2010) uses Bayesian classifier to classify the noisy speech into a target and masker dominant T–F units. Dharmalingam et al. (2017) proposed a successful model-based system which uses SVM (support vector machine) model to segregate the target speech from the noisy mixture. All these approaches need some form of clean speech and noise but this may not be possible in real scenarios. Alternatively, a feature-based approach has been used for speech segregation since it does not need a prior information of the speech or noise. After Auditory Scene Analysis (ASA) has been proposed and, many researchers started to use ASA concepts to build feature-based speech segregation system are known as Computational Auditory Scene Analysis (CASA) systems (Brown and Cooke 1994; Brown and Wang 2005; Donald and Wang 2017; Harish and Rajavel 2014; Hu and Wang 2006).

A typical CASA system consists of auditory filtering, feature extraction, segmentation, grouping and re-synthesis. In auditory filtering, the acoustic mixture is decomposed into many T–F units using bandpass filtering and windowing technique. In feature extraction, auditory cues such as correlogram, cross-correlogram, pitch and onset/offset are extracted as features. In segmentation, each T–F units are segmented into speech T–F units and noise T–F units based on the extracted features. In grouping, the speech segments are grouped together and noise segments are grouped together to form speech group and noise group respectively. Each T–F unit in the speech group is denoted by ‘1’ and similarly each T–F unit in noise group is denoted by ‘0’ this constitute a binary mask. Finally, the obtained binary mask is used to synthesize the noisy speech into speech and

noise in the re-synthesis stage. Various CASA systems have been proposed in the literature over the past three decades for monaural speech segregation (Brown and Cooke 1994; Brown and Wang 2005; Donald and Wang 2017; Harish and Rajavel 2014; Hu and Wang 2006). A CASA system proposed by Hu and Wang (2004) uses auto-correlation, cross-correlation, pitch tracking and amplitude modulation to separate the monaural voiced speech from the noisy speech signal. Wang (2012) proposed another CASA system namely tandem algorithm to separate the monaural voiced speech from the noisy speech mixture and showed significant improvement in the speech quality. The above two CASA systems concentrate mainly on separating the voiced speech from the monaural noisy speech mixture and do not deal with unvoiced speech. A CASA system proposed Hu and Wang (2007) segment the speech into voiced and unvoiced speech segments using onset and offset analysis. Later, Hu and Wang (2013) proposed an unsupervised model-based CASA system for co-channel speech segregation in which tandem algorithm and clustering technique are used to produce voiced binary mask. Onset and offset based segmentation scheme is used to generate the segments for unvoiced speech. Later, the voiced T–F binary mask and unvoiced segments are grouped together and an overall T–F binary mask has been generated for co-channel speech segregation. Wang et al. (2013) proposed a CASA based monaural speech segregation system to segregate both voiced and unvoiced speech from the noisy speech mixture. This system uses auditory cues such as correlogram, cross-correlogram, response energy for voiced speech and onset/offset for unvoiced speech to construct a binary mask. The author proposed a morphological image processing technique to smooth and retrieve the binary mask. The obtained binary mask contains some broken auditory fragments which will degrade the quality of the resynthesized speech.

This research work proposes a new method to compute the T–F binary mask of unvoiced speech and voiced speech and then fuse them effectively using Genetic Algorithm (GA). The systems proposed by Hu and Wang (2013) and Wang et al. (2013) use onset and offset analysis to obtain the unvoiced segments and merged with the voiced T–F binary mask. The proposed system uses the same onset and offset analysis to extract the voiced and unvoiced segments and then mapped into its equivalent T–F binary mask. Another T–F binary mask is obtained from the voiced speech binary mask computation via feature extraction and initial segmentation. The two T–F binary masks representing voiced and unvoiced speech are fused using a Genetic Algorithm (GA). In the fusion process, a Genetic Algorithm is used to appropriately weight the voiced and unvoiced T–F binary mask such that the quality and intelligibility of the speech is increased. The proposed system is evaluated using Cooke speech database (Cooke 1993) which consist of 10 speech

utterances and 17 noises (Wang et al. 2013). The proposed system performance is measured in terms of Signal to Noise Ratio (SNR), Perceptual Evaluation of Speech Quality (PESQ) as a quality measure, Percentage of noise residue (P_{nr}) and Percentage of energy loss (P_{el}) as complementary error measure and coherence speech intelligibility index (CSII), normalized covariance metric (NCM) and short-time objective intelligibility (STOI) as the intelligibility measure (Hu and Loizou 2007; Ma et al. 2009; Taal et al. 2011). The outcome of this research work enhances the speech quality by improving the SNR and PESQ and improves the speech intelligibility in terms of CSII, NCM and STOI with respect to other existing systems.

The rest of the research article is prepared as follows. Section 2 describes the computation of voiced T–F binary mask, a new scheme of computing T–F binary mask for unvoiced speech, and finally the proposed GA based T–F binary mask fusion. In Sect. 3, the quality, intelligibility and complementary error measures are evaluated and their results are reported. Finally, Sect. 4 concludes the paper with the possible future extension of this research work.

2 Speech segregation system

The main contribution of this research work is to effectively fuse the T–F binary mask of voiced and unvoiced speech using the Genetic Algorithm such that the quality and intelligibility of the speech is improved. The proposed T–F

binary mask fusion using a Genetic Algorithm is illustrated in Fig. 1 contains auditory filter bank, voiced binary mask computation, unvoiced binary mask computation; Genetic Algorithm based T–F binary mask fusion and speech reconstruction of the signal.

2.1 Auditory filter bank

The proposed speech segregation system illustrated in Fig. 1 contains an auditory filter bank as the first element where the noisy speech mixture is processed through 128 channel gammatone filterbanks centered from 50 Hz to 8 KHz and its response is given in (Meddis 1988; Patterson et al. 1988) as

$$g_c(f_c, t) = \begin{cases} b_w^N t^{N-1} e^{2\pi b_w t} \cos(2\pi f_c t), & \text{if } t \geq 0, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

where b_w determines the bandwidth, f_c is the center frequency and N is the order of the filter. The response from the gammatone filter produces a two-dimensional Time–Frequency (T–F) mask called cochleagram.

2.2 Voiced binary mask computation

The main idea of this research work is to employ a feature extraction process to obtain the T–F binary mask for voiced and unvoiced speech. The schematic block diagram to obtain the voiced T–F binary mask is shown in Fig. 2. It consists of feature extraction, initial segmentation, pitch estimation, segments grouping & unit labeling

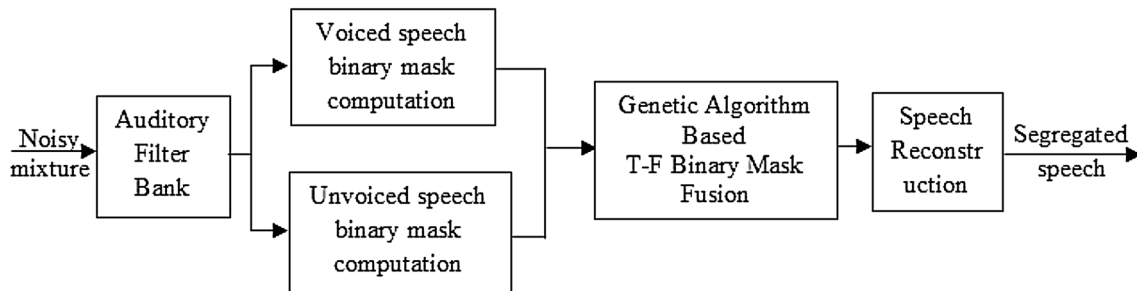


Fig. 1 Proposed GA speech segregation system

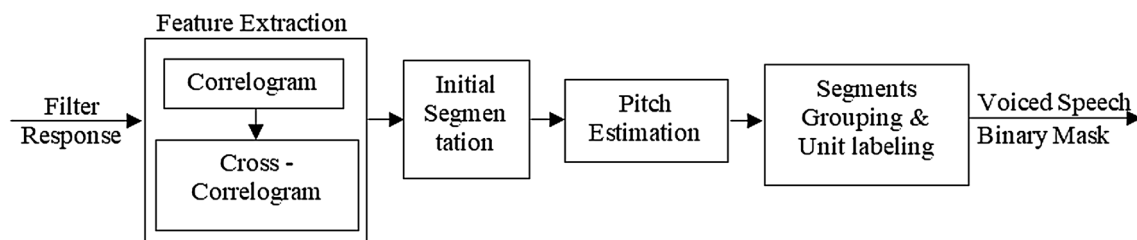


Fig. 2 Voiced binary mask computation

segments grouping and unit labeling. The following subsections briefly explain in detail of the above-mentioned components of voiced binary mask computation.

2.2.1 Feature extraction

Feature extraction is one of the important stages in the voiced binary mask computation where the response of the filter is given as an input to extract speech features such as energy, correlogram and cross-correlogram (Hu and Wang 2004). The reliable extracted speech features are used in the initial segmentation stage to segment each T–F unit into voiced dominant or noise dominant T–F unit.

2.2.2 Initial segmentation

The initial segmentation is the intermediate stage in voiced binary mask computation in which every individual T–F unit is segmented into voiced dominant or noise dominant T–F unit. According to Hu–Wang (2013), if the energy $E(F_c, T_m)$ of a T–F unit is greater than a constant threshold θ_H^2 (where $\theta_H^2 = 50$) the corresponding T–F unit is represented as voiced dominant and denoted as ‘1’ and the remaining T–F unit is represented as noise dominant and denoted as ‘0’. Likewise, the cross-correlogram of a T–F unit is greater than threshold θ_c [0.985 as in Wang and Kun (2013)], the corresponding T–F unit is denoted as ‘1’ otherwise denoted as ‘0’. This approach reduces the quality of speech by missing some T–F units where voiced speech is less dominant than noise. It is known, the energy and cross-correlogram are the reliable features to segregate the voiced speech from the noisy speech mixture. By combining the energy and the cross-correlogram, a T–F binary mask $M(F_c, T_m)$ for voiced speech is created as in Shoba and Rajavel (2017a) as

$$M(F_c, T_m) = \begin{cases} 1 & \text{if } E(F_c, T_m) \geq \theta_{AT}(F_c) \& C(F_c, T_m) \geq \theta_c, \\ 0 & \text{otherwise,} \end{cases} \quad (2)$$

The binary mask $M(C_H, f_R)$ will represent the voiced segment of speech and failed to represent the unvoiced segment of speech. The above T–F binary mask $M(F_c, T_m)$ will represent the voiced speech and denoted as ‘1’ and remaining T–F units are denoted as ‘0’.

2.2.3 Segments grouping and Unit Labeling

Segments grouping and unit labeling is the last stage in the voiced binary mask computation. The above stage captures the majority of the voiced speech T–F units and misses few of the T–F units where noise is more dominant. The missed T–F units are recovered by the process called as pitch estimation and are done by the summary of the correlogram of T–F units over all channels in Hu and Wang (2004) as

$$\xi_F(T_m, \tau) = \sum_c A_H(F_c, T_m, \tau) \quad (3)$$

The pitch estimation proposed in Hu and Wang (2004) estimates the pitch period at frame T_m in plausible pitch range of 2 to 12.5ms. By this estimation, most of the voiced speech T–F units are captured accurately by eliminating the noise components of the speech. The voiced speech T–F units captured from the pitch estimation is grouped and labeled as in Hu and Wang (2004) as

$$M_v(F_c, T_m) = \begin{cases} 1 & \text{if } \frac{A_H(F_c, T_m, \tau_s(T_m))}{A_H(F_c, T_m, \tau_p(F_c, T_m))} > \theta_T \\ 0 & \text{otherwise,} \end{cases} \quad (4)$$

where θ_T is the threshold equal to 0.85 which is the most appropriate value taken from Wang and Brown (1999) for labeling the particular T–F unit that determines the target speech as voiced speech denoted as ‘1’ and remaining background T–F units as noise denoted as ‘0’. The voiced binary mask computation stages finally produce a voiced T–F binary mask $M_v(F_c, T_m)$ that captures almost all the voiced speech T–F units and also eliminates the noise present in the noisy speech mixture.

2.3 Unvoiced binary mask computation

The previous section explains various steps involved in the computation of the T–F binary mask for voiced speech. This section explains how a T–F binary mask is acquired in this research work for unvoiced speech. The schematic block diagram to obtain the T–F binary mask for unvoiced speech is shown in Fig. 3. It consists of a smoothing process, onset/offset detection and matching and multiscale integration and mapping of segments into the T–F binary mask. The

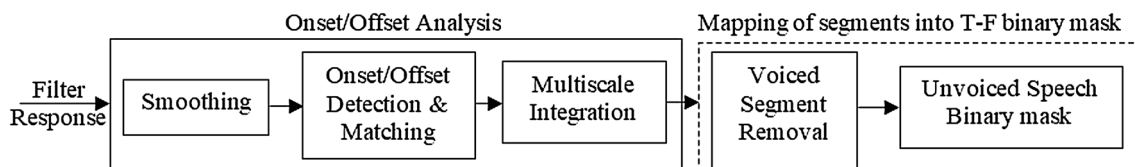


Fig. 3 Unvoiced binary mask computation

following subsections explain the computation of the T–F binary mask for unvoiced speech.

2.3.1 Onset/offset analysis

The first step in the unvoiced binary mask computation is the onset/offset analysis. The sudden increase in intensity is called onset and sudden decrease in intensity is called offset. This can be obtained by taking the first derivative of the intensity with respect to time and then finds the peaks and valleys of the derivative. Actual onset/offset cannot be determined due to intensity fluctuations produced by the peaks and valleys of the derivate. To reduce the intensity fluctuations, smoothing is done which smooth the intensity over time and frequency. After smoothing, onset/offset detection and matching detects the onset and offset in every filter channel and merge them into onset and offset fronts. Each onset /offset fronts are paired to form segments. Finally the multiscale integration form segments at different scales (Hu and Wang 2007). The following subsections explain in detail about the onset/offset analysis.

Smoothing The first stage in the onset and offset analysis is smoothing. Smoothing is a process where the fluctuations will be reduced to produce a better solution. The initial intensity $\vartheta(c, t, 0, 0)$ is smoothed with a low pass filter $h(s_t)$ over time and then $\vartheta(c, t, 0, s_t)$ is smoothed with a Guassian kernel $g(0, s_c)$ over frequency. The response of the smoothing is given as in Hu and Wang (2007) as

$$\vartheta(c, t, 0, s_t) = \vartheta(c, t, 0, 0) * h(s_t), \tag{5}$$

$$\vartheta(c, t, s_c, s_t) = \vartheta(c, t, 0, s_t) * g(0, s_c), \tag{6}$$

where * represents convolution, the smoothing degree is denoted by (s_c, s_t) . At the end of the smoothing process, the intensity fluctuations are reduced considerably but at the same time, major intensity changes of onset and offset are preserved.

Onset and offset detection and matching The next step after smoothing is the onset and offset detection and matching, which detects the onset and offset by finding the peaks and valleys of the derivative of the intensity as given in Hu and Wang (2007) as

$$\frac{d}{dt}\vartheta(c, t, s_c, s_t) = \vartheta(c, t, 0, 0) * \left[\frac{d}{dt}h(s_t) \right] * g(0, s_c), \tag{7}$$

The peak and valleys are determined by setting a threshold value $\theta_T = \mu + \sigma$ where μ denotes mean and σ denotes standard deviation. A true onset is found by its corresponding peak higher than threshold θ_T and offset is found between two onset peaks. In this way, onset and offset are determined for each filter channels. The common onset and offset found in adjacent channels are merged to form onset and

offset fronts. Then the individual onset and offset fronts are matched to form segments (Hu and Wang 2007).

Multiscale integration The final step in the onset and offset analysis is the multiscale integration. The segments generated in the above step captures the voiced and unvoiced speech segments roughly but may miss some small voiced and unvoiced speech segments. To generate the segments accurately different scales are integrated from coarser to finer scale (Hu and Wang 2007). The coarser scale captures a majority of the segments containing speech events but misses some of the small segments. Segments locate more onset and offset positions at finer scale but do not form any new speech T–F segments. The finer scale captures more T–F speech segments and also more segments from the interference. Thus this process identifies voiced, unvoiced and some noisy speech segments as denoted as $M_{v/uv}(F_c, T_m)$ which will be an important cue for speech segregation.

2.3.2 Mapping of segments into an unvoiced T–F binary mask

One of the proposals of this research work is to obtain the T–F binary mask for the unvoiced speech from segments $M_{v/uv}(F_c, T_m)$ generated through the onset and offset analysis. Each segment of onset and offset analysis contains a unique label. The steps involved in the computation of the T–F binary mask for an unvoiced speech from these segments are as follows

Proposed algorithm

- Step 1:** Replace all the non-zero label value in the segments produced by the onset and offset analysis by binary 1.
- Step 2:** Next, compare the segment map of voiced speech $M_v(F_c, T_m)$ obtained in Sect. 2.2 with the segment map of unvoiced and voiced speech obtained using onset and offset analysis $M_{v/uv}(F_c, T_m)$.
- Step 3:** Then, identify the overlapping region of $M_v(F_c, T_m)$ and $M_{v/uv}(F_c, T_m)$ and complement all the T–F units of $M_{v/uv}(F_c, T_m)$ within that overlap region. This process will remove all the voiced segments of $M_{v/uv}(F_c, T_m)$ as follows

$$\begin{aligned} M_{vr/l}(F_c, T_m) &= M_{v/uv}(F_c, T_m) - M_v(F_c, T_m) \quad \text{and} \\ M_{uv/l}(F_c, T_m) &= 0 \quad \text{if } M_{vr/l}(F_c, T_m) < 1, \end{aligned} \tag{8}$$

- Step 4:** The resultant binary matrix $M_{uv/l}(F_c, T_m)$ now contains only the unvoiced speech segments with a little bit of interference. To further remove the interferences, the energy of the T–F frame corresponding to the binary 1 in $M_{uv/l}(F_c, T_m)$ is

computed via correlogram similar in (Hu and Wang 2004) as follows

$$A_{uv}(F_c, T_m, \tau) = \frac{1}{N_c} \sum_{n=0}^{N_c-1} h_{uv}(F_c, T_m T - n) * h_{uv}(F_c, T_m T - n - \tau), \quad (9)$$

where h_{uv} is the hair cell output corresponding to binary 1 in $M_{uv/I}(F_c, T_m)$ and * denotes convolution.

Step 5: If the energy of the T–F units corresponding to the binary 1 in $M_{uv/I}(F_c, T_m)$ is greater than a threshold θ_A those T–F units are marked as 1 to represent it as unvoiced speech segments otherwise it is zero as follows

$$M_{uv}(F_c, T_m) = \begin{cases} 1 & \text{if } A_{uv}(F_c, T_m, \tau) > \theta_A \\ 0 & \text{otherwise,} \end{cases} \quad (10)$$

The value for θ_A is determined via experimentation and the resultant matrix $M_{uv}(F_c, T_m)$ represent the T–F binary mask for unvoiced speech.

2.4 Proposed GA based T–F binary mask fusion

Another proposal of this research work is to effectively fuse the T–F binary mask representing the voiced and unvoiced speech such that the speech is enhanced using a Genetic Algorithm. The Genetic Algorithm (GA) is an optimization and search technique based on the principles of genetics and random selection. An initial population of individuals (weights for fusion) is made and by iterative use of the genetic operators (selection, crossover, mutation) an optimal weight is reached according to the defined fitness function

(Rajavel and Sathidevi 2012, 2011). The GA used in this work to obtain the correct weight factor α to fuse the T–F binary mask of voiced and unvoiced speech and in turn improves the speech quality and intelligibility according to the following equation as

$$Opt_{BM_{T-F}}(F_c, T_m) = \alpha * M_v(F_c, T_m) + (1 - \alpha) * M_{uv}(F_c, T_m) \quad \text{for } 0 \leq \alpha \leq 1, \quad (11)$$

Initially, α has chosen as 0.5 to give equal weights to voiced and unvoiced T–F binary mask and conducted the experiment to separate the estimated speech from the noisy speech mixture with various noises. The experimental result does not show any significant improvement in speech as compared to the existing systems (Shoba and Rajavel 2017b, 2018a, b; Wang et al. 2013). It is known, the low-frequency filter channel may represent the voiced speech more dominantly than high-frequency filter channel and vice versa for the unvoiced speech. Hence there is a need for a scheme to give adaptive weights to the T–F units in the filter channel which in turn change the Eq. (11) as

$$Opt_{BM_{T-F}}(F_c, T_m) = \alpha * M_v(F_c, T_m) + (1 - \alpha) * M_{uv}(F_c, T_m) \quad \text{for } 0 \leq \alpha \leq 1 \quad \text{for } 1 \dots N_c \quad \text{and} \quad (12)$$

$$Opt_{BM_{T-F}}(F_c, T_m) = \alpha * M_{uv}(F_c, T_m) + (1 - \alpha) * M_v(F_c, T_m) \quad \text{for } 0 \leq \alpha \leq 1 \quad \text{for } N_c + 1 \dots N_T,$$

where N_c is the number of filter channel where voiced speech dominates which is determined via experimentation and N_T is the total number of filter channels in $Opt_{BM_{T-F}}(F_c, T_m)$. The proposed Genetic Algorithm (GA) based scheme to effectively fuse the T–F units of voiced and unvoiced T–F binary mask is as follows:

Algorithm**Initialization:**

Generate initial population (weight factor α) of size M

Population set $P_K = (\alpha_1, \alpha_2, \alpha_3, \alpha_4, \dots, \alpha_M)$ where K denotes iteration

Fitness Calculation:

For each weight (α_i) $i \in 1, \dots, M$, compute the binary T-F mask using the following equation

$$Opt_{BM}(F_c, T_m) = \alpha_i * M_v(F_c, T_m) + (1 - \alpha_i) * M_{uv}(F_c, T_m) \text{ for } 1 \dots N \text{ and}$$

$$Opt_{BM}(F_c, T_m) = \alpha_i * M_{uv}(F_c, T_m) + (1 - \alpha_i) * M_v(F_c, T_m) \text{ for } N + 1 \dots N_T;$$

Calculate fitness (SNR_i)(quality) (Wang et al. 2013) and ($STOI_i$) (intelligibility) (Taal et al. 2011) $i \in 1, \dots, M$, for each $i \in P_K$

do

{

New Population:

Selection: Select two best weights (α_{b1}, α_{b2}) $\in \{ \alpha_1, \alpha_2, \alpha_3, \alpha_4, \dots, \alpha_M \}$ of P_K and place into P_{K+1}

Crossover: Select two best weights (α_{b1}, α_{b2}) of P_K ; pair them up; produce offspring; place the offspring into P_{K+1}

Mutation: Select two random weights (α_{r1}, α_{r2}) $\in \{ \alpha_1, \alpha_2, \alpha_3, \alpha_4, \dots, \alpha_M \}$ of (P_{K+1}); place a randomly chosen bit in each α_{r1}, α_{r2} .

Evaluate P_{K+1} :

Calculate fitness (SNR_i)(quality) [41] and ($STOI_i$)(intelligibility) [35] $i \in 1, \dots, M$, for each $i \in P_{K+1}$

Increment Iteration:

$K := K + 1;$

}

While

$K \geq 100$; Terminate the program

return

The individual optimal weight factor α_i in P_K

Finally, the above proposed algorithm returns the optimal weight factor α which produce the speech with good quality and intelligibility.

2.5 Reconstruction of the speech signal

The final stage in the proposed research is to reconstruct the speech signal from the T-F binary mask obtained

using a Genetic Algorithm (GA) based fusion scheme. The re-construction of the speech signal used in this work follows the same procedure proposed by Weintraub (1985). Figure 4 represents the spectrogram of the (a) Clean speech signal “I ll willingly marry Marilyn”. (b) Noisy speech mixture (White noise) at 0dB SNR (c) Enhanced speech signal of using IASG system. (d) Enhanced speech using the proposed Genetic Algorithm (GA) based fusion scheme. From the observation, the final segregated speech from the proposed algorithm contains both voiced and unvoiced speech.

3 Evaluation results and comparison

The evaluation of the proposed speech segregation system is done using 10 speech utterances from Cooke (1993) database and 17 noises (Wang et al. 2013) are mixed at specific SNR ranges to yield 170 ($17 \times 10 \times 1$) noisy speech utterances and at different SNR ranges ($-5, 0, 5, 10$ and 15 dB) to yield 835 ($17 \times 10 \times 5$) noisy speech utterances. Figure 5 shows the noises used for this research work and the proposed system is evaluated with quality measures such as Signal-to-noise-ratio (SNR) and Perceptual Evaluation of Speech Quality (PESQ) and intelligibility measures such as short-time objective intelligibility (STOI), coherence

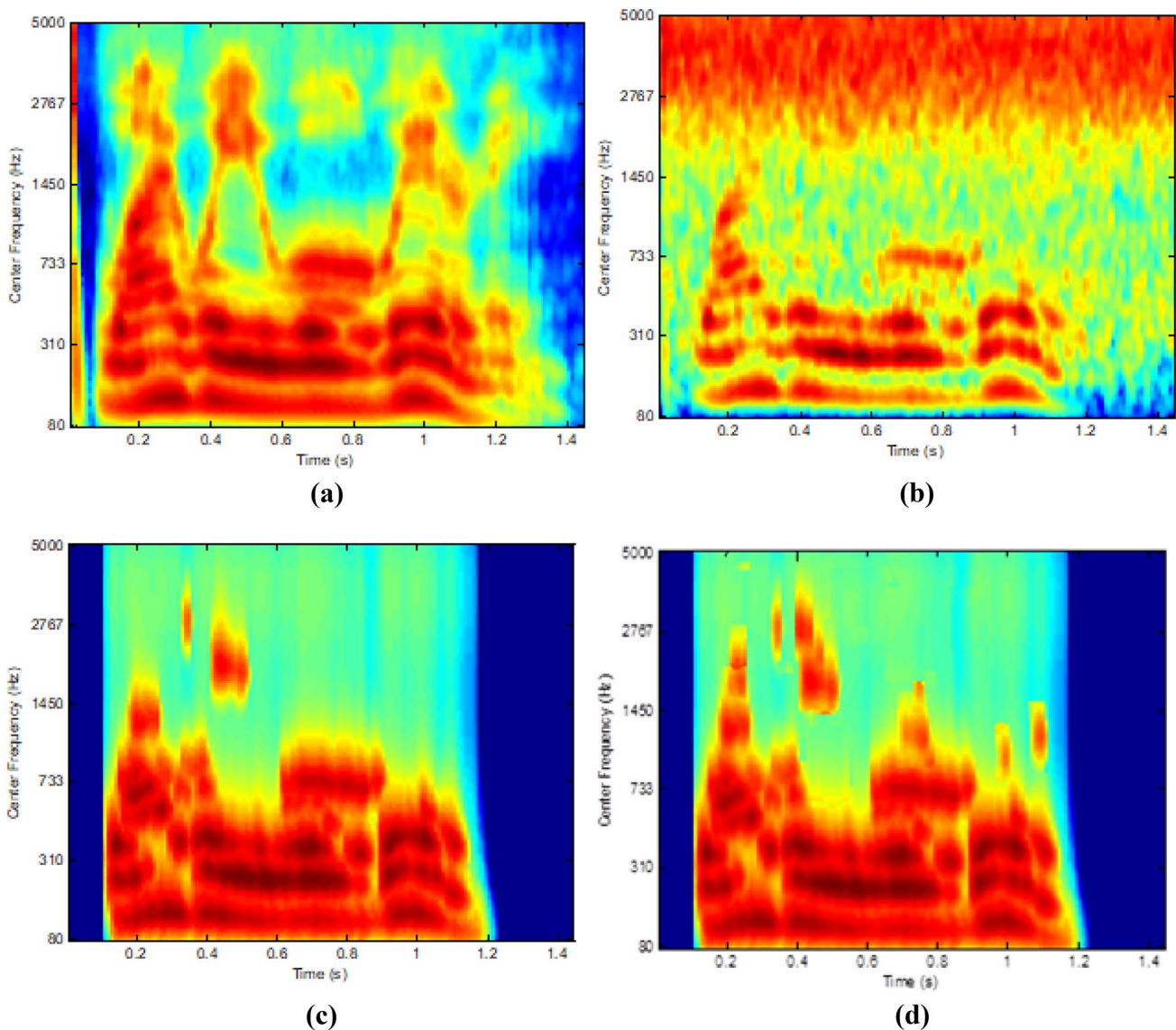


Fig. 4 Spectrogram results of the (a) clean speech - “I ll willingly marry Marilyn” (b) noisy signal with white noise at 0 dB SNR (c) enhanced speech using IASG system (d) the proposed GA based fusion scheme

Noise	Description
N ₁	White Noise
N ₂	Rock Music
N ₃	Siren
N ₄	Telephone
N ₅	Electric Fan
N ₆	Alarm Clock
N ₇	Traffic noise
N ₈	Bird Chirp with Water Flow
N ₉	Wind Noise
N ₁₀	Rain
N ₁₁	Cocktail Party
N ₁₂	Crowd Noise at a Playground
N ₁₃	Crowd Noise with Music
N ₁₄	Crowd Noise with Clap
N ₁₅	Babble Noise
N ₁₆	Male Speech
N ₁₇	Female Speech

Fig. 5 Description of noise

speech intelligibility index (CSII), and normalized covariance metric (NCM) (Hu and Loizou 2007; Ma et al. 2009; Taal et al. 2011). The performance of the proposed system is also measured using two error measures such as Percentage of energy loss (P_{el}), Percentage of noise residue (P_{nr}). The proposed system is measured in terms of both quality and intelligibility. Since the existing (Wang–Yu) Morphological

(MIPS) system (Wang et al. 2013) and (Hu–Wang) pitch tracking (PT) system (Hu and Wang 2004) evaluation results might be an appropriate one for comparing the quality measure with the proposed system and may not be an appropriate one for comparing intelligibility measure with the proposed system. Hence, these two systems have been re-evaluated using the Cooke Speech database and 17 noise and then the results (quality and intelligibility) are compared with the proposed system. The proposed work is implemented using a personal computer with an Intel® Core™ (i5-3230M) 2.60 GHz processor, 8 GB RAM. The software used is MATLAB (R2016b) installed in Windows 10 64 bit operating system.

The SNR is a direct criterion to measure the quality of speech and is determined as in Wang et al. (2013)

$$SNR = 10 \log \left(\frac{\sum_n S_{IN}(n)^2}{\sum_n (S_{IN}(n) - S_{OUT}(n))^2} \right), \tag{13}$$

where the original speech is named as $S_{IN}(n)$ and the segregated speech is named as $S_{OUT}(n)$. Figure 6 depicts that there is an improvement in SNR over all noises when compared with the existing system. The proposed system improves the SNR with an average value of 10.74 dB with input mixture, 3.47 dB with (Hu–Wang) pitch tracking (PT) system (Hu and Wang 2004), 2.19 dB with (Wang–Yu) Morphological (MIPS) system (Wang et al. 2013) and 0.83 dB with (Shoba–Rajavel) Image processing (IASG) system (Shoba and Rajavel 2017b) respectively. From the observation, the proposed system shows a high SNR improvement for noises

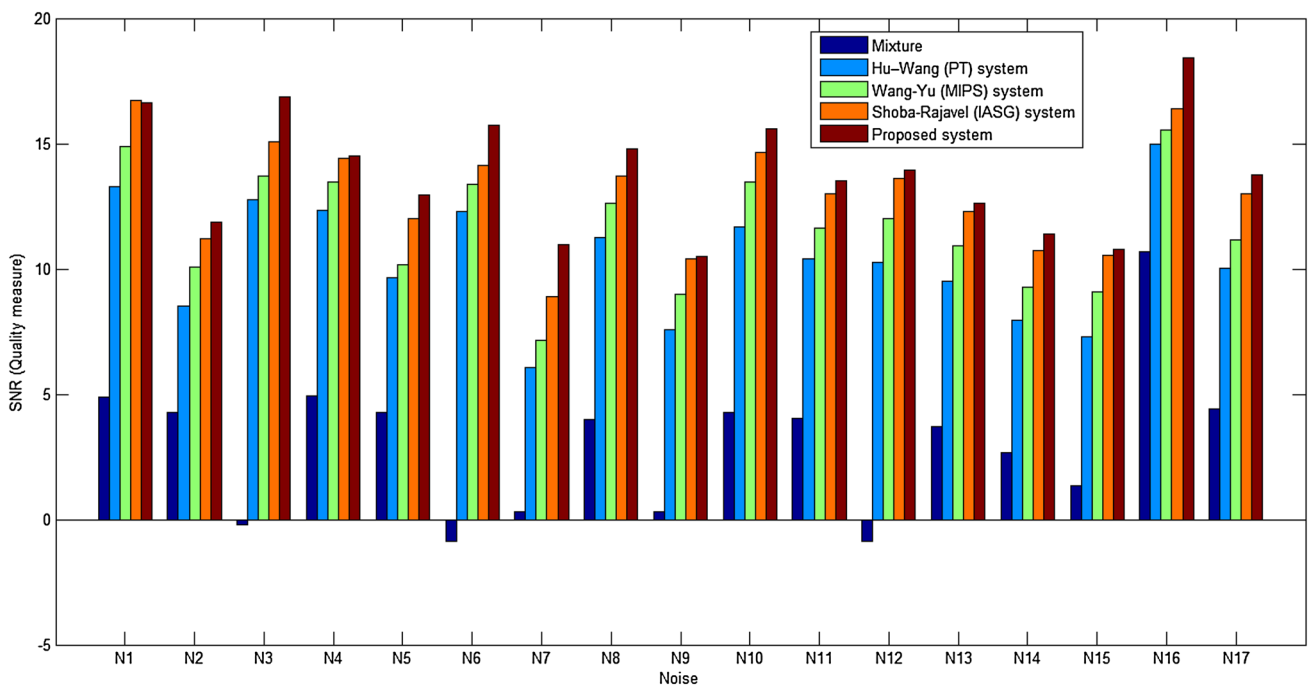


Fig. 6 Proposed system performance with the other systems (SNR(dB)-Quality)

N_3, N_7, N_{10}, N_{14} as it retains most of the speech information and there is a modest SNR improvement for noises $N_{11}, N_{15}, N_{16}, N_{17}$ due to significant spectral overlap between speech and noise. The proposed algorithm merges the unvoiced and voiced speech effectively using a Genetic Algorithm and in turn increases the quality of the speech when compared with the other systems in terms of SNR measure (Hu and Wang 2004; Wang et al. 2013).

An objective measure called perceptual evaluation of speech quality (PESQ) (Pichevar and Rouat 2005) is used to measure the segregated quality of speech for the proposed system. The bigger the PESQ worth suggests a greater performance approach. Figure 7 depicts the proposed system performance in comparison with other systems (Wang et al. 2013). The PESQ improvement for the proposed system increases with an average value of 0.4 with (Hu–Wang) pitch tracking (PT) system (Hu and Wang 2004), 0.2 with (Wang–Yu) Morphological (MIPS) system (Wang et al. 2013) and 0.08 with (Shoba–Rajavel) Image processing (IASG) system (Shoba and Rajavel 2017b) respectively. From the comparison, the PESQ results are superior for all types of noises and produce good quality of speech.

The above two common criterion SNR and PESQ relate to the speech quality measure and does not provide more information approximately how the segregated target speech is different from that of the clean speech signal. To do so, two complementary error measures P_{el} and P_{nr} are evaluated for the proposed speech segregation system to achieve low errors.

The percentage of energy loss P_{el} is measured by the quantity of target speech excluded from the segregated speech and determined by

$$P_{el} = \frac{\sum_n e_{s_1}^2(n)}{\sum_n I_s^2(n)}, \quad (14)$$

The percentage of noise residue P_{nr} is measured by the quantity of noise included in the segregated speech and determined by

$$P_{nr} = \frac{\sum_n e_{s_2}^2(n)}{\sum_n O_s^2(n)}, \quad (15)$$

$I_s(n)$ is the re-synthesized speech from IBM (Ideal Binary Mask). $O_s(n)$ is the re-synthesized speech from the proposed system.

e_{s_1} is the signal energy present in $I_s(n)$ but missing in $O_s(n)$.

e_{s_2} is the signal energy present in $O_s(n)$ but missing in $I_s(n)$.

The P_{el} and P_{nr} of the proposed system are evaluated and compared with other existing systems (Hu and Wang 2004; Wang et al. 2013) as shown in Table 1. Table 1 depicts that the energy loss is very less for all types of noises and the average value of energy loss P_{el} is 2.21% and the average value of noise residue P_{nr} is 2.23% which is less than (Hu–Wang) pitch tracking (PT) system (Hu and Wang

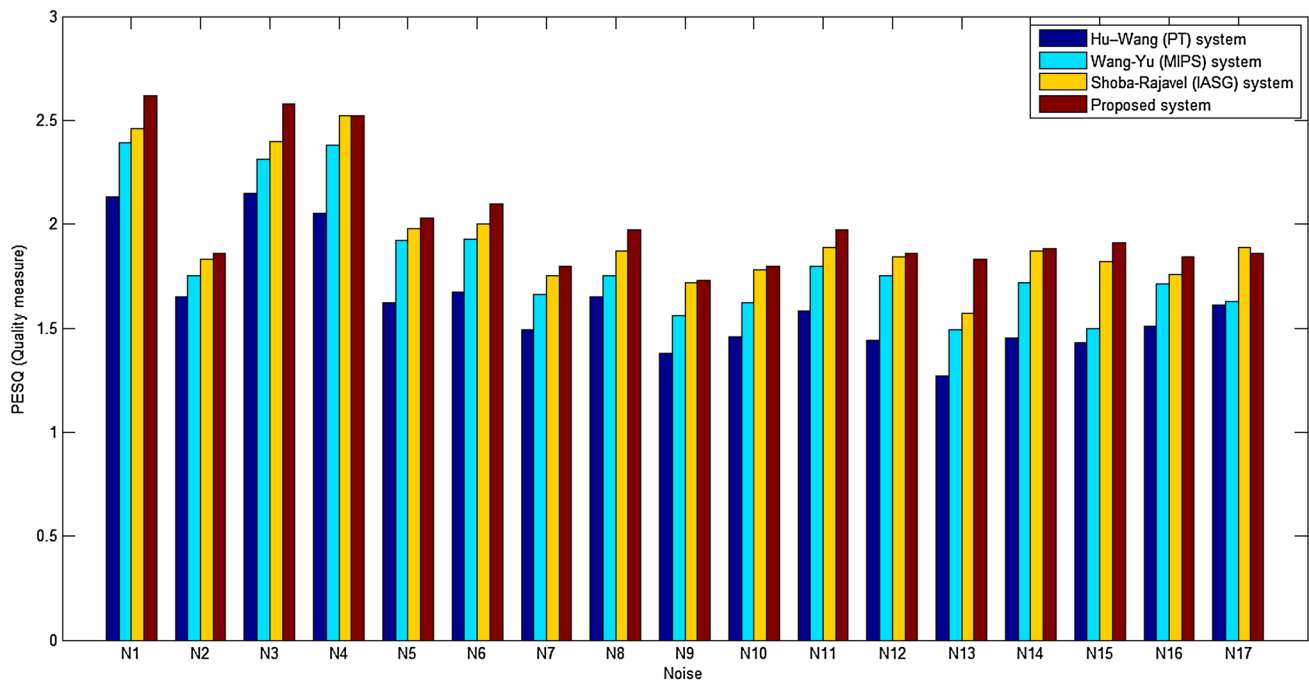


Fig. 7 Proposed system performance with the other systems (PESQ-Quality)

Table 1 Error measure - P_{el} and P_{nr} performance results of the proposed system and the other systems

Noise	Hu–Wang (PT) system		Wang–Yu (MIPS) System		Shoba–Rajavel (IASG) System		Proposed system		Mixture P_{nr} (%)
	P_{el} (%)	P_{nr} (%)	P_{el} (%)	P_{nr} (%)	P_{el} (%)	P_{nr} (%)	P_{el} (%)	P_{nr} (%)	
N_1	3.81	2.72	2.77	1.94	2.51	1.74	1.43	1.45	27.11
N_2	4.77	6.1	5.03	3.95	4.64	3.80	2.80	2.99	29.5
N_3	6.50	5.78	5.92	4.08	5.58	3.90	3.93	2.64	40.56
N_4	3.39	6.09	2.11	4.73	1.97	4.34	1.44	4.21	22.01
N_5	2.93	1.97	1.93	1.01	1.74	0.76	1.20	0.98	9.77
N_6	5.27	2.3	1.7	1.88	1.43	1.49	1.07	1.15	46.31
N_7	6.59	6.91	5.21	6.23	5.03	5.93	3.26	5.86	36.51
N_8	6.32	3.51	5.77	0.69	5.18	0.65	3.70	0.97	18.43
N_9	7.40	5.83	3.95	4.72	3.38	4.35	2.92	3.99	29.46
N_{10}	4.25	0.84	2.06	0.42	1.65	0.31	1.56	0.28	11.77
N_{11}	6.41	3.00	3.22	1.89	3.01	1.69	2.11	1.35	20.73
N_{12}	4.82	3.93	2.01	2.19	1.75	2.42	1.22	0.82	76.79
N_{13}	4.61	3.45	3.14	2.25	2.89	1.77	1.94	1.62	13.76
N_{14}	6.03	4.98	4.96	3.62	4.66	3.18	2.74	2.99	23.16
N_{15}	4.52	5.52	3.83	3.72	3.65	3.32	2.81	3.02	26.84
N_{16}	3.07	2.91	1.84	1.87	0.87	1.72	0.48	0.77	6.64
N_{17}	6.98	7.36	4.06	6.1	4.05	5.94	3.01	2.87	43.07
Average	5.16	4.3	3.5	3.02	3.18	2.78	2.21	2.23	28.38

2004), (Wang–Yu) Morphological (MIPS) system (Wang et al. 2013) and (Shoba–Rajavel) Image processing (IASG) system (Shoba and Rajavel 2017b). The proposed system of P_{nr} retains 97.77% ((100- P_{nr})%) of the original speech information.

The STOI intelligibility measure shown in Table 2 observed that the average STOI value for the proposed system shows improvement [(i.e) 0.91 for the proposed system, 0.84 for (Shoba–Rajavel) Image processing (IASG) system (Shoba and Rajavel 2017b), 0.79 for (Wang–Yu) Morphological(MIPS) system (Wang et al. 2013) and 0.74 for (Hu–Wang) pitch tracking (PT) system (Hu and Wang 2004)] when compared with other existing system. Also observed that, there is a very low improvement in STOI particularly for noises such as N_{11} , N_{16} , N_{17} since it contains noise components present in the target speech and there is a decent improvement in STOI for noises such as N_4 , N_7 , N_{10} and N_{14} since most of the target speech is segregated accurately by eliminating the interference completely. Another experiment has been conducted with different SNRs ranging from - 5, 0, 5, 10 and 15 dB for the proposed system and the other existing system for noise N_1 as shown in Fig 8. From the graphical representation of the Fig. 8, it is observed that the proposed system shows a gradual improvement in STOI for low SNR ranges such as -5dB, 0dB and 5dB and there is a very low improvement in STOI for high SNR ranges. The observation of Table 2 and Fig 8 concludes that the intelligibility (STOI) of the speech is increased even at low SNR ranges.

Table 2 The STOI (Intelligibility) of the proposed system and other systems at specific SNR range

Noise	Mixture	Hu–Wang (PT) system	Wang–Yu (MIPS) system	Shoba–Rajavel (IASG) system	Proposed system
N_1	4.88	0.81	0.82	0.83	0.88
N_2	4.29	0.71	0.75	0.8	0.86
N_3	- 0.20	0.76	0.85	0.90	0.97
N_4	4.95	0.83	0.87	0.89	0.97
N_5	4.27	0.74	0.83	0.90	0.97
N_6	- 0.85	0.78	0.87	0.91	0.98
N_7	0.3	0.81	0.85	0.87	0.95
N_8	3.97	0.75	0.79	0.81	0.87
N_9	0.33	0.69	0.71	0.77	0.81
N_{10}	4.25	0.81	0.84	0.88	0.94
N_{11}	4.04	0.63	0.70	0.77	0.81
N_{12}	- 0.86	0.65	0.79	0.85	0.91
N_{13}	3.69	0.7	0.75	0.77	0.83
N_{14}	1.36	0.71	0.75	0.79	0.92
N_{15}	2.67	0.66	0.73	0.80	0.87
N_{16}	10.69	0.82	0.82	0.86	0.88
N_{17}	4.41	0.71	0.74	0.87	0.91
Average	3.07	0.74	0.79	0.84	0.91

The proposed system is evaluated using another measure (intelligibility) called CSII as shown in Table 3. It is observed from Table 3 the average CSII value for the

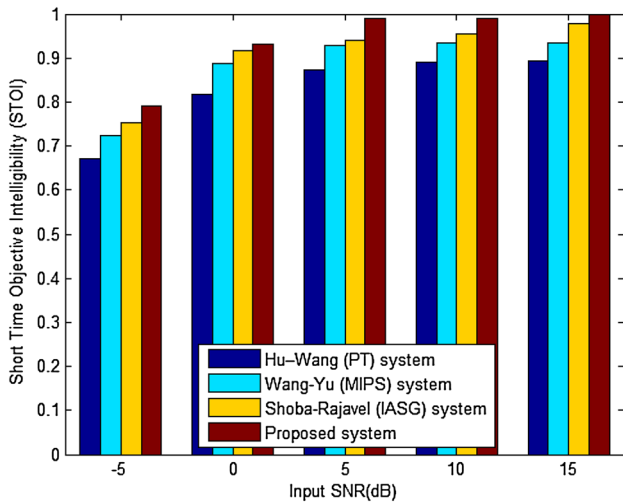


Fig. 8 The performance of the proposed system and other system (STOI) intelligibility measures at different SNR ranges

Table 3 The CSII (Intelligibility) of the proposed system and other systems at specific SNR range

Noise	Mixture	Hu-Wang (PT) system	Wang Yu (MIPS) System	Shoba-Rajavel (IASG) System	Proposed System
N_1	4.88	0.65	0.75	0.81	0.9
N_2	4.29	0.70	0.75	0.83	0.87
N_3	-	0.20	0.69	0.85	0.90
N_4	4.95	0.67	0.76	0.86	0.89
N_5	4.27	0.68	0.69	0.79	0.93
N_6	-	0.85	0.51	0.60	0.67
N_7	0.3	0.73	0.81	0.87	0.91
N_8	3.97	0.71	0.79	0.87	0.92
N_9	0.33	0.68	0.76	0.85	0.88
N_{10}	4.25	0.73	0.81	0.87	0.96
N_{11}	4.04	0.57	0.69	0.81	0.86
N_{12}	-	0.86	0.66	0.73	0.83
N_{13}	3.69	0.68	0.77	0.84	0.88
N_{14}	2.67	0.67	0.73	0.86	0.92
N_{15}	1.36	0.64	0.79	0.86	0.89
N_{16}	10.69	0.75	0.87	0.92	0.94
N_{17}	4.41	0.64	0.76	0.88	0.88
Average	3.07	0.67	0.75	0.83	0.89

proposed system shows improvement ((i.e) 0.89 for the proposed system, 0.83 for (Shoba-Rajavel) Image processing (IASG) system (Shoba and Rajavel 2017b), 0.75 for (Wang-Yu) Morphological (MIPS) system (Wang et al. 2013) and 0.67 for (Hu-Wang) pitch tracking (PT) system (Hu and Wang 2004)) when compared with other existing

system. The keen observation of Table 3 depicts that the target speech components are segregated accurately from the noisy mixture for noises such as N_5 , N_6 and N_{10} as compared with other noises. The performance of the proposed system and other existing system for noise N_1 is measured in a different view by conducting experiments at different SNR ranges such as -5 to 15dB in steps of 5dB as shown in Fig 9. The Fig 9 of the proposed system shows a gradual improvement over SNR ranges of -5dB, 0dB and 5dB and if SNR range increases 5dB, the improvement of CSII value is low because it fails to suppress the noise-dominant units at high SNR levels.

Apart from intelligibility measures such as STOI and CSII, another intelligibility measure called NCM is evaluated as shown in Table 4. The performance of the NCM measure in Table 4 shows an average value of 0.82 for the proposed system, 0.76 for (Shoba-Rajavel) Image processing (IASG) system (Shoba and Rajavel 2017b), 0.68 for (Wang-Yu) Morphological (MIPS) system (Wang et al. 2013) and 0.61 for (Hu-Wang) pitch tracking (PT) system (Hu and Wang 2004) respectively. Similar to Figs. 8 and 9, the NCM measure has been evaluated at different SNR ranges from -5 to 15 dB. Figure 10 depicts that there is a significant improvement in NCM value for low SNR ranges when compared with high SNR ranges. From the experimental results, it is proved that the unvoiced and voiced speech T-F units are segregated effectively by the proposed Genetic Algorithm (GA) based fusion and in turn improves both the quality and intelligibility of the speech.

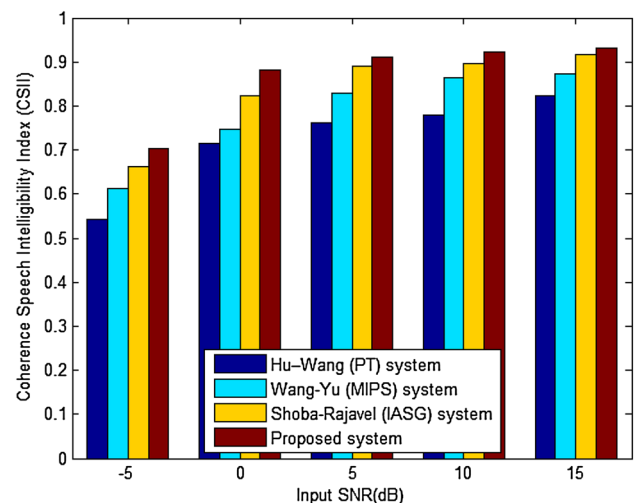
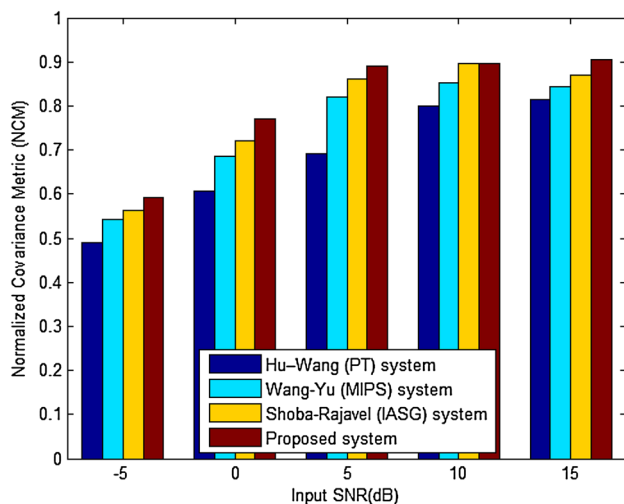


Fig. 9 The performance of the proposed system and other system (CSII) intelligibility measures at different SNR ranges

Table 4 The NCM (Intelligibility) of the proposed system and other systems at specific SNR range

Noise	Mixture	Hu–Wang (PT) system	Wang–Yu (MIPS) system	Shoba–Rajavel (IASG) system	Proposed system
N_1	4.88	0.64	0.72	0.85	0.89
N_2	4.29	0.51	0.6	0.69	0.73
N_3	-0.20	0.65	0.7	0.78	0.82
N_4	4.95	0.8	0.87	0.92	0.95
N_5	4.27	0.69	0.72	0.82	0.86
N_6	-0.85	0.52	0.6	0.67	0.72
N_7	0.3	0.67	0.68	0.76	0.89
N_8	3.97	0.57	0.65	0.71	0.75
N_9	0.33	0.61	0.65	0.73	0.79
N_{10}	4.25	0.69	0.73	0.85	0.9
N_{11}	4.04	0.56	0.67	0.74	0.8
N_{12}	-0.86	0.57	0.66	0.75	0.81
N_{13}	3.69	0.63	0.67	0.70	0.76
N_{14}	2.67	0.65	0.72	0.76	0.87
N_{15}	1.36	0.52	0.6	0.66	0.72
N_{16}	10.69	0.54	0.69	0.72	0.77
N_{17}	4.41	0.68	0.78	0.85	0.89
Average	3.07	0.61	0.68	0.76	0.82

**Fig. 10** The performance of the proposed system and other system (NCM) intelligibility measures at different SNR ranges

4 Conclusion

This research work proposed a Genetic Algorithm (GA) based fusion scheme to segregate the target speech from the noisy speech mixture. The proposed work computes the T–F binary mask representing the unvoiced speech

and effectively fused this binary mask with another mask representing the voiced speech using a Genetic Algorithm. The Genetic Algorithm (GA) based fusion enhances the T–F binary mask by representing both voiced and unvoiced speech and in turn, shows significant improvement in speech quality and intelligibility. The experimental results of the proposed system show improvement in SNR and PESQ by increasing the quality of speech and considerably reduce the noise residue and energy loss by deteriorating the P_{el} and P_{nr} value and also show a significant improvement in speech intelligibility as compared with the other existing systems. The Genetic Algorithm (GA) based fusion scheme gives the accurate optimal solution for fusing the voiced and unvoiced speech for a known set of speech and noise samples and may not be an optimal one for unknown speech and noise samples. The future course of this research will focus on this issue and propose a training based framework utilizing a neural network to deliver the optimal solution for all speech samples.

References

- Boll SF (1979) Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans Audio Speech Signal Process* 27:113–120
- Brown GJ, Cooke MP (1994) Computational auditory scene analysis. *Comput Speech Lang* 8:297–336
- Brown GJ, Wang DL (2005) Separation of speech by computational auditory scene analysis. In: Benesty J, Makino S, Chen J (eds) *Speech enhancement*. Springer, Berlin, pp 371–402
- Cooke MP (1993) *Modeling auditory processing and organization*. Dissertation, University of Sheffield, UK
- Dharmalingam M, JohnWiselin MC (2017) CASA for improving speech intelligibility in monaural speech separation. *Int J Perform Eng* 13(3):259–263
- Donald S, Wang D (2017) Time-frequency masking in the complex domain for speech dereverberation and denoising. *IEEE/ACM Trans Audio Speech Lang Process* 25(7):1492–1501
- Ephraim Y, Trees HL (1995) A signal subspace approach for speech enhancement. *IEEE Trans Speech Audio Process* 3:251–266
- Ellis DPW, Weiss RJ (2006) Model-based monaural source separation using a vector-quantized phase-vocoder representation. In *Proceedings on IEEE international conference on acoust speech and signal processing (ICASSP)*, pp 957–960
- Gibak K, Loizou PC (2010) Improving speech intelligibility in noise using environment-optimized algorithms. *IEEE Trans Audio Speech Lang Process* 18(8):2080–2090
- Harish N, Rajavel R (2014) Monaural speech separation system based on optimum soft mask. *IEEE Int Conf Comput Intell Comput Res*. <https://doi.org/10.1109/ICCIC.2014.7238420>
- Hu G, Wang D (2006) An auditory scene analysis approach to monaural speech segregation. In: Hansler E, Schmidt G (eds) *Topics in acoustic echo and noise control*. Springer, New York, pp 485–515
- Hu G, Wang D (2007) Auditory segmentation based on onset and offset analysis. *IEEE Trans Audio Speech Lang Process* 15(2):396–405
- Hu K, Wang D (2011) Unvoiced speech segregation from non-speech interference via CASA and spectral subtraction. *IEEE Trans Audio Speech Lang Process* 19(6):1600–1609

- Hu K, Wang D (2004) Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Trans Neural Netw* 15(5):1135–1150
- Hu Y, Loizou PC (2007) A comparative intelligibility study of speech enhancement algorithms. In: *Proceedings of IEEE international conference on acoustics speech and signal processing (ICASSP)*, pp 561–564
- Hyvarinen A, Karhunen J, Oja E (2001) *Independent component analysis*. Wiley, New York
- Jensen J, Hansen HL (2001) Speech enhancement using a constrained iterative sinusoidal model. *IEEE Trans Speech Audio Process* 9:731–740
- Hu K, Wang D (2013) An unsupervised approach to cochannel speech separation. *IEEE Trans Audio Speech Lang Process* 21(1):122–131
- Yi-nan L, xiong-wei Zhang, Zeng L, Huang JJ (2014) An improved monaural speech enhancement algorithm based on sparse dictionary learning. *J Signal Process* 30(1):44–50
- Ma J, Hu Y, Loizou P (2009) Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *J Acoust Soc Am* 125(5):3387–3405
- Meddis R (1988) Simulation of auditory-neural transduction: further studies. *J Acoust Soc Am* 83(3):1056–1063
- Naik R, Ganesh R, Wang W (2012) Audio analysis of statistically instantaneous signals with mixed Gaussian probability distributions. *Int J Electron* 99(10):1333–1350
- Naik R, Ganesh R (2012) Measure of quality of source separation for sub and super-Gaussian audio mixtures. *Informatica* 23(4):581–599
- Nilesh M, Ann S et al (2013) The potential for speech intelligibility improvement using the ideal binary mask and the ideal Wiener filter in single channel noise reduction systems: application to auditory prostheses. *IEEE Trans Audio Speech Lang Process* 21(1):63–72
- Patterson RD, Nimmo-Smith I, Holdsworth J et al (1988) An efficient auditory filterbank based on the gammatone function. *MRC Applied Psych Unit*
- Phapatanaburi K, Wang L, Oo Z et al (2017) Noise robust voice activity detection using joint phase and magnitude based feature enhancement. *J Ambient Intell Hum Comput* 8(6):845–859
- Pichevar R, Rouat J (2005) A quantitative evaluation of a bio-inspired sound segregation technique for two and three-source mixtures. In: *Chollet G, Esposito A, Faundez-Zanuy M, Marinaro M (eds) Nonlinear speech modeling and applications, vol 3445. Lecture notes in computer science*. Springer, Berlin, pp 430–435
- Qazi KA, Nawaz T, Mehmood Z, Rashid M, Habib HA (2018) A hybrid technique for speech segregation and classification using a sophisticated deep neural network. *PLoS ONE* 13(3):e0194151. <https://doi.org/10.1371/journal.pone.0194151>
- Rajavel R, Sathidevi PS (2012) Adaptive reliability measure and optimum integration weight for decision fusion audio-visual speech recognition. *J Signal Process System* 68(1):83–93
- Rajavel R, Sathidevi PS (2011) A new GA optimised reliability ratio based integration weight estimation scheme for decision fusion audio-visual speech recognition. *Int J Signal Imaging Syst Eng* 4(2):123–131
- Sameti H, Sheikhzadeh H, Deng L, Brennan RL (1998) HMM-based strategies for enhancement of speech signals embedded in non-stationary noise. *IEEE Trans Speech Audio Process* 6:445–455
- Shoba S, Rajavel R (2017) Adaptive energy threshold selection for monaural speech separation. In: *International conference on communication and signal processing (ICCSP)*, India, pp 905–908
- Shoba S, Rajavel R (2017) Image processing techniques for segments grouping in monaural speech separation. *Circ Syst Signal Process* 37(8):3651–3670
- Shoba S, Rajavel R (2018) Improving speech intelligibility in monaural segregation system by fusing voiced and unvoiced speech segments circuits systems and signal process. *Circ Syst Signal Process*. <https://doi.org/10.1007/s00034-018-1005-3>
- Shoba S, Rajavel R (2018) Performance improvement of monaural speech separation system using image analysis techniques. *IET Signal Process* 12(7):896–906
- Singhal S, Passricha V, Sharma P et al (2018) Multi-level region-of-interest CNNs for end to end speech recognition. *J Ambient Intell Hum Comput*. <https://doi.org/10.1007/s12652-018-1146-z>
- Taal CH, Hendriks RC, Heusdens R et al (2011) An algorithm for intelligibility prediction of time frequency weighted noisy speech. *IEEE Trans Audio Speech Lang Process* 19(7):2125–2136
- Therese SS, Lingam C (2017) A linear visual assessment tendency based clustering with power normalized cepstral coefficients for audio signal recognition system. *J Ambient Intell Hum Comput*. <https://doi.org/10.1007/s12652-017-0653-7>
- Trowitzsch Ivo (2017) Robust detection of environmental sounds in binaural auditory scenes. *IEEE/ACM Trans Audio Speech Lang Process* 25(6):1344–1356
- Wang DL, Kun H (2013) Towards generalizing classification based speech separation. *IEEE Trans Audio Speech Lang Process* 21(1):68–77
- Wang D (2012) Tandem algorithm for pitch estimation and voiced speech segregation. *IEEE Trans Audio Speech Lang Process* 18(8):2067–2079
- Wang DL, Brown GJ (1999) Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Trans Neural Netw* 10:684–697
- Wang Y, Lin J, Chen N, Yuan W (2013) Improved monaural speech segregation based on computational auditory scene analysis. *J Audio Speech Music Process*. <https://doi.org/10.1186/1687-4722-2013-2>
- Weintraub M (1985) A theory and computational model of auditory monaural sound separation. Ph.D. dissertation, Dept Elect Eng, Stanford University
- Yu J, Xie L, Xiao X et al (2017) A hybrid neural network hidden Markov model approach for automatic story segmentation. *J Ambient Intell Hum Comput* 8(6):925–936
- Zhang X, Wang DL (2017) Deep learning based binaural speech separation in reverberant environments. *IEEE/ACM Trans Audio Speech Lang Process* 25(5):1075–1084

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.