IET Signal Processing

Research Article

IET Journals
The Institution of Engineering and Technology

Performance improvement of monaural speech separation system using image analysis techniques

ISSN 1751-9675
Received on 4th August 2017
Revised 22nd March 2018
Accepted on 28th March 2018
E-First on 6th June 2018
doi: 10.1049/iet-spr.2017.0375
www.ietdl.org

Shoba Sivapatham[1] ✉, *Rajavel Ramadoss*[1]

[1]*Department of Electronics and Communication Engineering, SSN College of Engineering, Tamil Nadu, India*

✉ *E-mail: shobansb@gmail.com*

Abstract: This research work proposes an image analysis-based algorithm to enhance the time–frequency (*T–F*) mask obtained in the initial segmentation of CASA-based monaural speech separation system to improve speech quality and intelligibility. It consists of labelling the initial segmentation mask, boundary extraction, active pixel detection and eliminating the non-active pixels related to noise. In labelling, the *T–F* mask obtained is labelled as periodicity pixel (***P***) matrix and non-periodicity pixel (***NP***) matrix. Next speech boundaries are created by connecting all the possible nearby ***P*** and ***NP*** matrix. Some speech boundary may include noisy *T–F* units as holes; these holes are treated using the proposed algorithm. The proposed algorithm is evaluated with the quality and intelligibility measures such as signal to noise ratio (SNR), perceptual evaluation of speech quality, P_{EL}, P_{NR}, coherence speech intelligibility index (CSII), normalised covariance metric (NCM), and short-time objective intelligibility (STOI). The experimental results show that the proposed algorithm improves the speech quality by increasing the SNR with an average value of 9.91 dB and reduces the P_{NR} by an average value of 25.6% and also improves the speech intelligibility in terms of CSII, NCM, and STOI when compared with the input noisy speech mixture

1 Introduction

In a natural environment, speech signal is accompanied with multiple sound sources. The process of separating the target speech from the noisy speech mixture remains a challenging task for machines. However, human being has an inherent capability to separate the speech from the noisy mixture. Researchers made tremendous effort to develop an effective monaural speech separation system that automatically separates the target speech from the noisy mixture recorded using single microphone. These systems have potential applications in automatic speech and speaker recognition, speaker identification, audio information retrieval, hearing aids, automatic music transcription, digital content management etc. For decades, various methods have been proposed for monaural speech separation and some of them are speech enhancement, subspace analysis, model-based approaches, and CASA. In speech enhancement, the statistical property of speech and noise is used to enhance the speech that is degraded by additive noise [1, 2]. In subspace analysis, eigen decomposition of acoustic mixture has been done and then either independent component analysis [3] or principal component analysis are used to remove the interference from the noisy mixture [4]. In model-based approaches, trained models of speech and noise are used for separation [5, 6]. However, all these methods require some form of prior knowledge about speech or interference, whereas in real scenario, prior knowledge of speech or interference is not possible. In 1990, Bregman has proposed the concept of auditory scene analysis (ASA) since then many researchers have adapted ASA for monaural speech separation is generally known as CASA systems [7–13]. Wang–Brown proposed a CASA system [14] which uses cross-channel correlation and temporal continuity as major cues to segregate the voiced speech. Another CASA system known as tandem algorithm [15] has been proposed by Wang for voiced speech segregation using pitch estimation. Similarly, Hu–Wang proposed a CASA system using energy of a time–frequency (*T–F*) unit as a primary cue [16] and onset and offset [17] to separate the target speech from the monaural noisy mixture. CASA systems for unvoiced speech segregation using spectral subtraction have been proposed by Hu–Wang and Boll [18, 19]. Recently, Wang–Yu proposed an improved CASA-based algorithm [20] for monaural speech separation using morphological image processing to separate the target speech.

All the above CASA-based system in general contains the following, (i) speech analysis, (ii) feature extraction, (iii) segmentation, and (iv) grouping and re-synthesis. Speech analysis means the input noisy speech signal is decomposed into various *T–F* units using filtering and windowing techniques. The auditory cues, such as cross-channel correlation, correlogram, energy etc., are extracted as features in the feature extraction stage. Based on the extracted features, *T–F* units are segmented into speech-dominant and noise-dominant in the segmentation stage. In the grouping stage, segments originated from the speech source are grouped together and similarly segments from the noise source are grouped together. Finally, speech signal is re-synthesised from the speech segments obtained from the grouping stage. The initial segmentation stage plays an important role in monaural speech separation process. The *T–F* mask obtained in the initial segmentation stage may not be appropriate due to wrongly interpreted *T–F* units. For example, sometimes, the speech *T–F* units may be interpreted as noisy *T–F* units and denoted as 0 and noisy *T–F* units may be interpreted as speech *T–F* units and denoted as 1 and this leads to poor speech separation. In order to solve this issue, this research work proposes an image analysis-based algorithm to enhance the mask obtained from the previous initial segmentation stage. The proposed algorithm removes the noise-dominant units using active pixel detection and complements the missing speech elements using boundary extraction which in turn improves the speech quality. The performance of the proposed algorithm is measured in terms of speech quality measures such as signal to noise ratio (SNR), percentage of energy loss (P_{EL}), percentage of noise ratio (P_{NR}), and perceptual evaluation of speech quality (PESQ) and intelligibility measures such as coherence speech intelligibility index (CSII), normalised covariance metric (NCM), and short-time objective intelligibility (STOI). The experimental results show that the proposed algorithm improves the speech quality in terms of increasing the SNR, PESQ, and reducing the P_{NR}. Similarly, the proposed algorithm also improves the speech intelligibility in terms of increasing the value of CSII, NCM, and STOI with respect to the input noisy speech mixture.

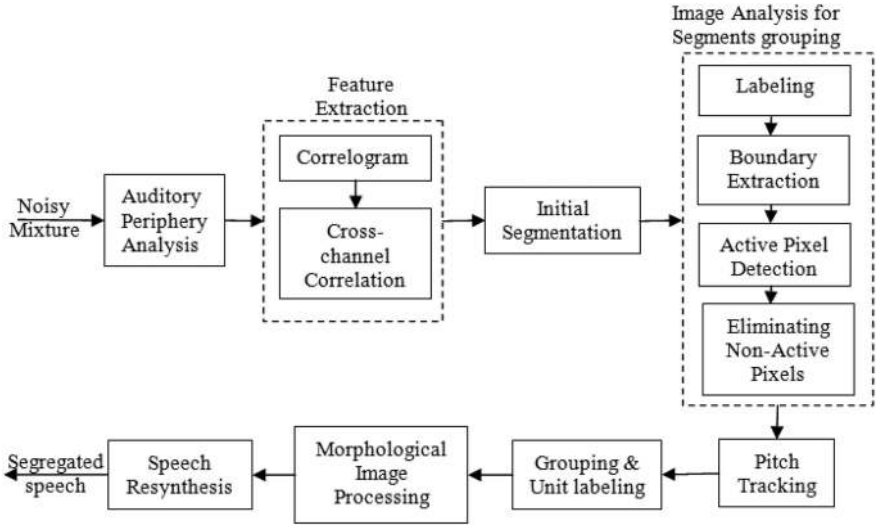

Fig. 1 *Proposed speech separation system*

The rest of the paper is organised as follows. Section 2 broadly explains the proposed speech separation system. Speech database, intelligibility and quality measures, and the experimental results are discussed in Section 3. Finally, the summary and the possible future extension of this research work are presented in Section 4.

2 Proposed speech separation systems

The block schematic of the proposed image analysis-based speech separation system is shown in Fig. 1. It consists of auditory periphery analysis, feature extraction, initial segmentation, image analysis for segments grouping, pitch tracking, grouping and unit labelling, morphological image processing, and finally the resynthesis of speech using synthesis filter bank. The following subsections briefly describe the above-mentioned stages of the proposed speech separation system.

2.1 Auditory periphery analysis

The auditory periphery is the first stage in the proposed speech separation system. It decomposes the noisy speech into various *T–F* units. It has been implemented through a bank of 128 Gammatone filterbank whose centre frequency varies from 80 to 5000 Hz and its impulse response is given as [21]

$$g(f,t)=\begin{cases} b^N t^{N-1}\mathrm{e}^{-2\pi bt} & \cos(2\pi f_c t), t\geq 0 \\ 0 & \text{Otherwise} \end{cases} \tag{1}$$

where $N=4$ is the order of the filter, b its equivalent rectangular bandwidth, and f_c the centre frequency. Let $s(t)$ be the input noisy speech signal and the response from the filter c is as [7]

$$y(c,t)=s(t)*g_c(f_c,t) \tag{2}$$

where * denotes linear convolution and the output of the filter is processed by Meddis model of hair cell transduction [22]. It simulates the non-linear process of the auditory nerve, such as rectification, saturation, and phase locking. The output of this model represents the firing rate of an auditory nerve fibre denoted as $h(c,t)$. The output from Meddis hair cell model of each filter channel is divided into 20 ms time-frames with 10 ms overlapping to generate a *T–F* unit which is denoted as $h(c,m)$, where c represents filter channel and m represents frame which constitute a two-dimensional *T–F* mask is called cochleogram.

2.2 Feature extraction

The next stage in the proposed speech separation system is the feature extraction. In this work, features such as energy, correlogram, and cross-channel correlation are extracted from each *T–F* units. The process of extracting these features is explained as follows.

2.2.1 Energy: The energy feature $E(c,m)$ is obtained from the autocorrelation of hair cell output $h(c,m)$ at zero lag, as follows:

$$E(c,m)=\frac{1}{N_c}\sum_{n=1}^{N_c-1} h(c,mT-n)*h(c,mT-n) \tag{3}$$

2.2.2 Correlogram: Correlogram is computed by auto correlating the hair cell response at the output of each cochlear filter channel [16]. The autocorrelation of filter response for each *T–F* unit is given as [7]

$$A_H(c,m,\tau)=\frac{1}{N_c}\sum_{n=1}^{N_c-1} h(c,mT-n)*h(c,mT-n-\tau) \tag{4}$$

where τ is the delay [0–1.25] ms, N_c the number of frames in the channel c, and n the sample index. An envelope correlogram is given as [7]

$$A_E(c,m,\tau)=\frac{1}{N_c}\sum_{n=1}^{N_c-1} h_E(c,mT-n)*h_E(c,mT-n-\tau) \tag{5}$$

where $h_E(c,n)$ is the envelope of the filter output in channel c and n the digitised time.

2.2.3 Cross-channel correlation: Cross-channel correlation between two adjacent filter channel response is measured in [16] as

$$C_H(c,m,\tau)=\frac{1}{N_c}\sum_{\tau=1}^{L-1} A_H(c,m,\tau)*A_H(c+1,m,\tau) \tag{6}$$

The cross-channel correlation of envelope is calculated in [16] as

$$C_E(c,m,\tau)=\frac{1}{N_c}\sum_{\tau=1}^{L-1} A_E(c,m,\tau)*A_E(c+1,m,\tau) \tag{7}$$

2.3 Initial segmentation

In this stage, the extracted features are used to segment each *T–F* unit into speech-dominant and noise-dominant *T–F* units. Hu–Wang proposed a CASA model [16] in which energy of a *T–F* unit is compared with a constant threshold θ_H^2. If the energy $E(c,m)$ of a particular *T–F* unit exceeds θ_H^2, the corresponding *T–F* unit is denoted by 1 to represent the speech-dominant, otherwise it is

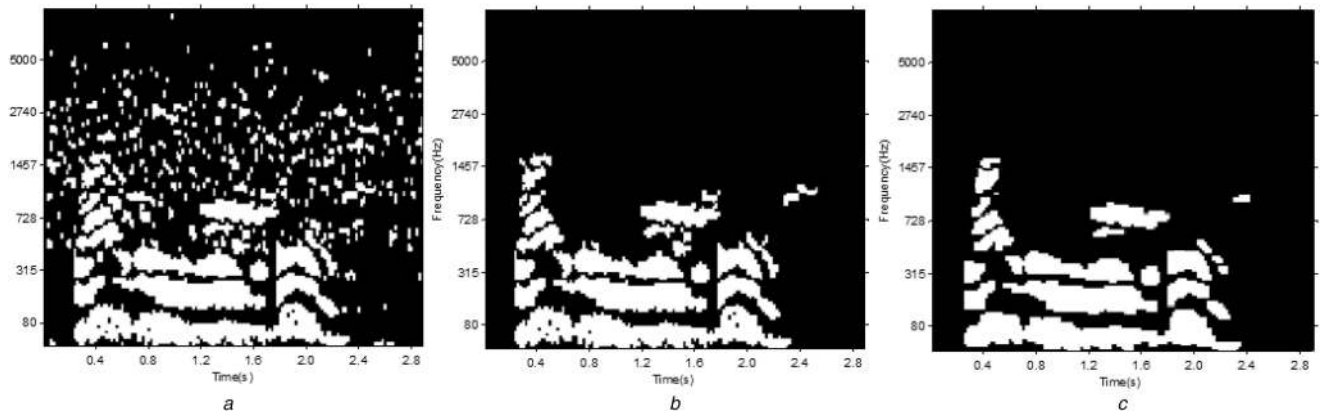

Fig. 2 *Binary mask*
(a) Before,
(b) After applying the proposed image analysis-based algorithm,
(c) After applying morphological image processing

denoted as 0 to represent the noise-dominant. Similarly, the cross-channel correlation is compared with a threshold $\theta_c = 0.985$ (chosen to be same as in [23]) and is 1 if it exceeds θ_c otherwise 0. This approach sometimes misses some of the *T–F* units in which speech is least dominant and reduces the speech quality and intelligibility. In order to solve this issue, an adaptive energy threshold selection algorithm is proposed in [24] to determine a particular *T–F* units belongs to speech-dominant or noise-dominant. The adaptive energy threshold $\theta_{\mathrm{AT}}(c)$ for each channel is computed as follows:

$$\theta_{\mathrm{AT}}(c) = \frac{1}{N_c * \beta} \sum_{m=1}^{N_c} E(c,m) \tag{8}$$

where N_c is the total number of time frame in each channel c and β the scaling constant which is determined via experimentation as $\beta = 1.7$. Finally, the binary mask $\boldsymbol{M}(c,m)$ is constituted as follows:

$$\boldsymbol{M}(c,m) = \begin{cases} 1 & \text{if } E(c,m) \geq \theta_{\mathrm{AT}}(c) \quad \text{and} \quad \boldsymbol{C}(c,m) \geq \theta_c, \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

In which 1 represents speech-dominant *T–F* units and 0 represents noise-dominant *T–F* unit as shown in Fig. 2*a*.

2.4 Image analysis for segments grouping

The binary mask generated in the initial segmentation may not be appropriate since some of the speech elements may be missed and some of the noise element may not be removed completely. This research work addresses this issue by enhancing the mask in the initial segmentation stage by an image analysis-based segment grouping algorithm. The proposed algorithm performs labelling the initial segmentation mask, boundary extraction, active pixel detection, and eliminating the non-active pixels to enhance the binary mask. The following subsection briefly explains each of the above steps in the proposed algorithm.

2.4.1 Labelling: The first step in the proposed algorithm is labelling the initial segmentation mask, in which the mask $\boldsymbol{M}(c,m)$ is considered as the initial mask. The process of labelling the initial segmentation mask $\boldsymbol{M}(c,m)$ is briefly explained as follows:

Initialisation

- Let $\boldsymbol{M}(c,m)$ be a binary image matrix where c represent channel and m represent frame and $\boldsymbol{B}_{\mathrm{SM}}$ is a square matrix of size 8×8 and is given as

$$\boldsymbol{B}_{\mathrm{SM}} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

- The binary submatrix obtained after multiplying with $\boldsymbol{B}_{\mathrm{SM}}$ with maximum number of 1 s and minimum number of 0 s is named as periodicity pixel matrix (***P***) and an example is shown in Fig. 3*a* highlighted in orange colour.
- In the same process, the binary submatrix minimum number of 1 s and maximum number of 0 s is named as non-periodicity pixel matrix (***NP***) and an example is shown in the same Fig. 3*a* highlighted in blue colour.

Labelling algorithm

i. Scan the binary image matrix $\boldsymbol{M}(c,m)$ by the square matrix $\boldsymbol{B}_{\mathrm{SM}}$ using raster scan technique and determine the resultant matrix as periodicity pixel matrix (***P***) or non-periodicity pixel matrix (***NP***).
ii. If the resultant matrix is a periodicity pixel matrix (***P***), then searches for its neighbouring periodicity pixel matrix (***P***) in all four directions.
iii. This scanning process continues until it finds the non-periodicity pixel matrix (***NP***) in any of the four directions and an example is shown in Fig. 3*a*.
iv. The submatrix having value 1 in both periodicity pixel matrix (***P***) and non-periodicity pixel matrix (***NP***) are grouped and labelled as N, initially $N = 1$.
v. Repeat steps (i)–(iv) for the remaining pixel values in the binary image matrix $\boldsymbol{M}(c,m)$ and name each new group with label N where $N = N + 1$.
vi. Finally, ensure that every pixel in the binary image matrix $\boldsymbol{M}(c,m)$ is grouped and labelled with unique label index.

The image matrix at the end of above labelling algorithm is denoted as $\boldsymbol{M}(c,m)$ which contains unique label for each group and an example is shown in Fig. 3*b*.

2.4.2 Boundary extraction: The next stage in the proposed segments grouping algorithm is the boundary extraction. In which the labelled matrix $\boldsymbol{M}_{\mathrm{L}}(c,m)$ is considered as the initial mask for further process in this stage. The labelled matrix $\boldsymbol{M}_{\mathrm{L}}(c,m)$ may lose some of the speech-dominant *T–F* units. This boundary extraction process focuses this issue and recovers the missing speech-dominant *T–F* units and in turn increases the speech intelligibility

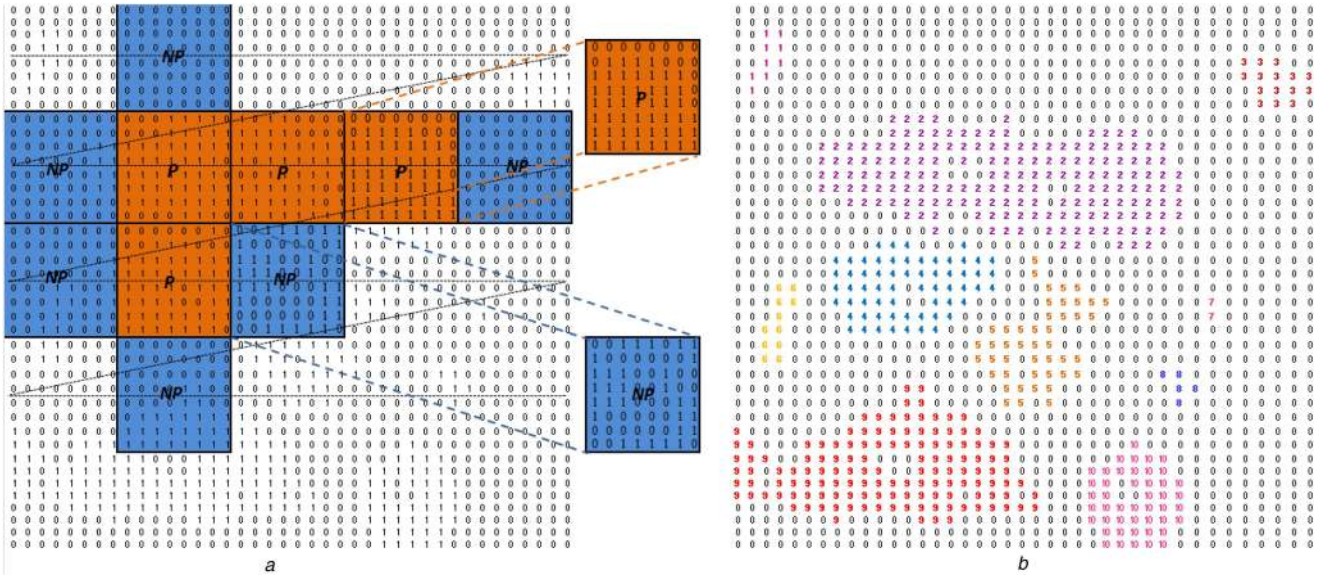

Fig. 3 *Labelling*

(a) Labelling: orange colour submatrix indicates P – periodicity pixel matrix and blue colour submatrix indicates NP – non-periodicity pixel matrix and black dotted lines in the matrix indicates the raster scan technique,

(b) Image labelled matrix $\boldsymbol{M}_{\mathrm{L}}(c, m)$ contains unique label for each group

and quality. The boundary extraction algorithm is briefly explained as follows:

Boundary extraction algorithm

i. Consider a labelled image matrix $\boldsymbol{M}_{\mathrm{L}}(c, m)$ as the initial mask.

ii. The first step in the boundary extraction algorithm is to create a concrete external boundary for each group by performing a dilation operation between the labelled matrix $\boldsymbol{M}_{\mathrm{L}}(c, m)$ and the square matrix $\boldsymbol{B}_{\mathrm{SM}}$. The resultant matrix is denoted as $\boldsymbol{M}_{D}(c, m)$ and is defined as follows:

$$\boldsymbol{M}_{D}(c, m) = \boldsymbol{M}_{\mathrm{L}}(c, m) \oplus \boldsymbol{B}_{\mathrm{SM}} \tag{10}$$

where $\oplus$ denotes dilation operation.

iii. In each group, there may be some pixels with zero values which are wrongly interpreted as noise-dominant *T–F* units since whose neighbouring pixels are speech-dominant *T–F* units as shown in Fig. 4*a*. These wrongly interpreted *T–F* units are called holes which can be represented in a better way by subtracting the labelled matrix $\boldsymbol{M}_{\mathrm{L}}(c, m)$ from the boundary extracted matrix $\boldsymbol{M}_{D}(c, m)$ and this process is defined as

$$\boldsymbol{M}_{H}(c, m) = \boldsymbol{M}_{D}(c, m) - \boldsymbol{M}_{\mathrm{L}}(c, m) \tag{11}$$

iv. Finally to recover the missing speech-dominant *T–F* units from the holes, a complement operation is performed with $\boldsymbol{M}_{H}(c, m)$ as in [20] as

$$\boldsymbol{M}_{\mathrm{CL}}(c, m) = (\boldsymbol{M}_{H}(c, m) \oplus \boldsymbol{B}_{\mathrm{SM}}) \ominus \boldsymbol{B}_{\mathrm{SM}} \tag{12}$$

where $\ominus$ denotes erosion operation and an example image matrix $\boldsymbol{M}_{\mathrm{CL}}(c, m)$ after boundary extraction process is shown in Fig. 4*b*.

2.4.3 Active pixel detection and elimination of non-active pixel: The next stage in the proposed algorithm is the active pixel detection and elimination of non-active pixels. In which, each group of the boundary extracted matrix $\boldsymbol{M}_{\mathrm{CL}}(c, m)$ is classified as active pixels group (speech-dominant *T–F* units) or non-active pixels group (noise-dominant *T–F* units) based on the number of layers and the number of active pixels. The process of active pixel detection and the elimination of non-active pixels are briefly explained as follows:

Initialisation

- The connected similar labelled index (for example, 1, 2, 3,… NL with different colour in Fig. 4*c*) in each group is named as layer.
- The maximum of the labelled index determines the number of layers (NL) in each group.

Active pixel detection and elimination of non-active pixel algorithm

i. Let $\boldsymbol{M}_{\mathrm{CL}}(c, m)$ be the boundary extracted image.

ii. Find the number of groups in $\boldsymbol{M}_{\mathrm{CL}}(c, m)$ and denote it as NG.

iii. In each group, find the number of layers NL_i and the number of active pixels (non-zero value) NP_i; $1 \leq i \leq \mathrm{NG}$.

iv. If the number of layers NL_i is >1 and the number of active pixels NP_i is greater than θ_{AP}, then the corresponding group $i \in 1 \leq i \leq \mathrm{NG}$, is called as active pixels group; otherwise, it is called non-active pixels group. The optimum value for θ_{AP} is determined by conducting various experiments with different speech and noise samples at different SNRs.

v. All the pixels in the active pixels group are replaced by 1 to represent it as speech-dominant and all the pixels in the non-active pixels group is replaced by 0 to represent it as noise-dominant and the resultant binary mask $\boldsymbol{L}_{M}(c, m)$ is shown in Fig. 4*d*.

The binary mask obtained before and after applying the proposed image analysis-based algorithm is shown in Figs. 2*a* and *b*, respectively.

2.5 Pitch tracking

The next stage in the proposed algorithm is the pitch tracking. Even though the proposed image analysis-based algorithm captures most of the speech-dominant *T–F* units but sometimes may miss one or two speech-dominant *T–F* units in which intrusion is more dominant than the target speech. This problem can be solved by performing the pitch tracking over the resultant binary image matrix $\boldsymbol{L}_{M}(c, m)$. The pitch of speech in every frame across each channel is computed by the summation of autocorrelation of the *T–F* units and is given in [16] as

$$\S_{F}(m, \tau) = \sum_{c} A_{H}(c, m, \tau) \tag{13}$$

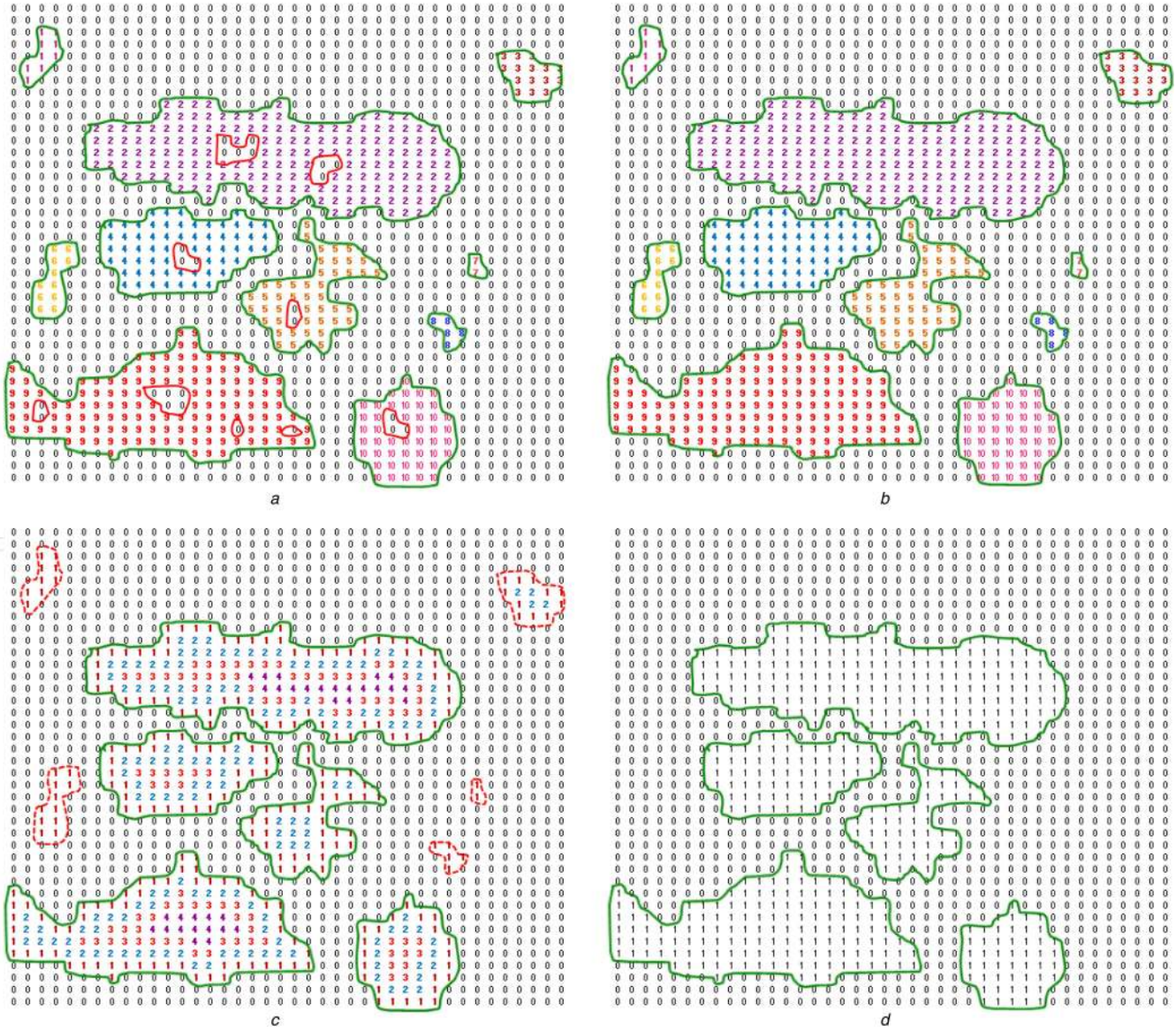

Fig. 4 *Boundary extraction*

(a) Speech-dominant units are indicated by green colour and holes indicated by red colour in each group shows the missing dominant speech units,

(b) Missing dominant speech units are recovered and substituted with the corresponding label index of that group,

(c) Active pixel detection,

(d) Elimination of non-active pixels

The pitch period of the target speech at frame m is $\tau_{s(m)}$, i.e. the lag corresponding to the maximum of $\S_F(m,\tau)$ in the possible pitch range [2–12.5 ms].

2.6 Grouping and unit labelling

The next stage is grouping and unit labelling of *T–F* units, in which each *T–F* unit is labelled as speech-dominant or noise-dominant based on the following rule proposed by Hu–Wang [16] as

$$\boldsymbol{M}_{\mathrm{GL}}(c,m)=\begin{cases}1 & \text{if } \dfrac{A_{H(c,m,\tau_{s(m)})}}{A_{H(c,m,\tau_{p(c,m)})}}>\theta_T\\ 0 & \text{otherwise}\end{cases} \tag{14}$$

where $\theta_T=0.85$ is chosen as same in [23], $\tau_s(m)$ the delay corresponding to the maximum of $A_H(c,m,\tau)$ within the possible pitch range [2–12.5 ms], and $\boldsymbol{M}_{\mathrm{GL}}(c,m)$ the binary mask obtained at the end of grouping and unit labelling.

2.7 Morphological image processing

The binary image matrix after grouping and unit labelling may contain some edges in which there will be an abrupt change of speech and noisy *T–F* units. To smoothen these edges and to improve the quality of speech, Yu *et al.* [20] used morphological image processing techniques. This research work also adopts the same technique as proposed by Yu *et al.* to improve the speech quality of the separated speech. The morphological image processing operations such as erosion, pruning, and complementing uses the following structuring element S_E as same in [20] as

$$S_E=\begin{pmatrix}1 & 1\\ 1 & 1\\ 1 & 1\end{pmatrix}$$

2.7.1 Erosion: Erosion is an operation that shrinks the inner and outer region binary image matrix $\boldsymbol{M}_{\mathrm{GL}}(c,m)$ with the structuring element S_E is as follows

$$\boldsymbol{E}(c,m)=\boldsymbol{M}_{\mathrm{GL}}(c,m)\ominus S_E \tag{15}$$

Table 1 O/P SNR of the proposed system for the specific input SNRs

Intrusion	Mixture	Hu–Wang pitch tracking [16]	Hu–Wang tandem system [15]	Wang–Yu morphological system [20]	Proposed (before morphological image proc.) system	Proposed (after morphological image proc.) system
N_1	4.88	13.28	13.98	14.88	15.78	16.72
N_2	4.29	8.53	8.92	10.08	10.22	11.19
N_3	−0.20	12.78	13.16	13.72	14.28	15.07
N_4	4.95	12.34	13.07	13.46	13.88	14.39
N_5	4.27	9.65	9.73	10.18	11.28	12.01
N_6	−0.85	12.29	12.56	13.37	13.20	14.11
N_7	0.3	6.07	6.22	7.13	7.96	8.88
N_8	3.97	11.23	11.31	12.64	12.45	13.71
N_9	0.33	7.56	7.86	8.98	9.10	10.38
N_{10}	4.25	11.69	12.08	13.45	13.37	14.64
N_{11}	4.04	10.39	10.81	11.62	11.78	13.01
N_{12}	−0.86	10.26	10.89	12.01	12.39	13.59
N_{13}	3.69	9.51	9.79	10.94	11.08	12.28
N_{14}	2.67	7.96	8.34	9.27	9.62	10.75
N_{15}	1.36	7.30	7.63	9.09	8.35	10.53
N_{16}	10.69	14.99	15.37	15.56	15.66	16.39
N_{17}	4.41	10.04	10.26	11.15	11.09	12.99
average	3.07	10.34	10.7	11.62	11.85	12.98

where $\ominus$ denotes the erosion operation and $\boldsymbol{E}(c,m)$ is the resultant binary image matrix after erosion with thinner inner and outer boundary.

2.7.2 Pruning: Pruning is the operation which smoothen the boundaries of the binary image matrix $\boldsymbol{M}_{\mathrm{GL}}(c,m)$. The pruning of the eroded image $\boldsymbol{E}(c,m)$ with the structuring element S_E as follows:

$$\boldsymbol{P}(c,m) = \boldsymbol{E}(c,m) \oplus S_E \tag{16}$$

where $\oplus$ denotes the dilation operation and $\boldsymbol{P}(c,m)$ is the resultant binary image matrix after pruning.

2.7.3 Complementing: The final stage is performing complementing operation which removes the discontinuous gaps in the binary image matrix $\boldsymbol{P}(c,m)$. The complementing operation with the structuring element S_E is defined as follows:

$$\boldsymbol{C}(c,m) = (\boldsymbol{P}(c,m) \oplus S_E) \ominus S_E \tag{17}$$

where $\boldsymbol{C}(c,m)$ is the binary image matrix after complementing which is shown in Fig. 2*c*.

2.8 Speech resynthesis

The final stage in the proposed algorithm is the speech re-synthesis which combines the processed speech signal from each filter channel of Gammatone filterbank. This research work follows the same re-synthesis procedure as proposed by Weintraub [8].

3 Experimental results and discussion

The proposed speech separation system shown in Fig. 1 is evaluated with a set of 170 speech mixtures which consists of 10 voiced utterances collected by Cooke [25] and 17 intrusions at different SNR levels as specified by Yu *et al.* [20]. The intrusions used to create noisy mixture are as follows: 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. The performance of the proposed system in terms of speech quality is measured in terms of SNR improvement [20], percentage of energy loss (P_{EL}) [16], percentage of noise residue (P_{NR}) [16], and PESQ [20]. Similarly, the performance of the proposed system in terms of speech intelligibility is measured in terms of CSII [26], NCM [27, 28], and STOI [27, 28].

The commonly used speech quality measure is SNR improvement which is defined as follows [20]:

$$\mathrm{SNR} = 10\log\left(\frac{\sum_n S(n)^2}{\sum_n (S(n) - S_{\mathrm{out}}(n))^2}\right) \tag{18}$$

where $S(n)$ denotes the original speech and $S_{\mathrm{out}}(n)$ the segregated speech. The proposed system improves the SNR significantly and is shown in Table 1. From Table 1, it is observed that the proposed system improves the SNR with an average value of 9.91, 2.64, and 1.36 dB when compared with input SNR, Hu–Wang pitch tracking system [16] and Yu *et al.* morphological image processing system [20], respectively, for a specific input SNR. From Table 1, it is also observed that, the SNR improvement by the proposed system for noises N_{15}, N_{16}, and N_{17} is low when compared with other noises since these noises include some form of voiced speech (male or female or both). Moreover, the morphological image processing done via pruning and complementing smoothenes the edges of the binary mask produced by the proposed system which shows slight improvement in SNR and it is easily understand from columns 6 and 7 of Table 1. The morphological image processing of course improves the SNR and speech quality, but it does not help for speech separation.

Further to show the superiority of the proposed algorithm, another experiment has been conducted with various SNRs in the range of −5, 0, 10, and 20 dB. The average SNR improvement of the proposed system for speech samples ($V_0 - V_9$) and noise samples ($N_1 - N_{17}$) is shown in Table 2. From Table 2, it is observed that the proposed system improves the average SNR by 9.59, 8.74, and 5.71 dB with respect to the input SNRs of −5, 0, and 10 dB. For closer observations, the performance of the proposed system in terms of average SNR improvement is measured with only five noises N_1, N_3, N_6, N_7, and N_{11} at different SNRs varies from −5, 0, 10, and 20 dB and is shown in Fig. 5*a*. From Table 2 and Fig. 5*a*, it is observed that, when the input SNR exceeds 15 dB, there is no significant improvement in the average SNR and even there is a slight decrement in the average SNR. It is mainly due to the loss of some low energy *T–F* units of voiced and unvoiced speech. The above observation shows that the proposed

Table 2 SNR improvement (IMP) of the proposed system at different SNR levels

Intrusion	I/P SNR, dB							
	−5		0		10		20	
	SNR O/P	SNR IMP	SNR O/P	SNR IMP	SNR O/P	SNR IMP	SNR O/P	SNR IMP
N_1	6.54	11.54	9.37	9.37	16.76	6.76	17.93	−2.07
N_2	2.36	7.36	4.58	4.58	15.71	5.71	16.88	−3.12
N_3	9.36	14.36	15.41	15.41	15.62	5.62	16.69	−3.31
N_4	2.72	7.72	5.84	5.84	14.53	4.53	16.41	−3.59
N_5	2.90	7.9	6.26	6.26	15.73	5.73	16.21	−3.79
N_6	0.11	5.11	14.93	14.93	17.77	7.77	17.19	−2.81
N_7	7.57	12.57	8.61	8.61	15.62	5.62	16.93	−3.07
N_8	4.59	9.59	8.99	8.99	16.95	6.95	17.20	−2.80
N_9	3.60	8.6	9.53	9.53	14.89	4.89	16.21	−3.79
N_{10}	10.06	15.06	11.48	11.48	15.61	5.61	16.34	−3.66
N_{11}	3.02	8.02	6.65	6.65	14.33	4.33	15.58	−4.42
N_{12}	6.90	11.9	13.66	13.66	17.05	7.05	17.75	−2.25
N_{13}	5.27	10.27	8.57	8.57	15.68	5.68	16.60	−3.40
N_{14}	4.53	9.53	6.27	6.27	15.64	5.64	16.49	−3.51
N_{15}	3.74	8.74	7.40	7.40	15.15	5.15	16.31	−3.69
N_{16}	2.75	7.75	4.18	4.18	15.27	5.27	16.82	−3.18
N_{17}	2.11	7.11	6.85	6.85	14.84	4.84	15.92	−4.08
average	4.59	9.59	8.74	8.74	15.71	5.71	16.67	−3.33

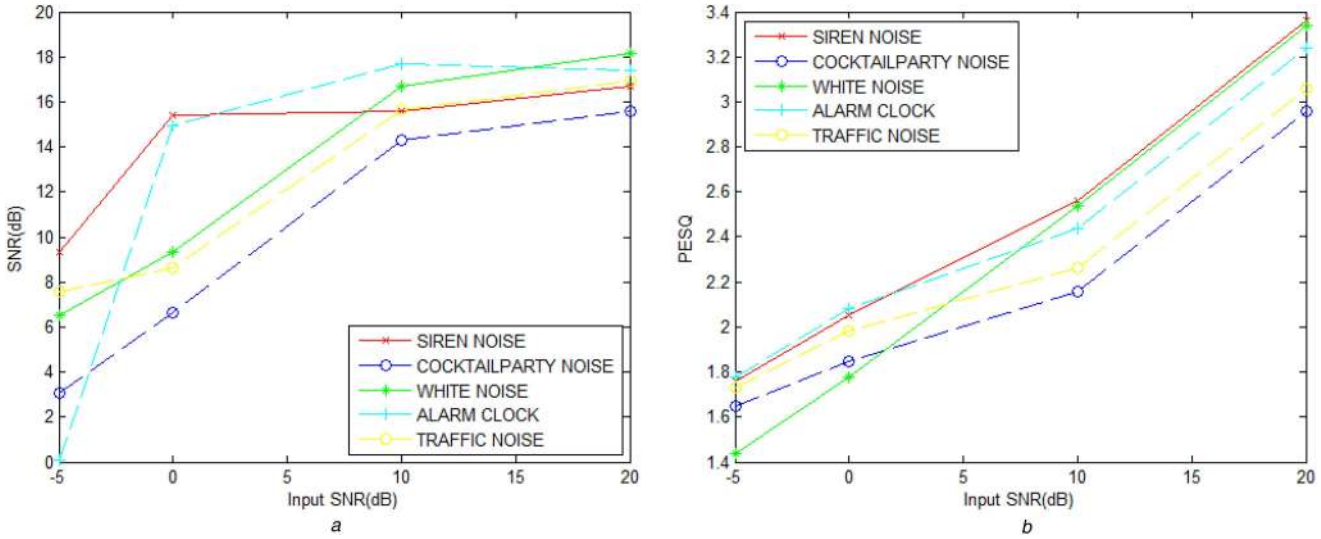

Fig. 5 *Quality measure SNR and PESQ*

(a) SNR improvement of the proposed systems at different SNRs over five various noises,

(b) PESQ (quality measure) of the proposed systems at different SNRs over five various noises

algorithm works well for the low SNR speech signal and failed to retain much of the *T*–*F* units of high SNR target speech signal which in turn degrades the output signal SNR.

In addition to the SNR improvement, two complementary error measures such as P_{EL} and P_{NR} are evaluated. The percentage of energy loss P_{EL} is measured as follows [16]:

$$P_{EL} = \frac{\sum_n e_1^2(n)}{\sum_n I^2(n)} \tag{19}$$

The percentage of noise residue P_{NR} is measured as follows [16]:

$$P_{NR} = \frac{\sum_n e_2^2(n)}{\sum_n O^2(n)} \tag{20}$$

where $I(n)$ is the re-synthesised speech from IBM (ideal binary mask), $O(n)$ the re-synthesised speech from the proposed system, $e_1(n)$ the signal energy in $I(n)$ but missing in $O(n)$, and $e_2(n)$ the signal energy in $O(n)$ but missing in $I(n)$.

The P_{EL} and P_{NR} of the proposed system are computed and shown in Table 3. From which, it is observed that the proposed system has P_{NR} of 2.78% which is less than the Hu–Wang pitch tracking system [16] and Yu *et al.* morphological image processing [20] system. Since noise residue is 2.78%, the proposed system retains 97.22% [(100−P_{NR})%] of the original target signal. Similar to P_{NR}, the proposed system has an average P_{EL} value of 3.24% which is less than the Hu–Wang pitch tracking system [16] and Yu *et al.* morphological image processing [20] system. From Table 3, it is also observed that the proposed system reduces the noise residue almost for all noises except N_{12}. The noise N_{12} is a crowd noise in playground, in which speech signal (shouting by a crowd of people) is present throughout. The proposed algorithm is more effective for segregating the voiced segment of speech and failed to suppress the noise in the unvoiced and silence region of the

Table 3 Energy loss P_{EL} and noise residue P_{NR} of the proposed system

Intrusion	Hu–Wang pitch tracking [16]		Hu–Wang tandem system [15]		Wang–Yu morphological system [20]		Proposed system		Mixture
	P_{EL}, %	P_{NR}, %	P_{EL}, %	P_{NR}, %	P_{EL}, %	P_{NR}, %	P_{EL}, %	P_{NR}, %	P_{NR}, %
N_1	3.81	2.72	3.34	2.46	2.77	1.94	2.51	1.74	27.11
N_2	4.77	6.1	4.76	6.08	5.03	3.95	4.64	3.80	29.5
N_3	6.50	5.78	6.24	5.62	5.92	4.08	5.58	3.90	40.56
N_4	3.39	6.09	3.03	5.82	2.11	4.73	1.97	4.34	22.01
N_5	2.93	1.97	2.39	1.64	1.93	1.01	1.74	0.76	9.77
N_6	5.27	2.3	4.65	2.17	1.7	1.88	1.43	1.49	46.31
N_7	6.59	6.91	6.19	6.63	5.21	6.23	5.03	5.93	36.51
N_8	6.32	3.51	7.07	3.23	5.77	0.69	5.18	0.65	18.43
N_9	7.40	5.83	7.49	5.21	3.95	4.72	3.38	4.35	29.46
N_{10}	4.25	0.84	3.84	0.58	2.06	0.42	1.65	0.31	11.77
N_{11}	6.41	3.00	6.38	3.02	3.22	1.89	3.01	1.69	20.73
N_{12}	4.82	3.93	4.67	3.71	2.01	2.19	1.75	2.42	76.79
N_{13}	4.61	3.45	5.09	3.12	3.14	2.25	2.89	1.77	13.76
N_{14}	6.03	4.98	6.03	4.8	4.96	3.62	4.66	3.18	23.16
N_{15}	4.52	5.52	4.18	5.78	3.83	3.72	3.65	3.32	26.84
N_{16}	3.07	2.91	2.03	2.89	1.84	1.87	0.87	1.72	6.64
N_{17}	6.98	7.36	6.37	6.94	4.06	6.1	4.05	5.94	43.07
average	5.16	4.3	4.93	4.1	3.5	3.02	3.18	2.78	28.38

Table 4 Comparison of PESQ (quality measure) of speech separation systems

Intrusion	Mixture	Hu–Wang pitch tracking [16]	Hu–Wang tandem system [15]	Wang–Yu morphological system [20]	Proposed system
N_1	4.88	2.13	2.34	2.39	2.45
N_2	4.29	1.65	1.68	1.75	1.83
N_3	−0.20	2.15	2.28	2.31	2.40
N_4	4.95	2.05	2.23	2.38	2.52
N_5	4.27	1.62	1.81	1.92	1.98
N_6	−0.85	1.67	1.83	1.93	2.00
N_7	0.3	1.49	1.59	1.66	1.75
N_8	3.97	1.65	1.73	1.75	1.87
N_9	0.33	1.38	1.53	1.56	1.72
N_{10}	4.25	1.46	1.58	1.62	1.78
N_{11}	4.04	1.58	1.67	1.80	1.89
N_{12}	−0.86	1.44	1.61	1.75	1.84
N_{13}	3.69	1.27	1.44	1.49	1.57
N_{14}	2.67	1.45	1.63	1.72	1.87
N_{15}	1.36	1.43	1.45	1.50	1.82
N_{16}	10.69	1.51	1.71	1.71	1.76
N_{17}	4.41	1.61	1.62	1.63	1.89
average	3.07	1.62	1.75	1.82	1.94

crowded noisy speech signal and this shows high P_{NR} value only for N_{12} noise.

The quality of the separated speech can also be evaluated through another measure is called PESQ proposed by the International Telecommunication Union (ITU) under the recommendation P.862 [29]. The PESQ value of the proposed system is shown in Table 4. From which it is observed that the proposed system improves the PESQ with an average value of 0.32, 0.19, and 0.12 when compared with Hu–Wang pitch tracking system [16], Hu–Wang Tandem algorithm [15], and Yu *et al.* morphological image processing system [20], respectively, for a specific input SNR. To show the effectiveness of the proposed system in terms of PESQ, another experiment has been conducted with various SNRs in the range of −5, 0, 10, and 20 dB. The average PESQ improvement for all speech samples ($V_0 - V_9$) and noise samples ($N_1 - N_{17}$) is shown in Table 5. It is observed that the proposed system improves the PESQ value by an average of 0.71, 0.62, 0.40, and 0.28 with respect to the noisy speech at −5, 0, 10, and 20 dB SNRs, respectively. Again for closer observations, the performance of the proposed system in terms of PESQ value is measured with only five noises N_1, N_3, N_6, N_7, and N_{11} at different SNRs varies from −5, 0, 10, and 20 dB and is shown in Fig. 5*b*. From Table 5 and Fig. 5*b*, it is observed that, when the input SNRs exceed 15 dB, there is only a slight improvement in the PESQ with respect to noisy speech. Similar observation has also been made with respect to SNR improvement >15 dB in Table 2, but the only difference is that the PESQ value does not decrease but shows only a slight improvement when compared with the improvement with respect to other SNRs.

The speech intelligibility of the proposed system is measured in terms of CSII, NCM, and STOI. The CSII, NCM, and STOI values for all noises and its average value at different SNRs is shown in Table 6 and its graphical representation with only average values is shown in Fig. 6. The speech intelligibility improvement in terms of

Table 5 PESQ quality measure of noisy speech (NS), separated speech (SS), and improvement (IMP) by the proposed system at different SNR levels

Intrusion	I/P SNR, dB											
	−5			0			10			20		
	PESQ			PESQ			PESQ			PESQ		
	NS	SS	IMP	NS	SS	IMP	NS	SS	IMP	NS	SS	IMP
N_1	0.85	1.44	0.59	0.88	1.78	0.9	2.18	2.54	0.36	3.28	3.34	0.06
N_2	0.87	1.33	0.46	1.05	1.53	0.48	1.65	2.09	0.44	2.21	2.89	0.68
N_3	1.04	1.76	0.72	1.17	2.05	0.88	2.1	2.56	0.46	3.12	3.36	0.24
N_4	1.01	1.63	0.62	1.38	1.84	0.46	2.06	2.42	0.36	3.16	3.22	0.06
N_5	1.02	1.51	0.49	1.49	1.98	0.49	2.18	2.54	0.36	3.08	3.34	0.26
N_6	1.13	1.78	0.65	1.29	2.08	0.79	2.08	2.44	0.36	3.01	3.24	0.24
N_7	0.91	1.73	0.82	1.03	1.98	0.95	1.98	2.26	0.28	2.78	3.06	0.28
N_8	0.81	1.63	0.82	1.04	1.74	0.7	2.01	2.47	0.46	2.82	3.27	0.45
N_9	0.92	1.55	0.63	1.11	1.77	0.66	2.19	2.55	0.36	3.30	3.35	0.05
N_{10}	0.43	1.56	1.13	1.19	1.66	0.47	1.82	2.05	0.23	2.26	2.85	0.59
N_{11}	0.81	1.65	0.84	1.24	1.85	0.61	1.89	2.16	0.27	2.58	2.96	0.38
N_{12}	0.63	1.35	0.72	1.09	1.64	0.55	2.09	2.45	0.36	3.19	3.25	0.06
N_{13}	0.78	1.54	0.76	1.07	1.75	0.68	1.73	2.15	0.42	2.54	2.95	0.41
N_{14}	0.79	1.64	0.85	1.15	1.75	0.6	1.82	2.48	0.66	3.04	3.28	0.24
N_{15}	0.71	1.53	0.82	1.29	1.73	0.44	1.79	2.15	0.36	2.63	2.95	0.32
N_{16}	0.72	1.33	0.61	1.20	1.54	0.34	1.27	1.70	0.43	2.21	2.50	0.29
N_{17}	0.89	1.43	0.54	1.01	1.64	0.63	1.41	2.09	0.68	2.72	2.89	0.17
average	0.84	1.55	0.71	1.16	1.78	0.62	2.18	2.30	0.4	3.28	3.10	0.28

Table 6a CSII, NCM, and STOI intelligibility measures of noisy speech (NS) and separated speech (SS) by the proposed system

Intrusion	Type	I/P SNR (dB)											
		−5			0			10			20		
		CSII	NCM	STOI	CSII	NCM	STOI	CSII	NCM	STOI	CSII	NCM	STOI
N_1	NS	0.23	0.19	0.43	0.29	0.32	0.59	0.33	0.36	0.73	0.41	0.48	0.82
	SS	0.75	0.54	0.72	0.88	0.85	0.81	0.91	0.89	0.94	0.95	0.94	1.00
N_2	NS	0.27	0.16	0.42	0.30	0.20	0.51	0.33	0.25	0.68	0.40	0.40	0.81
	SS	0.67	0.53	0.66	0.83	0.69	0.85	0.85	0.80	0.91	0.97	0.92	0.99
N_3	NS	0.27	0.17	0.67	0.31	0.19	0.69	0.41	0.28	0.81	0.44	0.42	0.76
	SS	0.78	0.61	0.78	0.93	0.88	0.90	0.90	0.93	0.95	0.95	0.83	0.98
N_4	NS	0.28	0.20	0.66	0.32	0.25	0.70	0.38	0.34	0.75	0.47	0.54	0.81
	SS	0.69	0.73	0.80	0.85	0.94	0.84	0.87	0.90	0.99	0.97	0.94	0.99
N_5	NS	0.27	0.30	0.33	0.31	0.36	0.58	0.35	0.39	0.79	0.43	0.55	0.83
	SS	0.79	0.58	0.79	0.94	0.92	0.90	0.93	0.96	0.93	0.98	0.96	0.99
N_6	NS	0.30	0.13	0.38	0.32	0.18	0.47	0.34	0.25	0.77	0.43	0.48	0.84
	SS	0.48	0.53	0.85	0.67	0.67	0.92	0.85	0.91	0.96	0.98	0.96	0.98
N_7	NS	0.34	0.31	0.67	0.36	0.42	0.76	0.41	0.43	0.82	0.50	0.59	0.84
	SS	0.79	0.79	0.80	0.94	0.80	0.90	0.93	0.96	0.99	0.98	0.96	1.00
N_8	NS	0.27	0.13	0.45	0.29	0.18	0.53	0.31	0.23	0.71	0.39	0.40	0.81
	SS	0.75	0.55	0.65	0.87	0.71	0.81	0.94	0.85	0.85	0.97	0.93	0.96
N_9	NS	0.26	0.07	0.37	0.28	0.11	0.51	0.29	0.18	0.76	0.35	0.39	0.83
	SS	0.69	0.55	0.65	0.85	0.78	0.77	0.91	0.90	0.97	0.98	0.95	0.98

CSII, NCM, and STOI is shown in Table 7. From Table 7, it is observed that the proposed system improves the CSII by an average value of 0.429, 0.541, 0.549, and 0.550 with respect to the noisy speech at SNRs of −5, 0, 10, and 20 dB, respectively. The NCM by an average value of 0.436, 0.555, 0.596, and 0.476 with respect to the noisy speech at SNRs of −5, 0, 10, and 20 dB and STOI by an average value of 0.266, 0.274, 0.211, and 0.169 with respect to the noisy speech at SNRs of −5, 0, 10, and 20 dB, respectively. It is also observed that, there is no significant improvement of CSII, NCM, and STOI values when the SNR exceeds 15 dB. In fact, the improvement of NCM and STOI value decreases when the input SNR exceeds 15 dB, this shows that the proposed system improves the speech intelligibility at low SNRs when compared with high SNRs, i.e. at >15 dB.

4 Conclusion and future work

This research work proposed an image processing-based algorithm to enhance the binary *T–F* mask obtained from the initial segmentation of a CASA system. The proposed algorithm effectively identifies and eliminates the holes which are wrongly interpreted as noise in the speech-dominant *T–F* units group. Also eliminates some noisy *T–F* units which are wrongly interpreted as speech and in turn improves the speech quality and intelligibility. The performance of the proposed system is evaluated in terms of

Table 6b

Intrusion	Type	I/P SNR (dB)											
		−5			0			10			20		
		CSII	NCM	STOI	CSII	NCM	STOI	CSII	NCM	STOI	CSII	NCM	STOI
N_{10}	NS	0.37	0.29	0.78	0.39	0.33	0.79	0.41	0.35	0.82	0.48	0.55	0.83
	SS	0.77	0.86	0.95	0.93	0.85	0.95	0.93	0.95	0.97	0.97	0.95	0.98
N_{11}	NS	0.27	0.07	0.26	0.29	0.11	0.40	0.30	0.17	0.75	0.37	0.40	0.84
	SS	0.79	0.61	0.56	0.74	0.74	0.70	0.88	0.90	0.93	0.97	0.96	0.97
N_{12}	NS	0.27	0.19	0.52	0.31	0.23	0.57	0.34	0.26	0.70	0.41	0.43	0.81
	SS	0.57	0.53	0.80	0.76	0.77	0.85	0.84	0.83	0.94	0.97	0.94	1.00
N_{13}	NS	0.27	0.11	0.43	0.30	0.16	0.51	0.31	0.21	0.70	0.38	0.37	0.82
	SS	0.70	0.54	0.70	0.84	0.70	0.77	0.87	0.83	0.92	0.97	0.95	1.00
N_{14}	NS	0.29	0.16	0.45	0.32	0.18	0.55	0.35	0.22	0.70	0.42	0.42	0.81
	SS	0.74	0.55	0.72	0.86	0.76	0.79	0.95	0.84	0.91	0.97	0.95	0.99
N_{15}	NS	0.27	0.06	0.37	0.29	0.12	0.48	0.32	0.19	0.68	0.38	0.38	0.81
	SS	0.69	0.50	0.62	0.86	0.66	0.80	0.87	0.81	0.89	0.96	0.93	0.98
N_{16}	NS	0.32	0.17	0.31	0.36	0.23	0.50	0.40	0.32	0.70	0.47	0.53	0.82
	SS	0.78	0.61	0.65	0.92	0.76	0.85	0.92	0.79	0.92	0.98	0.93	0.99
N_{17}	NS	0.28	0.16	0.37	0.31	0.19	0.50	0.34	0.24	0.69	0.42	0.47	0.81
	SS	0.68	0.68	0.68	0.88	0.74	0.87	0.90	0.76	0.94	0.96	0.92	1.00
average	NS	0.28	0.17	0.46	0.31	0.22	0.57	0.35	0.28	0.74	0.42	0.46	0.82
	SS	0.71	0.60	0.73	0.86	0.78	0.84	0.90	0.87	0.94	0.97	0.94	0.99

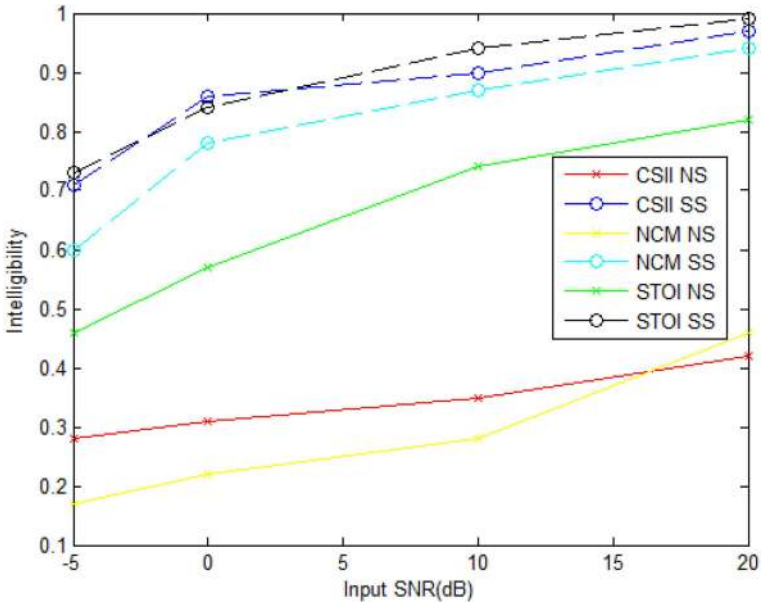

Fig. 6 *Average values of CSII, NCM, and STOI measures of the proposed systems at different SNR levels*

SNR, P_{EL}, P_{NR}, and PESQ for speech quality and CSII, NCM, and STOI for speech intelligibility. The experimental results show that the proposed system improves the speech quality and intelligibility when compared with the other existing systems. However, the proposed system uses many parameters, for example, $\theta_H, \theta_c, \theta_T, \theta_{AT}(c)$, and θ_{AP}, to effectively classify the *T–F* units into speech-dominant and noise-dominant. The value for these parameters is obtained either from the corresponding references or by conducting experiments with different speech and noise samples. The change in the value of these parameters may slightly affect the performance of the system. Also observed that, when the input SNR exceeds 15 dB, there is no significant improvement in the output SNR and even there is a slight reduction in SNR. It is mainly due to the loss of some low energy *T–F* units of voiced and unvoiced speech. The above observation shows that the proposed algorithm works well for the low SNR speech signals and failed to retain much of the *T–F* units of high SNR target speech signal which in turn degrades the output signal SNR. The future work of this research will concentrate on this issue and develop a system to show SNR improvement in all SNRs range for both voiced and unvoiced speech utterances.

Table 7 CSII, NCM, and STOI improvement of the proposed system at various SNR levels

Intrusion	I/P SNR, dB											
	−5			0			10			20		
	CSII	NCM	STOI	CSII	NCM	STOI	CSII	NCM	STOI	CSII	NCM	STOI
N_1	0.527	0.357	0.286	0.597	0.528	0.219	0.580	0.521	0.214	0.546	0.453	0.173
N_2	0.401	0.371	0.240	0.527	0.487	0.338	0.523	0.547	0.232	0.577	0.521	0.182
N_3	0.512	0.440	0.117	0.620	0.684	0.207	0.492	0.646	0.139	0.515	0.406	0.221
N_4	0.416	0.531	0.146	0.529	0.687	0.147	0.486	0.559	0.234	0.491	0.391	0.179
N_5	0.522	0.284	0.452	0.631	0.566	0.323	0.582	0.572	0.141	0.548	0.402	0.155
N_6	0.177	0.397	0.468	0.351	0.483	0.451	0.508	0.661	0.187	0.552	0.481	0.140
N_7	0.453	0.479	0.131	0.573	0.386	0.146	0.521	0.532	0.162	0.475	0.374	0.157
N_8	0.481	0.420	0.204	0.582	0.531	0.279	0.630	0.618	0.146	0.577	0.531	0.149
N_9	0.426	0.479	0.283	0.572	0.663	0.268	0.622	0.719	0.206	0.629	0.562	0.142
N_{10}	0.407	0.571	0.169	0.536	0.520	0.157	0.514	0.599	0.149	0.491	0.397	0.151
N_{11}	0.522	0.541	0.302	0.449	0.630	0.298	0.587	0.727	0.181	0.607	0.555	0.132
N_{12}	0.303	0.341	0.283	0.456	0.541	0.273	0.495	0.574	0.238	0.557	0.512	0.189
N_{13}	0.432	0.428	0.270	0.546	0.541	0.262	0.556	0.628	0.221	0.597	0.578	0.179
N_{14}	0.449	0.390	0.268	0.545	0.573	0.237	0.604	0.625	0.219	0.551	0.522	0.174
N_{15}	0.420	0.434	0.250	0.564	0.541	0.325	0.557	0.615	0.207	0.580	0.557	0.169
N_{16}	0.457	0.440	0.339	0.555	0.527	0.359	0.523	0.469	0.222	0.511	0.401	0.174
N_{17}	0.396	0.516	0.313	0.569	0.548	0.376	0.553	0.520	0.244	0.543	0.456	0.185
average	0.429	0.436	0.266	0.541	0.555	0.275	0.549	0.596	0.197	0.550	0.476	0.168

5 References

[1] Jensen, J., Hansen, J.H.L.: 'Speech enhancement using a constrained iterative sinusoidal model', *IEEE Trans. Speech Audio Process.*, 2001, **9**, (7), pp. 731–740

[2] Zhang, X., Wang, Z., Wang, D.: 'A speech enhancement algorithm by iterating single- and multi-microphone processing and its application to robust ASR'. Proc. ICASSP, New Orleans, USA, March 2017, pp. 276–280

[3] Hyvarinen, A., Karhunen, J., Oja, E.: '*Independent component analysis*' (Wiley Press, New York, 2001)

[4] Ephraim, Y., Trees, H.L.: 'A signal subspace approach for speech enhancement', *IEEE Trans. Speech Audio Process.*, 1995, **3**, (4), pp. 251–266

[5] Sameti, H., Sheikhzadeh, H., Deng, L., *et al.*: 'HMM-based strategies for enhancement of speech signals embedded in non-stationary noise', *IEEE Trans. Speech Audio Process.*, 1998, **6**, (5), pp. 445–455

[6] Wang, D.L., Kun, H.: 'Towards generalizing classification based speech separation', *IEEE Trans. Audio, Speech Lang. Process.*, 2013, **21**, (1), pp. 68–177

[7] Hu, G., Wang, D.: 'An auditory scene analysis approach to monaural speech segregation', in Hansler, E., Schmidt, G. (EDS.): '*Topics in acoustic echo and noise control*' (Springer Press, Heidelberg, 2006), pp. 485–515

[8] Weintraub, M.: 'A theory and computational model of auditory monaural sound separation'. PhD dissertation, Stanford University, Stanford, CA, 1985

[9] Rabiee, A., Setayeshi, S., Lee, S.Y.: 'CASA: biologically inspired approaches for auditory scene analysis', *Natural Intelligence*, 2012, **1**, (2), pp. 50–58

[10] Brown, G.J., Cooke, M.P.: 'Computational auditory scene analysis', *Comput. Speech Lang.*, 1994, **8**, (4), pp. 297–336

[11] Harish, N., Rajavel, R.: 'Monaural speech separation system based on optimum soft mask'. Proc. IEEE Int. Conf. Computational Intelligence and Computing Research, Coimbatore, India, December 2014, pp. 1–5

[12] Rabiee, A., Setayeshi, S., Lee, S.Y.: 'A harmonic-based biologically inspired approach to monaural speech separation', *IEEE Signal Process. Lett.*, 2012, **19**, (9), pp. 559–562

[13] Donald, S., Wang, D.: 'Time-Frequency masking in the complex domain for speech dereverberation and denoising', *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2017, **25**, (7), pp. 1492–1501

[14] Brown, G.J., Wang, D.L.: 'Separation of speech by computational auditory scene analysis', in Benesty, J., Makino, S., Chen, J. (EDS.): '*Speech enhancement*' (Springer, New York, 2005), pp. 371–402

[15] Wang, D.: 'Tandem algorithm for pitch estimation and voiced speech segregation', *IEEE Trans. Audio, Speech, Lang. Process.*, 2012, **18**, (8), pp. 2067–2079

[16] Hu, G., Wang, D.: 'Monaural speech segregation based on pitch tracking and amplitude modulation', *IEEE Trans. Neural Netw.*, 2004, **15**, (5), pp. 1135–1150

[17] Hu, G., Wang, D.: 'Auditory segmentation based on onset and offset analysis', *IEEE Trans. Audio Speech Lang. Process.*, 2007, **15**, (2), pp. 396–405

[18] Boll, S.F.: 'Suppression of acoustic noise in speech using spectral subtraction', *IEEE Trans. Acoust. Speech Signal Process.*, 1979, **27**, pp. 113–120

[19] Hu, K., Wang, D.: 'Unvoiced speech segregation from non-speech interference via CASA and spectral subtraction', *IEEE Trans. Audio Speech Lang. Process.*, 2011, **19**, (6), pp. 1600–1609

[20] Yu, W., Jiajun, L., Ning, C., *et al.*: 'Improved monaural speech segregation based on computational auditory scene analysis', *J. Audio Speech Music Process.*, 2013, **2**, doi: 10.1186/1687-4722-2013-2, pp. 1–15

[21] Patterson, R.D., Nimmo-Smith, I., Holdsworth, J., *et al.*: 'An efficient auditory filterbank based on the gammatone function', MRC Applied Psychology Unit, 1988

[22] Meddis, R.: 'Simulation of auditory-neural transduction: further studies', *J. Acoust. Soc. Am.*, 1988, **83**, (3), pp. 1056–1063

[23] Wang, D.L., Brown, G.J.: 'Separation of speech from interfering sounds based on oscillatory correlation', *IEEE Trans. Neural Netw.*, 1999, **10**, (3), pp. 684–697

[24] Shoba, S., Rajavel, R.: 'Adaptive energy threshold selection for monaural speech separation'. Proc. IEEE Int. Conf. Communication and Signal Processing, Melmaruvathur, India, April 2017

[25] Cooke, M.P.: 'Modeling auditory processing and organization'. PhD dissertation, University of Sheffield, Sheffield, UK, 1993

[26] Kates, J.M., Arehart, K.H.: 'Coherence and the speech intelligibility index', *J. Acoust. Soc. Am.*, 2005, **17**, (4), pp. 2224–2237

[27] Taal, C.H., Hendriks, R.C., Heusdens, R., *et al.*: 'An algorithm for intelligibility prediction of time frequency weighted noisy speech', *IEEE Trans. Audio Speech Lang. Process.*, 2011, **19**, (7), pp. 2125–2136

[28] Ma, J., Hu, Y., Loizou, P.: 'Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions', *J. Acoust. Soc. Am.*, 2009, **125**, (5), pp. 3387–3405

[29] Pichevar, R., Rouat, J.: '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*' (Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2005), pp. 430–435