Journal of King Saud University – Computer and Information Sciences (2015) 27, 269–277

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

Cuckoo search based optimal mask generation for noise suppression and enhancement of speech signal

Anil Garg *, **O.P. Sahu**

Department of ECE, NIT Kurukshetra, India

Received 25 April 2013; revised 7 March 2014; accepted 3 April 2014
Available online 18 June 2015

KEYWORDS

Noise suppression;
Enhancement of speech signal;
AMS feature extraction;
Cuckoo search;
Waveform synthesis;
Optimal mask

Abstract In this paper, an effective noise suppression technique for enhancement of speech signals using optimized mask is proposed. Initially, the noisy speech signal is broken down into various time–frequency (TF) units and the features are extracted by finding out the Amplitude Magnitude Spectrogram (AMS). The signals are then classified based on quality ratio into different classes to generate the initial set of solutions. Subsequently, the optimal mask for each class is generated based on Cuckoo search algorithm. Subsequently, in the waveform synthesis stage, filtered waveforms are windowed and then multiplied by the optimal mask value and summed up to get the enhanced target signal. The experimentation of the proposed technique was carried out using various datasets and the performance is compared with the previous techniques using SNR. The results obtained proved the effectiveness of the proposed technique and its ability to suppress noise and enhance the speech signal.

1. Introduction

The problem of speech enhancement has received a significant amount of research attention over the past several decades (Hu and Loizou, 2007). Particularly, it focuses on improving the performance of speech communication system in noisy environments such as traffic and crowd (Hong et al., 2009). Many speech enhancement algorithms such as spectral subtraction, subspace, statistical-model based and wiener type have been reported (Hu and Loizou, 2007; Kim and Loizou, 2011). Spectral subtraction is based on principle of obtaining the estimate of clean speech signal by subtracting the average of noise spectrum from noisy speech spectrum (Boll, 1979). The noise spectrum is estimated initially in the absence of speech signal (Boll, 1979). The performance of the speech enhancement algorithms is usually measured in terms of intelligibility and signal-to-noise ratio (SNR) (Kim and Loizou, 2011; Chirstiansen et al., 2010; Ma et al., 2010). Several researchers and professionals have developed various algorithms for estimating and improving intelligibility and SNR (Hu and Loizou, 2007; Chirstiansen et al., 2010). In many speech enhancement and noise reduction algorithms, the decision is based on the apriori SNR (Loizou, 2006), and the classic algorithms like spectral subtraction, Wiener filtering, and maximum likelihood, can be formulated as a function of this

* Corresponding author.
E-mail addresses: anilgarg0778@gmail.com, agarg001@yahoo.com (A. Garg).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2014.04.006

a priori SNR (Scalart and Filho, 1996). In real-time applications, the apriori SNR estimation is useful, but in the ideal situation the local SNR is preferable instead of the apriori SNR (Wolfe and Godsill, 2003). For example, Ephraim and Malah used the decision directed approach for signal-to noise ratio estimation by using the weighted average of the past SNR estimate and the present SNR estimate (Ephraim and Malah, 1984; Chen and Loizou, 2011). The posteriori and a priori SNRs are main function for computing gain function using modified decision-directed approach (Ephraim and Malah, 1984). The gain function used in ideal binary mask for computational auditory scene analysis is identical to the gain function of the Maximum a posterior (MAP) estimators Lu and Loizou (2011). Another significant research was presented by Kim et al. (2009) and Kim and Loizou (2010), where the input signals were broken down into time–frequency units and the features were extracted by the AMS feature extraction technique. In this approach, binary decisions (weight value zero or one) were taken based on the Bayesian classifier, as to whether each T–F unit is dominated by the target or the masker. These speech enhancement algorithms/approaches have been reported to estimate the original speech, degraded by various types of noises (Lu and Loizou, 2011; Kim et al., 2009; Kim and Loizou, 2010; Muhammad, 2010). However, the degree of improvement, measured in terms of intelligibility and SNR, is not easy (Kim and Loizou, 2011; Chirstiansen et al., 2010; Ma et al., 2010). This is primarily due to lack of good estimation of the noise spectrum, especially when it is non stationary (Kim and Loizou, 2011). However, a high signal–to-noise ratio is always desirable to increase speech intelligibility (Kim and Loizou, 2011; Chirstiansen et al., 2010; Ma et al., 2010). In recent studies, the binary mask (Kim and Loizou, 2010) retains the time–frequency (T–F) regions where the target speech dominates the masker (noise) (e.g., local SNR $>$ 0 dB) and removes T–F units where the masker dominates (e.g., local SNR $<$ 0 dB) (Kim and Loizou, 2010). Although, speech produced in the presence of noise called "Lombard speech" has been found to be easily understandable than speech produced during silence (Lu and Cooke, 2009). In previous studies, large gain in intelligibility can be obtained by multiplying the noisy signal with the ideal binary mask signal, even at extremely low (5, 10 dB) SNR levels (Brungart et al., 2006; Li and Loizou, 2008). Kim et al. (2009) and Kim and Loizou (2010) presented the generation of binary mask with the help of Bayesian classifier technique that is lazy classification technique. Since the classification with the lazy classifier, the generation of binary mask will not be an optimal one. If the binary mask is not an optimal one, it will affect the performance of the speech enhancement. This paper presents optimal mask generation using cuckoo search algorithm (Yang, 2009) which is a kind of optimization algorithm (Mandal, 2012; Venkata Rao and Waghmare, 2014) for speech enhancement to improve the SNR and thus intelligibility. The proposed algorithm optimizes the masking parameters in order to suppress the noise effectively for enhancement of speech signal. Comparison and simulation results of our proposed method are better in terms of SNR than the Bayesian classifier technique.

The rest of the paper is organized as follows: A brief description of Cuckoo search algorithm is given in Section 2. The cuckoo search based optimal mask generation is explained in Section 3. The simulation results and discussions are presented in Section 4. The paper is concluded in Section 5.

2. Cuckoo search algorithms

Cuckoo search (CS) Yang, 2009; Valian et al., 2011 is one of the latest optimization algorithms and was developed from the inspiration that the obligate brood parasitism of some cuckoo species lay their eggs in the nests of other host birds which are of other species. In Cuckoo Search, three idealized rules are considered which say that each cuckoo lays one egg at a time, and dumps its egg in a randomly chosen nest. The second rule states that best nests with high quality of eggs will carry over to the next generations and the third one says that the number of available host nests is fixed, and the egg laid by a cuckoo is discovered by the host bird with a probability in the range 0–1. In this case, the host bird can either throw the egg away or abandon the nest, and build a completely new nest. It is also assumed that a definite fraction of the nests are replaced by new nests. For a maximization problem, the quality or fitness of a solution can simply be proportional to the value of the objective function. The algorithm is based on the obligate brood parasitic behavior of some cuckoo species in combination with the Levy flight behavior of some birds and fruit flies.

In the algorithm, updation is carried out using Levy flight and comparison is made with the use of fitness functions and suitable substitutions are made. Levi flight is carried out on ym_i to yield to get a new cuckoo ym_i^* which is given by: $ym_{i1}^* = ym_{i1}^{(t+1)} = ym_{i1}^{(t)} + \Delta \otimes Levy(y)$, where the levy sharing is specified by: $Levy(y) = \sqrt{\frac{c}{2\pi}} \cdot \frac{e^{-\frac{1}{2y}}}{y^{3/2}}$, where c is arbitrary constant. Consequently, some other nest is observed and its fitness function is found out. If the fitness of the Levy flight made nest is superior to the fitness of the nest in consideration, then substitute nest signal values by the host nest Levy performed values. For each iteration, a portion of the utmost horrifying nests are done away with and fresh nests are constructed as replacement.

Based on the above mentioned rules, the basic steps of the Cuckoo search can be summarized as the pseudo code as follows (Yang, 2009; Valian et al., 2011):

Pseudo code:
Objective Function: Maximize the SNR ratio and to obtain the optimal mask weight for each class
Start
For every class Cl_i for $0 < I \leqslant 3$ perform:
The initial population of the class cl_i in consideration is $G_i = \{g_{i1}, g_{i2}. \ldots . \ldots . \ldots . . gi_{Nci}\}$
Generate 25 host nests $H = \{h_1, h_2. \ldots . \ldots . \ldots . . h_{25}\}$ and consider the signals $Y_i = \{y_{i1}, y_{i2}. \ldots . \ldots . . yi_{Nh}\}$ in the i^{th} host nest for $0 < i \leqslant 25$
While (stop criteria)
Perform the levy flight $y_{i1}^* = y_{i1}^{(t+1)} = y_{i1}^{(t)} + \Lambda \otimes Levy(x)$ for all signals in the i^{th} host nest
Find the fitness of the new solution F_i where fitness is the SNR ratio
Choose another random nest j and find the fitness value F_j
If ($F_i > F_j$)

Replace the nest j with the new solution of nest i
End
Fraction of worst nests F_{ra} are abandoned and new ones are built
Best solutions are kept which are ranked and current best is taken
End while
The SNR ratio of the best solution is taken as mask for the class
End

3. Cuckoo searches based optimal mask generation

The approach used in this paper for noise suppression and speech enhancement technique consists of three major modules namely; Feature extraction module (Kim et al., 2009), optimal mask generation module and waveform synthesis module. Initially, the original and noise speech signal is given as input to extract features and subsequently, optimal mask is generated with the use of cuckoo search. Subsequently, in the waveform synthesis module, filtered waveforms are windowed and then multiplied by the optimal mask value and summed up to get the enhanced signal. The block diagram of the proposed technique is given in Fig. 1.

3.1. Feature extraction module

In this module, features are extracted from the input speech corpus with the aid of the Amplitude Magnitude Spectrogram (AMS) Kim et al., 2009. The input speech signal will be a mixture of clean speech signal and the noisy signal. The input signal is initially processed by performing sampling, quantization and then, pre-emphasized to make the signal fit for further processing. Block diagram of the AMS feature extraction is given in Fig. 2.

The processed signals are then decomposed into various TF (Time–Frequency) units with the use of the band pass filters. In this module (Kim et al., 2009), we split the signals into 25 TF units; each contributing to a channel which is represented by C_i, *where* $1 \leqslant i \leqslant 25$. Band-pass filter has the characteristics of passing the signals within the prescribed range of frequencies while attenuating other signals. Therefore in all of the 25 band channels in consideration, each will have signals lying in the range of frequencies defined for the respective channel. Here, every channel is defined by the upper limit frequency U_i and the lower limit frequency L_i. After forming the channel bands, envelope of each band is calculated by the full wave rectification and subsequently, the envelope is decimated by a factor of 3 which is later segmented into overlapping segments of 128 samples of 32 ms with an overlap of 64 samples (Lu and Loizou, 2011). Let each of the segments be represented by S_{ij}, *where* $1 \leqslant i \leqslant 25$, $1 \leqslant j \leqslant N_i$ and N_i is the number of segments formed by the i^{th} channel. The sampled signals obtained after the segmentation are Hanning windowed (Salivahanan, 2010) in order to remove unwanted signal components and get sharper peaks. The windowed signals are initially zero-padded and taken Fourier transformed (256 point FFT) to obtain the modulation spectrum of each channel having frequency resolution of 15.6 Hz (Kim et al., 2009).

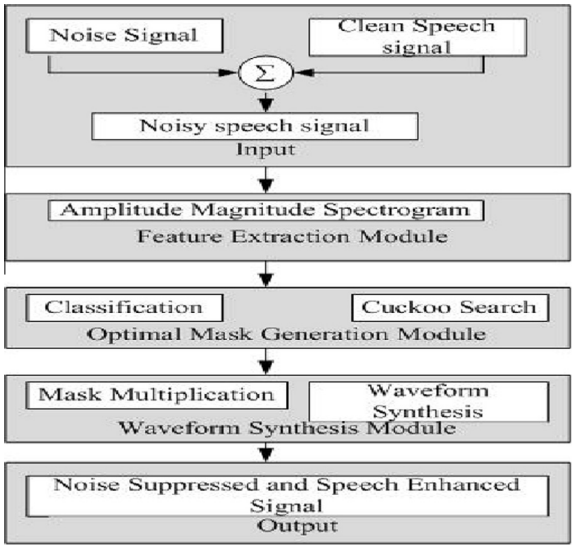

Figure 1 Block diagram of the proposed technique.

Hence, the modulation spectrum for all the 25 channels is obtained by the use of FFT and subsequently, every channel is then multiplied by fifteen triangular-shaped windows spaced uniformly across the 15.6–400 Hz range (Kim et al., 2009). All these are summed up to produce 15 modulation spectrum amplitudes and each of this represents the AMS feature vector (Kim et al., 2009). Use of AMS results in having better extraction of features form the noisy speech signal when compared to other conventional feature extraction techniques. This is due to the combined effort of segment separation, windowing, FFT and multiplication with triangular function. Let the feature vector is represented by $A_F(\lambda, \phi)$ where ϕ represents the time slot and λ represents the sub-band (Kim et al., 2009). Considering the small changes that may occur in the time and the frequency domains, we also take in the delta functions to the features extracted. The time delta functions ΔA_T as given below (Kim et al., 2009):

$$\Delta A_T(\lambda, \phi) = A_F(\lambda, \phi) - A_F(\lambda, \phi - 1), \text{ where } \phi = 2, ..., T \quad (1)$$

The frequency delta function ΔA_S is as given below:

$$\Delta A_s(\lambda, \phi) = A_F(\lambda, \phi) - A_F(\lambda - 1, \phi) \text{ where } \lambda = 2, ..., B \quad (2)$$

The overall feature vector $A(\lambda, \phi)$ including the delta functions can be defined as:

$$A(\lambda, \phi) = [A_F(\lambda, \phi), \Delta A_T(\lambda, \phi), \Delta A_S(\lambda, \phi)] \quad (3)$$

Hence, we have extracted the features from a large speech signal corpus using AMS feature extraction (Kim et al., 2009).

3.2. Optimal weight generation module

In this module, each of the individual TF units is classified into various classes by comparing with the original signal and later an optimal mask is found by the use of cuckoo search (Yang, 2009; Valian et al., 2011).

(a) Classification:

Here, the input TF unit is classified into the respective class with the use of original signal and noisy signal. The classification of the speech signal to different classes is based on the Quality Ratio which is the ratio of the estimated speech magnitude $\bar{M}$ to the true speech magnitude T for each T–F unit.

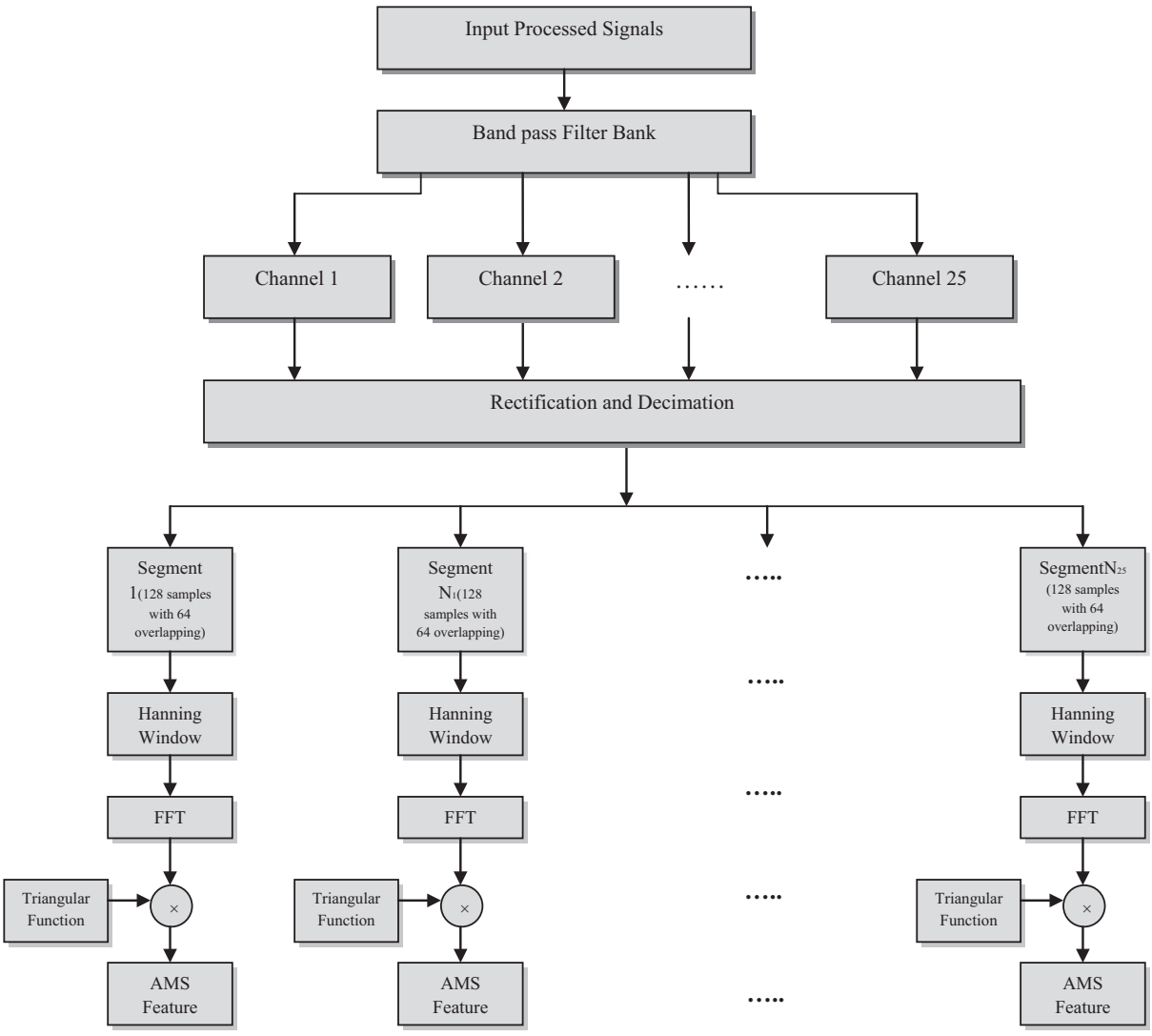

Figure 2 Block diagram of AMS feature extraction.

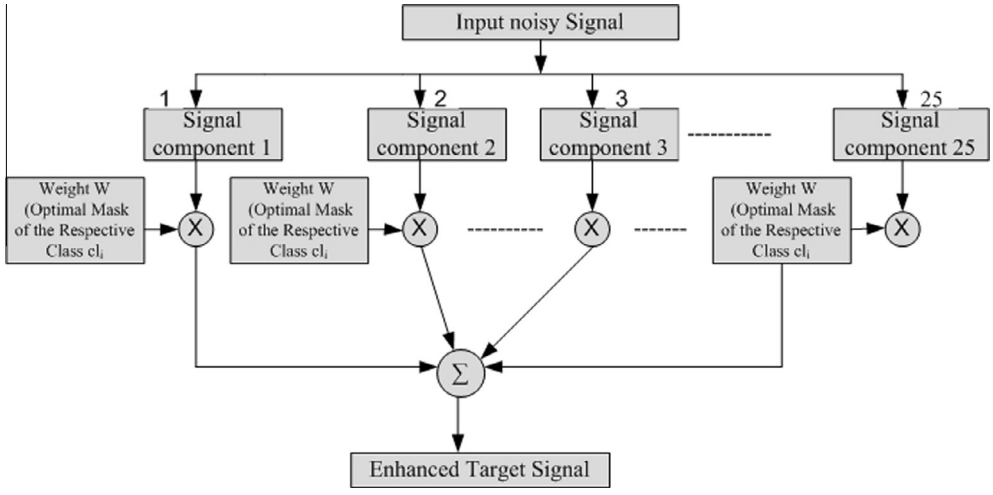

Figure 3 Block diagram of the waveform synthesis module.

Here the spectrum at time slot ϕ and sub-band λ is considered; hence the quality ratio R_Q can be defined by:

$$R_Q = \frac{|\overline{M}(\lambda,\phi)|}{|T(\lambda,\phi)|} \tag{4}$$

where estimated signal spectrum $\bar{M}$ is obtained by the product of spectrum M with the gain function G_A which is shown in the equation below:

$$\overline{M}(\lambda,\phi) = G(\lambda,\phi).|M(\lambda,\phi)| \tag{5}$$

where Gain can be found out from the Eq. (3):

$$G_A(\lambda,\phi) = \sqrt{\frac{\psi(\lambda,\phi)}{1+\psi(\lambda,\phi)}} \tag{6}$$

where ψ is the priori signal to noise ratio given by the equation ($\eta = 0.98$ is a smoothing constant and ε_N is the estimate of the background noise variance) (Loizou, 2007):

$$\psi(\lambda,\phi) = \frac{\eta.|\overline{M}(\lambda,\phi-1)|^2}{\varepsilon_N(\lambda,\phi-1)} + (1-\eta).\max\left[0, \frac{|M(\lambda,\phi)|^2}{\varepsilon_N(\lambda,\phi)} - 1\right] \tag{7}$$

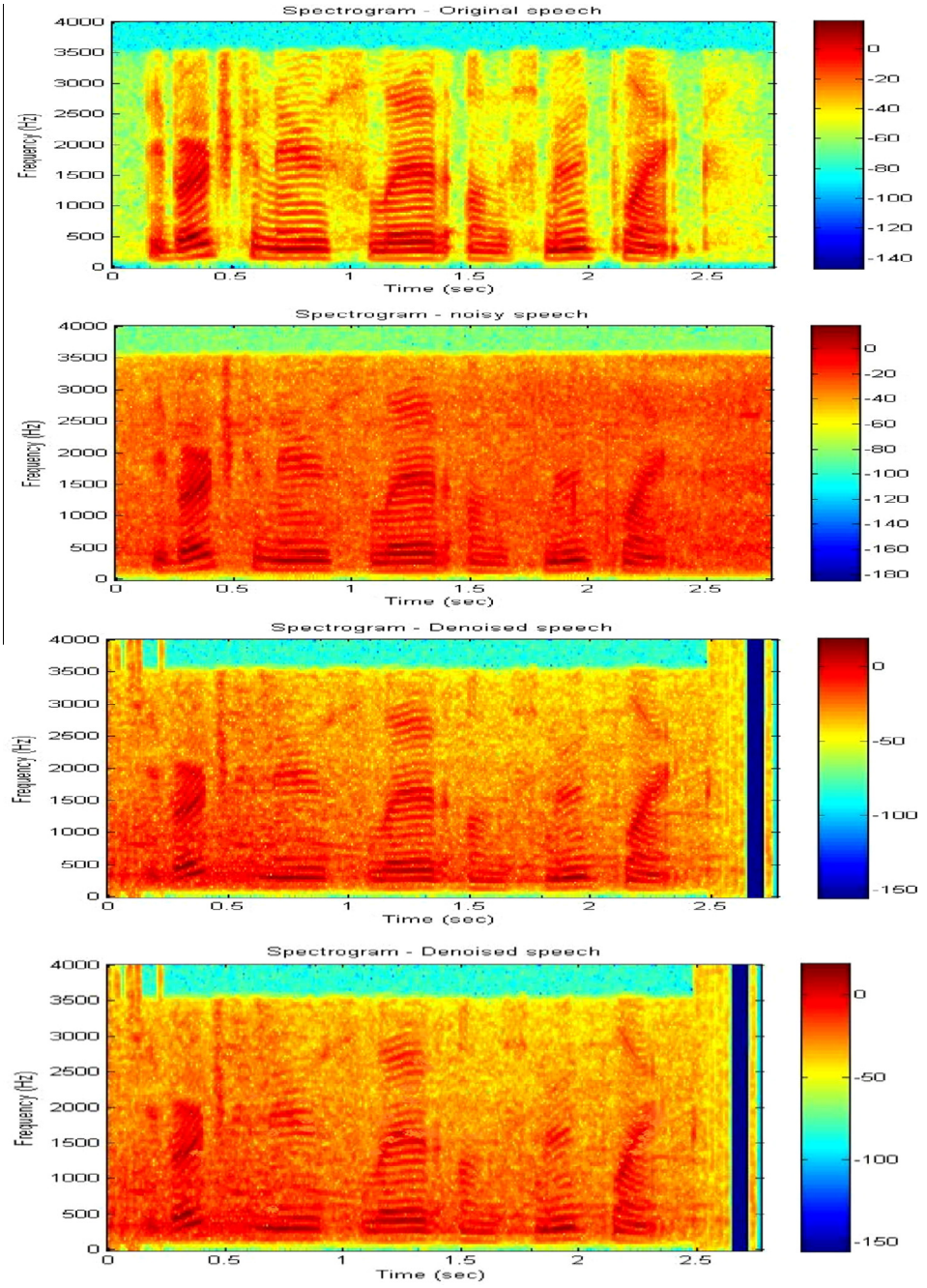

Figure 4 (a) Spectrogram of an original speech signal (b) Spectrogram of a signal corrupted by street at 10 dB SNR (c) Spectrogram of the estimated speech signal using optimal mask generation (d) Spectrogram of the estimated speech signal for a similar signal using optimal mask generation.

Subsequently, based on the quality ratio value R_Q, the speech spectrum of $M(\lambda,\phi)$ is classified into various classes Cl_1, Cl_2, Cl_3. If the ratio R_Q comes in below T_1, it is classified as Cl_1, else if between T_1 and T_2 it is classified as Cl_2, else it is classified as Cl_3.That is, it can be represented as:

$$M(\lambda,\phi) \in \left\{ \begin{array}{l} class\ Cl_1\ ,\ if\ R_Q \leqslant T_1 \\ class\ Cl_2\ ,\ if\ R_Q \leqslant T_2 \\ class\ Cl_3\ ,\ if\ R_Q > T_2 \end{array} \right\} \tag{8}$$

(b) Generation of optimal weight by cuckoo search:

Here the optimal weight mask is generated for each of the classes making use of the cuckoo search algorithm (Yang, 2009).

3.2.1. Initial population

Let the noisy speech input signal be represented by M, which is defined by $M = \{m_1, m_2, ..., m_{Ns}\}$, where Ns is the total number of input signals. The input signal is classified into class Cl_1, Cl_2 *or* $Cl_3$with the use of quality ratio. In order to obtain the best optimal binary mask with less iteration, first classify the units into different classes and generate the initial mask with the help classification module. Then, fitness (SNR) is computed for the initial population to find whether it is fixed to synthesis speech enhance signal.

3.2.2. New solutions

Then, with the help of initial mask, generate the new mask based on the equation of cuckoo search. Levi flight is performed on Y_i (initial mask) to yield to get a new cuckoo Y_i^*. Considering the signal y_{i1} in Y_i, then the changed value (new solution) y_{i1}^* is given by Yang (2009) and Valian et al. (2011):

$$y_{i1}^* = y_{i1}^{(t+1)} = y_{i1}^{(t)} + \Lambda \otimes Levy(x). \tag{9}$$

Here $\Lambda > 0$ is the step size which is greater than zero and normally it is taken as one and $\otimes$ means entry-wise multiplication. The Levi flight equation represents the stochastic equation for random walk as it depends on the current position and the transition probability (second term in the equation). Here, the levy distribution is given by:

$$Levy(x) = \sqrt{\frac{c}{2\pi}} \cdot \frac{e^{-\frac{1}{2}\left(\frac{c}{x}\right)}}{x^{3/2}} \tag{10}$$

where c is arbitrary constant. Hence, by performing Levi search, we obtain new solutions and then the fitness value (SNR value) of the new solution is found out. Let the fitness of the Levi performed nest be F_i.

Subsequently, some other nest is considered other than the i^{th} host nest and let the nest in consideration be represented by $Y_j = \{y_{j1}, y_{j2}, ..., y_{jNh}\}$ representing j^{th} host nest. The fitness of the j^{th} nest is found using the fitness function and is represented by F_j. If the fitness of the Levy flight performed i^{th} nest F_i is greater than fitness of the j^{th} nest F_j, then replace j^{th} nest signal values $Y_j = \{y_{j1}, y_{j2}, ..., y_{jNh}\}$ by the i^{th} host nest Levy performed values $Y_i^* = \{y_{i1}^*, y_{i2}^*, ..., y_{iNh}^*\}$. Initially when Levi flight is performed, corresponding fitness is found out F_i, compared to fitness of some other nest F_j and the replacement is carried out if the condition $F_i > F_j$ is satisfied.

3.2.3. Termination

After the comparison and replacements, we have to abandon a fraction of worst nests and build new nests in their place. This is done by finding the quality of all the current nests and analysing it. That is, keeping the best solutions and replacing the worst nests by newly built nests. Subsequently the solutions are ranked and the current best is found out. The full loop is continued till some stop criteria are met and the current best in the last loop performed will be the best solution. The optimal mask weight for the training signals will be the fitness function value obtained for the best solution.

3.3. Waveform synthesis module

In the enhancement module (testing phase), the test noisy speech signal is multiplied by the corresponding optimal binary mask obtained from the cuckoo search in the training module. Subsequently the resultant signals are synthesized to produce the enhanced speech waveforms. Fig. 3 shows the block diagram of the waveform synthesis module. Here,

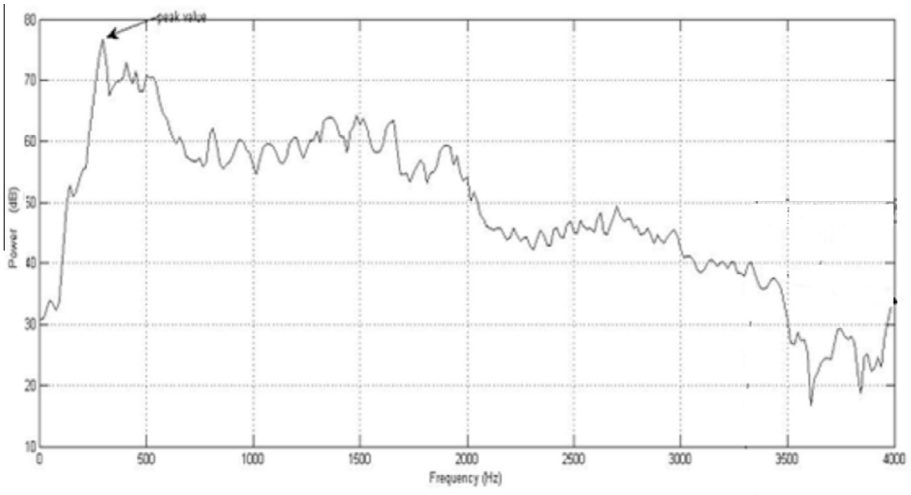

Figure 5 Estimation of PSD.

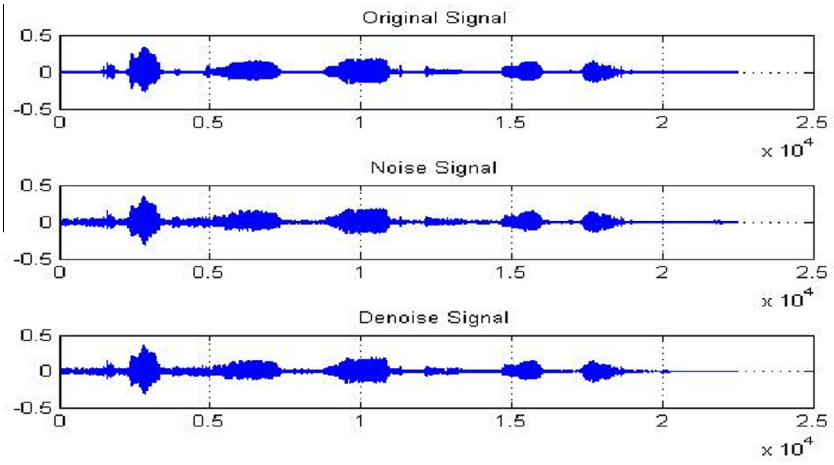

Figure 6 Input signal, noisy signal and denoized signal.

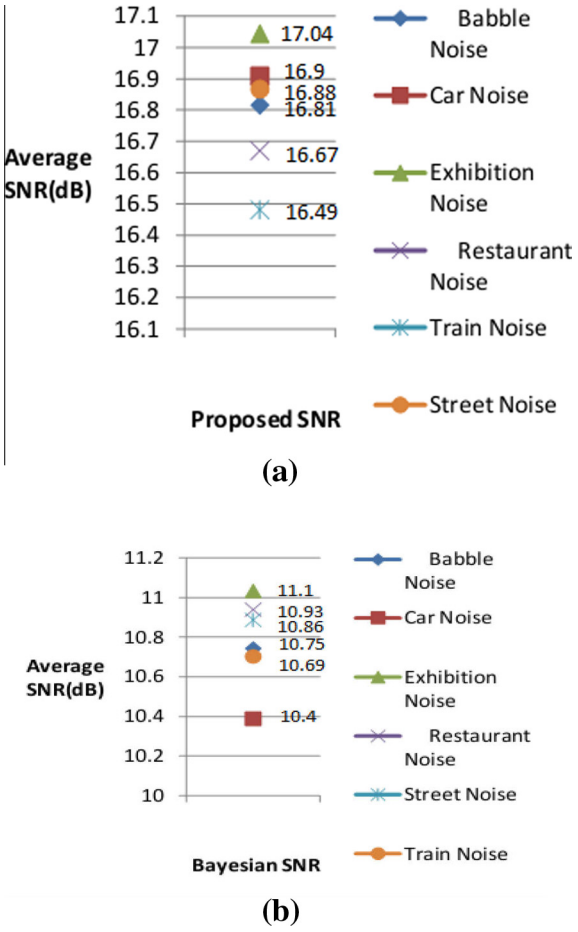

Figure 7 Plot of average SNR values for various noises and at various levels 0 dB, 5 dB, 10 dB, 15 dB using proposed approach (a) Bayesian approach (b).

initially the noisy speech signal is multiplied with the optimal mask generated from cuckoo search algorithm directly.

Let the noisy speech signal given as input for speech enhancement be represented as $T(k,t)$ and the optimal mask generated be represented as $O(k,t)$. The enhanced signal (represented as $E(k,t)$) is given by the following equation:

$$E(k,t) = O(k,t) * T(k,t) \tag{11}$$

So, finally the original speech signal is estimated after summing the weighted responses of the 25 signal components. Fig. 4 shows an example spectrogram of a synthesized signal using the proposed approach for speech enhancement (b) Spectrogram of a signal corrupted by street at 10 dB SNR (c) Spectrogram of the estimated speech signal using optimal mask generation. The spectrogram of the estimated speech signal using optimal mask generation shows the level of energy similar to the original speech signal energy level at the corresponding frequencies.

Fig. 5 shows the power spectrum magnitude (dB) vs frequency (Hertz). The power spectral density (PSD) describes how the power of a signal or time series is distributed with the frequency. PSD shows the energy of the signal as a function of frequency, which is the square of magnitude of absolute value of FFT of estimated signal. Power spectral density is used to describe the energy of the signal at various frequencies. It also signifies the variance which should be as small as possible to increase signal-to-noise ratio. The total power can be calculated after knowing the PSD and system bandwidth. The main contribution of the paper is the employment of cuckoo search for generating optimal mask for each class. Optimal mask generation results in having higher speech enhancement and noise reduction in comparison to existing techniques. Feature extraction using AMS also adds to the effectiveness of the proposed technique. Optimal mask is important as the enhanced signal is derived by multiplying the mask with the noisy signal. Hence finding the correct mask is very important. In our proposed technique, we employ cuckoo search which is effective for obtaining good optimal mask so as to obtain good results.

Pseudo code:
Input-noisy signal
Output-enhanced speech signal
Start
Extract features from the input speech corpus using Amplitude Magnitude Spectrogram using the equation:
$A(\lambda,\phi) = [A_F(\lambda,\phi), \Delta A_T(\lambda,\phi), \Delta A_S(\lambda,\phi)]$

(*continued on next page*)

Table 1 SNR for different cases.

Noise level (dB)	Babble noise		Car noise		Exhibition noise		Restaurant noise		Street noise		Train noise	
	Proposed SNR	Bayesian SNR	Proposed SNR	Bayesian SNR	Proposed SNR	Bayesian SNR	Proposed SNR	Bayesian SNR	Proposed SNR	Bayesian SNR	Proposed SNR	Bayesian SNR
0	5.874	2.171	5.6502	1.337	5.9468	1.640	5.5359	1.806	5.2881	1.699	5.7519	1.987
5	11.468	6.991	12.038	7.211	11.539	8.940	11.49	7.508	11.367	7.906	11.453	7.167
10	20.284	9.922	19.997	9.584	20.266	9.662	19.625	9.765	19.714	9.989	19.966	9.709
15	29.635	23.88	29.94	23.42	30.418	23.88	30.026	24.67	31.097	23.95	28.747	23.94

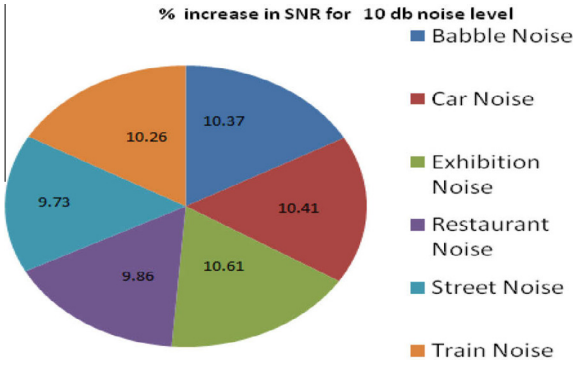

Figure 8 Percentage increase in SNR for 10 dB street noise level.

Classify each of the individual TF units by comparing with the original signal using: $M(\lambda,\phi) \in \left\{ \begin{array}{l} class\ Cl_1\ ,\ if\ R_Q \leqslant T_1 \\ class\ Cl_2\ ,\ if\ R_Q \leqslant T_2 \\ class\ Cl_3\ ,\ if\ R_Q > T_2 \end{array} \right\}$
Generate an optimal mask using cuckoo search
Multiply test noisy speech signal with the corresponding optimal binary mask obtained from the cuckoo search
$E(k,t) = O(k,t) * T(k,t)$
Synthesize the resultant signals to produce the enhanced speech waveforms given by
Stop

4. Experimental results and discussions

The proposed technique for speech enhancement and noise reduction is implemented in MATLAB Version 2012 and COLEA (Kim et al., 2009) on a system having 4 GB RAM with 32 bit operating system having i5 Processor. Dataset description is given in Section 4.1 and experimental results are given in Section 4.2.

4.1. Database description

The database used for the experimentation is taken from the Loizou's database given in Kim et al. (2009). The database was introduced to ease the assessment of speech improvement techniques. The noisy database comprises of thirty IEEE sentences degraded by eight diverse real-world noises at different SNRs. The noise was taken from the AURORA database (Hirsch and Pearce, 2000) and comprises suburban train noise, babble, car, exhibition hall, restaurant, street, airport and train-station noise. The IEEE sentence database was recorded in a sound-proof booth using Tucker Davis Technologies (TDT) recording equipment. The sentences were covered by three male and three female speakers. The sentences were originally sampled at 25 kHz and downsampled to 8 kHz.

4.2. Experimental results

The simulation results include plots of input signal, noisy signal and the de-noised signal shown in fig. 6. The signal power is plotted for the corresponding frequency, having a frequency range between 0 and 2.5 kHz. For this, various types of noise such as babble noise, car noise, exhibition noise, restaurant noise, street noise and train noise at different levels of 0 dB, 5 dB, 10 dB, 15 dB were used as maskers. Subjects participated in a total of 24 conditions [4 SNR levels (0 dB, 5 dB, 10 dB, 15 dB) × 6 types of maskers].

The results obtained proved the effectiveness of the proposed technique and its ability to suppress noise and enhance the speech signal. The graphical representation of percentage increase in SNR for various maskers at 10 dB level is shown in Fig. 8.

4.2.1. Inference of comparative analysis from (tables 1 and Figs. 7 and 8)

We have compared the proposed technique with the Bayesian Classifier using standard evaluation metrics of SNR. Various

Table 2 SSNR for different cases.

Noise level (dB)	Babble noise		Car noise		Exhibition noise		Restaurant noise		street noise		Train noise	
	Proposed SSNR	Bayesian SSNR	Proposed SSNR	Bayesian SSNR	Proposed SSNR	Bayesian SSNR	Proposed SSNR	Bayesian SSNR	Proposed SSNR	Bayesian SSNR	Proposed SSNR	Bayesian SSNR
0	−4.55	−7.13	−5.05	−7.68	−4.88	−7.46	−4.54	−7.06	−4.75	−7.64	−4.52	−7.35
5	−1.80	−5.39	−2.33	−5.40	−1.19	−4.99	−2.23	−5.20	−1.07	−5.28	−1.84	−5.43
10	1.09	−4.82	0.77	−4.83	0.96	−4.87	0.93	−4.75	1.55	−4.62	0.69	−4.78
15	4.17	−3.00	3.45	−3.16	3.70	−3.13	4.33	−2.95	4.61	−3.10	4.31	−3.07

types of noise taken include babble noise, train noise, car noise, exhibition noise, restaurant noise and street noise. In all the cases, noise at level of 0 dB, 5 dB, 10 dB and 15 dB has been considered. Fig. 7 gives the average SNR for the proposed and the Bayesian technique. Comparing with Bayesian the proposed technique has got better results which show the efficiency of the technique. Best SNR value obtained for the proposed technique is 31.0977 dB when compared to 24.67 dB for Bayesian technique. Average SNR value came about 16.79 dB with the proposed approach when compared to 10.78 dB for Bayesian technique. Fig. 8 gives the percentage increase in SNR for 10 dB noise level. The use of optimal mask has resulted in having better performance for the proposed technique. It is because the mask value is of great importance as this value is being multiplied to get the

Segmental signal-to-noise ratio (SSNR) computation is also carried out. Here, the techniques divides target and masker signals into segments. It subsequently computes segment energies, then SNRs, and returns mean segmental SNR (dB).

Table 2 gives the Segmented SNR values for the proposed and the Bayesian technique. From the values, we can observe that the proposed technique has achieved better SSNR values. The net average SSNR for the proposed technique came about 0.02 when compared to -5.31 for the Bayesian technique.

5. Conclusion

In this paper, cuckoo search based optimal mask generation for noise suppression and enhancement of speech signal is presented. The technique has three modules: Feature extraction module, optimal mask generation module and the waveform synthesis module. Feature extraction is carried out using AMS and classification of signals is done to generate the initial population of cuckoo search algorithm. The Simulation of the proposed technique was carried out using various datasets. It was also compared with the previous techniques using SNR parameter. The results obtained proved the effectiveness of the proposed technique and its ability to suppress noise and enhance the speech signal. Best SNR value obtained for the proposed technique is 31.0977 dB whereas it is 24.67 dB using Bayesian technique. Average SNR value came about 16.79 dB with the proposed approach when compared to 10.78 dB for Bayesian technique. Large gains in intelligibility were achieved with the proposed approach using a limited amount of training data. Overall, the summary of finding using proposed approach suggests that speech intelligibility can be improved by estimating the signal-to-noise ratio in each time–frequency unit.

References

Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process 27, 113–120.

Brungart, D., Chang, P., Simpson, B., Wang, D., 2006. Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. J. Acoust. Soc. Amer. 120, 4007–4018.

Chen, F., Loizou, C., 2011. Impact of SNR and gain function over – and under-estimation on speech intelligibility. Speech Commun. 54, 272–281.

Chirstiansen, C., Pedersen, M.S., Dau, T., 2010. Prediction of speech intelligibility based on an auditory preprocessing model. Speech Commun. 52, 678–692.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans. Acoust Speech Signal Process ASSP-32 (6), 1109–1121.

Hirsch, H., Pearce, D., 2000. The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions. ISCA ITRW ASR, September 18–20.

Hong Y.L., Qing H.Z., Guang L.R., Bao J.X., Speech Enhancement algorithm Based on Independent Component Analysis. 5th IEEE International Conference on Natural Computation, 2009, pp. 598–602.

Hu, Y., Loizou, P., 2007. Subjective comparison of speech enhancement algorithms. Speech Commun. 49, 588–601.

Kim, Gibak, Loizou, Philipos C., 2010. Improving speech intelligibility in noise using environment optimized algorithms. IEEE Trans. Audio Speech Lang. Process. 18 (8), 2080–2090.

Kim, G., Loizou, C., 2010. A new binary mask based on noise constraints for improved speech intelligibility. Interspeech, Chiba, Japan, 1632–1635.

Kim, Gibak., Loizou, Philipos C., 2011. Reasons why speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. IEEE Trans. Audio Speech Lang. Process. 19 (1), 47–56.

Kim, Gibak, Yang, Lu, Yi, Hu, Loizoua, Philipos C., 2009. An algorithm that improves speech intelligibility in noise for normal-hearing listeners. J. Acoust. Soc. Am. 126 (3), 1486–1492.

Li, N., Loizou, P.C., 2008. Factors influencing intelligibility of ideal binary-masked speech: implications for noise reduction. J. Acoust. Soc. Amer. 123 (3), 1673–1682.

P.C. Loizou, 2006, Speech processing invocoder-centric cochlear implants, In: Møller, A.R. (Ed.), Cochlear and Brainstem Implants, Advances in Oto- Rhino-Laryngology, Karger, Basel, Switzerland, 64, pp. 109–143.

Loizou, P.C., 2007. Speech Enhancement: Theory and Practice. CRC Press.

Youyi, Lu, Cooke, Martin, 2009. The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise. Speech Commun. 51, 1253–1262.

Lu, Y., Loizou, P., 2011. Estimators of the magnitude-squared spectrum and methods for incorporating SNR uncertainty. IEEE Trans. Audio Speech Lang. Process. 19 (5), 1123–1137.

Jianfen, Ma, Loizou, P.C., 2010. SNR loss: a new objective measure for predicting the intelligibility of noise-suppressed speech. Speech Commun. 53, 340–354.

Mandal, Sangeeta, Ghoshal, Sakti Prasad, Kar, Rajib, Mandal, Durbadal, 2012. Design of optimal linear phase FIR high pass filter using craziness based particle swarm optimization technique. J. King Saud Univ. Comp. Inform. Sci. 24 (1), 83–92.

Muhammad, Ghulam, 2010. Noise-robust pitch detection using auto-correlation function with enhancements. J. King Saud Univ. – Comp. Inform. Sci. 22, 13–28.

Salivahanan, Gnanapriya, 2010. Digital signal processing, second ed. Tata McGraw Hill.

Scalart, P., Filho, J.V., 1996. Speech enhancement based on apriori signal to noise estimation, In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2. IEEE, pp. 629–632.

Valian, E., Mohanha, S., Tavakoi, S., 2011. Improved Cuckoo search algorithm for feed forward neural network training. Int. J. Artificial Intelligence Appl. 2 (3), 36–42.

Venkata Rao, R., Waghmare, G.G., 2014. A comparative study of a teaching–learning-based optimization algorithm on multi-objective unconstrained and constrained functions. J. King Saud Univ. Comp. Inform. Sci. 26 (3).

Wolfe, P.J., Godsill, S.J., 2003. Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement. EURASIP J. Appl. Signal Process. 2003 (10), 1043–1051.

Yang, Xin.-She, 2009. Cuckoo Search via Lévy flights. Nat. Biol. Inspire. Comput., 210–214