Retraction Notice: Speech enhancement method using deep learning approach for hearing-impaired listeners

At the request of the Journal Editor and SAGE Publishing, the following article has been retracted.

Khaleelur Rahiman PF, Jayanthi VS and Jayanthi AN (2020) Speech enhancement method using deep learning approach for hearing-impaired listeners. *Health Informatics Journal*. Epub ahead of print 23 January 2020. DOI: 10.1177/1460458219893850.

This article contains substantial unreferenced overlap with material from other sources (1-4). An earlier version of this article (4) was previously retracted from *Medical & Biological Engineering & Computing* and the authors had not disclosed this to the Editor on submission.

In addition, this article contains manipulated data that reduces the validity of the reported findings.

The unattributed excerpts in the article were taken from the following sources:

1. Goehring Tobias, Bolner Federico, Monaghan Jessica JM, van Dijk Bas, Zarowski Andrzej, Bleeck Stefan (2017) Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users. *Hearing Research* 344: 183-194.
2. Acharya U. Rajendra, Oh Shu Lih, Hagiwara Yuki, Tan, Jen Hong, Adam Muhammad, Gertych Arkadiusz and Tan, Ru San (2017) A Deep Convolutional Neural Network Model to Classify Heartbeats. *Computers in Biology and Medicine*. 89(1): 389-396.
3. Monaghan Jessica JM, Goehring Tobias and Xin Yang (2017) Auditory inspired machine learning techniques can improve speech intelligibility and quality for hearing-impaired listeners. *The Journal of the Acoustical Society of America* 141(3). DOI: 10.1177/1477370819839620.
4. Khaleelur Rahiman PF, Jayanthi VS and Jayanthi AN (2019) Deep convolutional neural network-based speech enhancement to improve speech intelligibility and quality for hearing-impaired listeners. *Medical & Biological Engineering & Computing* 57 (4):757-759.
RETRACTED: Speech enhancement method using deep learning approach for hearing-impaired listeners

PF Khaleelur Rahiman
Hindusthan College of Engineering and Technology, India

VS Jayanthi
Rajagiri School of Engineering and Technology, India

AN Jayanthi
Sri Ramakrishna Institute of Technology, India

Abstract
A deep learning-based speech enhancement method is proposed to aid hearing-impaired listeners by improving speech intelligibility. The algorithm decomposes the noisy speech signal into frames (as features). Subsequently, a deep convolutional neural network is fed with decomposed noisy speech signal frames to produce frequency channel estimation. However, a higher signal-to-noise ratio information is contained in produced frequency channel estimation. Using this estimate, speech-dominated cochlear implant channels are taken to produce electrical stimulation. This process is the same as that of the conventional n-of-m cochlear implant coding strategies. To determine the speech-in-noise performance of 12 cochlear implant users, the fan and music sound applied are considered as background noises. Performance of the proposed algorithm is evaluated by considering these background noises. Low processing delay and reliable architecture are the best characteristics of the deep learning-based speech enhancement algorithm; hence, this can be suitably applied for all applications of hearing devices. Experimental results demonstrate that deep convolutional neural network approach appeared more promising than conventional approaches.

Keywords
cochlear implant, convolutional neural networks, impaired listener, speech intelligibility

Corresponding author:
PF Khaleelur Rahiman, Electronics and Communication Engineering, Hindusthan College of Engineering and Technology, Coimbatore, Tamil Nadu 641050, India.
Email: khaleelurphd@gmail.com
Introduction

Speech recognition in background noise seems to be complex for hearing impairment individuals. Many users using conventional cochlear implant (CI) devices in quiet acoustic conditions have achieved near-to-normal speech understanding situation. However, speech understanding ability of the CI users is negatively affected with the presence of competing talkers and environmental sounds (i.e. background noises). Speech reception threshold (SRT) is applied to measure the performance degradation. In other words, this can be defined as signal-to-noise ratio (SNR) where speech intelligibility achieved is 50 percent. When compared to the SRTs of normal-hearing (NH) listeners, the SRTs of CI users can typically vary in higher (worse) range from 10 to 25 dB. Based on the reports in literature, slow amplitude fluctuations or temporal gaps of stationary noise masker benefits have been slightly enjoyed by the CI recipients compared to the speech intelligibility of NH listeners. Nonetheless, release from masking has defined this process. Most powerful spectral channels in small quantity are formed by reducing the spectral information suggested by a CI. Temporal information is largely used by the CI users but when compared to NH listeners the modulated masking noise is highly susceptible by the CI users. Probably, CI user’s speech understanding performance is decreased through combining increased modulation interference and decreased spectral resolution. This is evident with CI simulations tested for NH listeners. Noise component of the noisy mixture can be attenuated using the proposed speech enhancement (SE) algorithms; hence, the perceived quality and intelligibility of the speech component is increased. Limited success has been achieved by adopting conventional single-channel SE techniques. The recognition ability of NH and hearing-impaired (HI) listeners is increased using the “auditory masked threshold noise suppression” technique under noisy conditions. Conversely, babble noise for HI listeners, speech-shaped quality and speech intelligibility in speech-shaped noise were improved using the laboratory-tested sparse code shrinkage algorithm. Past works have reported that using single-channel enhancement algorithms for HI listeners has shown poor performance in word recognition but not for listener preference.

In the current trend, machine learning approaches have proved its great strength in speech intelligibility improvement for CI users, NH listeners, HI listeners, and medical signals. Normally, based on the incoming signal, gain function for noise and speech statistics is estimated. Nonetheless, the optimal gain function was estimated using the traditional machine learning approaches by means of incorporating prior knowledge of speech and noise patterns. This estimation is done to apply for the incoming signal. Speech intelligibility of both CI users and NH listeners is improved by applying Gaussian mixture models. NH and HI listener’s speech intelligibility scores are highly improved using a deep neural network (DNN) algorithm. In most of the research field, a machine learning method called deep learning is applied widely due to its improved classification performance. In other words, deep learning is mostly used for speech separation, SE, and speech recognition. Depending on time–frequency units, the binary or soft classification decision is made using some data-driven methods to achieve SE. Example for this is monaural speech denoising achieved through estimating smoothed ideal ratio mask (IRM) or ideal binary mask (IBM).

Speech intelligibility can be effectively improved using the hard targets IBM. Alternatively, speech quality can be improved by providing better prediction against soft targets IRM. At each frequency unit, a suppression gain can be inferred on IRM in the range of [0, 1]. Noisy features and estimated IRM are multiplied in an element-wise manner to obtain the improved features as a final output. Reduced speech distortion is achieved by suppressing noise in some degrees using efficient soft masking algorithms. Apart from this IRM direct prediction, joint optimization of DNNs having an additional masking layer and masking functions is examined and analyzed in
Furthermore, speech spectral is mapped directly using deep learning approaches for defeating the time–frequency mask prediction. Xu et al. have used a large collection of heterogeneous data to train deep and extensive neural network architecture as well as to propose a DNN-based regression framework. The reason for the introduction of DNN-based regression approach is that it has acquired the ability to maintain both the high non-stationary noises and non-linear noises and to avoid the unnecessary assumptions on statistical properties of signals.

Estimated clean speech signal in some cases has suffered from distortions due to the considerable noise removal from the noisy speech using the regression DNN. However, removing such distortions turns to be a tedious and challenging task. To solve this tedious task, variance equalization of features was applied for further post-processing the regression DNN. Thus, the distortions included in estimated clean features are removed effectively. Unseen noise conditions can be determined using the generalization capacity of deep learning approaches. Xu et al. have used dynamic noise aware training method to enhance the generalization capability of deep learning approaches. Notably, the multi-objective framework has been developed through extending the DNN architecture. For speech recognition, Kim and Smaragdis have used the adaptation methods and improved the denoising autoencoder (DAE) performance at the test stage by means of applying a fine-tuning scheme. In low SNR condition, degradation of performance is considered to be another major challenging task. Under critical noise conditions, the speech intelligibility was enhanced in the work by Gao et al. by means of integrating SE and voice activity detection (VAD) to a proposed joint DNN framework. In view of research point, SE using the more complex neural network is still in need to be studied and analyzed. Long short-term memory (LSTM) is explored in the work by Weninger et al. Conversely, Fu et al. have analyzed the convolutional neural network (CNN). Inferring source spectral and target source spectral are learned using the architecture with dual outputs by Tu et al.

In this work, we studied the speech intelligibility improvement of CI users in different background noises through combining deep convolutional neural networks (DCNNs) with an SE algorithm. The noisy speech signal is decomposed into features (frames) using an SE algorithm. Then, a DCNN is fed with decomposed noisy speech signal frames to produce frequency channel estimation. However, a higher SNR information is contained in produced frequency channel estimation. Using this estimate, speech-dominated CI channels are taken to produce electrical stimulation. This process is the same as that of the conventional n-of-m CI coding strategies. To determine the speech-in-noise performance of 12 CI users, the fan and music sound applied are considered as background noises. Performance of the proposed algorithm is evaluated by considering these background noises. The short summarizations of this work are as follows:

Our main contributions are summarized as follows:

- We proposed a nine-layer DCNN to evaluate the SE for improving speech intelligibility in noise for CI users.
- Group participation is made with 12 CI users having Cochlear Nucleus® CI implanted, and all are native Tamil speakers.
- Performance of the proposed algorithm is evaluated through conducting extensive simulation, compared, and analyzed the outcomes with the conventional methods.

Rest of this article is organized as follows: deep learning-based speech enhancement (DLSE)-based SE algorithm is explained in section “Proposed algorithm description.” Experimental setup of the proposed technique is described in section “Materials and methods.” Results and discussions are explained in section “Result and Discussion.” Section “Conclusion” concludes the article.
Proposed algorithm description

CI speech processing strategy implementation of Advanced Combination Encoder (ACE™) is integrated with the DLSE algorithm. MATLAB simulation of the proposed system is illustrated in Figure 1.

Reference strategy

An implementation of a research ACE strategy acts as the reference strategy. To an automatic gain control (AGC), the noisy speech signals that are passing using a pre-emphasis filter and downsam-pled to 16 kHz are transmitted. Essentially, the acoustic dynamic range is compressed using AGC. Then, to a CI recipient having a smaller electrical dynamic range, this can be conveyed easily. Subsequently, to the compressed signal, a filter bank based on a fast Fourier transform (FFT) was deployed. For each of the \( M \) frequency channels (typically, \( M = 22 \)), an estimate of the envelope was provided by the output of the complex FFT magnitude. Then, only for one electrode, each channel was allocated. To retain the subset of \( N \) channels having greater envelope magnitudes (while setting the subjects of CI processor, an audiologist set \( N < M \)), the envelope for each channel is mapped instantaneously with the loudness growth function (LGF). Also for electrical stimulation, the subject’s dynamic range varies between the threshold level (THL) and maximum comfortable loudness level (MCL) (the parameters THL and MCL are used from the subject’s CI processor). At last, the cycle of stimulation can be completed through sequentially stimulating the electrodes related to the selected channels. Notably, the channel stimulation rate is defined as a number of cycles per second, and the channel stimulation rate is \( N \) times the total stimulation rate.

Deep CNN classifier for SE

To an electrical output, the frequency channel envelope is transformed directly with CI processing and also it does not demand any of the reconstruction stages. Instead of doing pre-processing of the noisy signal, we prefer to combine the DLSE directly into the CI signal path. However, redundant synthesis stage is avoided and also delay and complexity of the system are increased with the introduction of additional noise. Essentially, pre-processing and DCNNs are the two main components of the DLSE algorithm.

Feature transformation (pre-processing). During pre-processing, elimination of silent frames from the signal and downsampling the speech signals to 8 kHz is performed. Amplitude scaling issues are addressed through normalizing each frame using Z-score normalization. Subsequently, training
and testing stage in one-dimensional deep learning CNN should be conducted only after feeding the speech segments by removing the offset. Significantly, 100ms hamming window included in 256-point short-time Fourier transform is applied to compute the spectral vectors. However, the window shift is fixed to 3ms (64-point). For each frequency bin, the frequency resolution was fixed to $4kHz/128 = 31.25$Hz. Symmetric half was removed to 129-point by reducing 256-point scale invariant feature transform (SIFT) magnitude vectors. Nonetheless, eight consecutive noisy SIFT magnitude vectors with duration 100ms and size $129 \times 8$ included in the input feature are used by CNN for further processing. To obtain the unit variance and zero mean, it is important to standardize both the input features.

**CNN architecture.** Among artificial neural networks, CNN plays a common role. A multilayer perceptron (MLP) is resembled by a CNN. An activation function included in every single neuron of MLP can perform mapping of weighted inputs to the outputs. In the network, more than one hidden layers are integrated to convert a single MLP into deep MLP. Likely, using a special structure, the CNN is assumed to behave the same as that of MLP. Following the architecture model, rotation invariant and translation properties can be shown by CNN with the support of special structure. CNN architecture includes a rectified linear activation function and additional three common layers, such as a convolutional layer, pooling layer, and fully connected layer. The architecture of the proposed CNN model is summarized in Table 1. Simply, CNN architecture is made up of nine layers (i.e. three convolutional layers, three max-pooling layers, and three fully connected layers). This arrangement of CNN layers is shown in Figure 2. Using equation (1), the kernel sizes (3, 4, and 4) are used by their respective layer for convolution. This is done for every convolution layer (Layers 1, 3, and 5). To the feature maps, a max-pooling operation is employed next to every convolution layer. Notably, the feature map size is reduced by the max-pooling operation. In this article, the brute force technique is applied to obtain the filter (kernel) size parameter. However, 1 and 2 are the stride fixed for convolution and max-pooling operation. Layers 1, 3, 5, 7, and 8 have obtained activation function using the leaky rectifier linear unit (LeakyRelu). Output neurons (i.e. 30, 20, and 5) are included in the fully connected layers. Moreover, Layer 9 (final layer) consists of $w$ output neurons. Wiener gain function is applied to compute the softmax function.

**Table 1. Measures of proposed CNN model.**

| Layers | Type               | No. of neurons (output layer) | Kernel size | Stride |
|--------|-------------------|-------------------------------|-------------|--------|
| 0–1    | Convolution       | $258 \times 5$                | 3           | 1      |
| 1–2    | Max pooling       | $129 \times 5$                | 2           | 2      |
| 2–3    | Convolution       | $126 \times 10$               | 4           | 1      |
| 3–4    | Max pooling       | $63 \times 10$                | 2           | 2      |
| 4–5    | Convolution       | $60 \times 20$                | 4           | 1      |
| 5–6    | Max pooling       | $30 \times 20$                | 2           | 2      |
| 6–7    | Fully connected   | 30                             | –           | –      |
| 7–8    | Fully connected   | 20                             | –           | –      |
| 8–9    | Fully connected   | 5                              | –           | –      |

Wiener gain function is applied to compute the estimated gain

$$G(b, \tau) = \frac{SNR_p(b, \tau)}{\sqrt{1 + SNR_p(b, \tau)}}$$  \(1\)
Figure 2. System design of proposed CNN model. CNN: convolutional neural network.
where $\tau$ is the frame, $b$ is the Gammatone frequency channel. Here, priori SNR is indicated as $SNR_p$ which is shown in the below equation as follows

$$SNR_p(b, \tau) = \frac{\alpha |X(b, \tau - 1)|^2}{\lambda_D(b, \tau - 1)} + (1 - \alpha) \max \left(\frac{|S(b, \tau)|^2}{\lambda_D(b, \tau)}, -1, 0\right)$$

(2)

where $\alpha = 0.98$ is a smoothing constant, $\lambda_D$ is the estimate of the background noise variance.

Using the below equation, the convolution operation is calculated

$$x_n = \sum_{k=0}^{N-1} y_k f_{n-k}$$

(3)

where $N$, $n$, and $y$ represent the number of elements in $y$, filter, and signal, respectively. The term $x$ denotes the output vector.

Considering the feature set (input sample size) in the back propagation algorithm, the training stage of CNN is conducted. Parametric measure for the hyper-parameters, namely momentum, learning rate, and regularization ($\lambda$) is fixed to 0.7, $3 \times 10^{-3}$, and 0.2, respectively. Benefits of using these hyper-parameters in the training stage are as follows: (a) controlling the learning speed, (b) assisting the data convergence, (c) and managing overfitting of data. Optimal performance can be achieved by adjusting the parameters depending on the brute force technique. Using equations (4) and (5), the weights and biases are updated

$$\Delta W_i(t + 1) = -\frac{x}{r} \frac{\partial C}{\partial W_i} + m \Delta W_i(t)$$

(4)

$$\Delta B_i(t + 1) = -\frac{x}{n} \frac{\partial C}{\partial B_i} + m \Delta B_i(t)$$

(5)

where $C$, $t$, $m$, $n$, $x$, $\lambda$, $l$, $B$, and $W$ indicate the cost function, updating step, momentum, the total number of training samples, learning rate, regularization parameter, layer number, bias, and weight, respectively.

Materials and methods

Hardware/software

In MATLAB platform, DLSE algorithm and research ACE strategy are implemented. The ACE strategy implemented in a computer is applied to process the stimuli. Then, the implant users are directly presented with the processed stimuli.

L34 experimental processor is used to connect the Cochlear NIC3 interface for delivering electrical stimulation. Subsequently, a coil receives radio frequency output delivered by the system. Ultimately, the subject’s implant contains the stimulus data that are transmitted by a coil.

Subjects

Group participation is made with 12 CI users having Cochlear Nucleus CI implanted, and all are native Tamil speakers. A total of 12 CI users (files) are included in the Tamil speech database.
However, 12 vowels sound (each sound extending to 3 s) and 18 consonant sounds comprised in each file. The size of sampling frequency is 16,000 Hz and each one holding 16 bits per sample.

During the initial study, 61 years was fixed as the mean age of the group; thereby, it may range from 23 to 81 years. For testing, single ear of each subject is considered. However, it is important to turn off the CI or hearing aid (HA) fixed to the contralateral side of each subject. At the initial stage of the study, 9.8 years was fixed as the implants mean duration ranging from 1.2 to 3.6 years. In ACE strategy, all subjects have acted like users. Table 2 indicates the subject’s demographic data.

**Table 2. Demographic data.**

| Parameter       | Values                        |
|-----------------|-------------------------------|
| Momentum parameters | 0.7                           |
| Learning rate   | \(3 \times 10^{-3}\)          |
| Regularization (\(\lambda\)) | 0.2                           |
| \(p\) value     | When compared to the selected significance level (i.e. 0.05), the obtained \(p\) value is small, from which it is evident that better significant results are yielded. |
| Vowels          | 12                            |
| Consonants      | 8                             |
| Gender          | 4 males, 8 females             |
| Speakers        | 12                            |
| Age group       | 23–81 years                   |
| Mother tongue   | Tamil                         |

Processing conditions and stimuli

Target speech material is assumed to the vowels and consonants of Tamil language. Noisy environment applied in this study is divided into two forms namely, fan and music background. Then, vowels and consonants (i.e. two speech materials) were used to test all CI users under these noisy environments.

Processing conditions are of two forms:

- Unprocessed condition (UN): ACE, unprocessed condition,
- DLSE: DLSE, processed condition.

**Used protocol**

An adaptive procedure is used to measure the type I and type II error rates for two environmental conditions (i.e. two processing conditions (UN, DLSE) \(\times\) two maskers (fan and music background)). This computation is performed in a sound-treated room with the help of an audiologist. During testing, the processing conditions, both the audiologist and subject, were kept in unseen condition (i.e. blind).

**Evaluation**

For each processing condition, the objective analysis is made before clinical testing. At different SNRs, the electrodograms were evaluated. Then, reference electrodogram is applied to compare the computed electrodograms using type I and type II error rates. Normalized values ranging from...
0 to 1 is contained in the stimuli of an electrodogram. These values indicate the electrical perception varying from comfort level and threshold included in the frequency channel and frames, correspondingly. In quiet environment, with the presence of ACE, the processing speech was created using reference electrodogram. However, not including DLSE in this environment can produce good noise reduction result by means of applying processing speech for generating reference electrodogram.

Type I and type II error rates were obtained by dividing the total number of possible errors to the summed type I and type II errors across all frames and frequency channels. With ACE maxima selection (11 chosen channels), the average error rates computed over 10dB SNR and 20 characters at −4, 0, 4 are considered as the error rates of processing condition. Thereby, both the error types include equal number of possible errors and also control the introduction of a bias toward 2.

Figure 3 displayed the outcomes of the objective analysis. UN providing the type I and type II error rates for music background changes from 36 to 66 percent and 95 to 15 percent which means SNR = −5 and 10 dB, respectively. Similar error rates are provided by the DLSE conditions with the cost of somewhat higher type II error rates (⩽14% and ⩽20% at −4 and 10 dB SNR, respectively) and with highly reduced type I error rates (⩽6% and ⩽17% at −4 and 10 dB SNR, respectively). UN providing type I error rates and type II error rates varying from 20 to 42 percent and 4 to 10 percent (i.e. SNR = −4 and 10 dB, respectively) is indicated for fan background noise. With the cost of higher type II error rates, type I error rates are reduced using both the DLSE conditions.

Ultimately, not only noise is highly reduced with the DLSE algorithm but also some speech removal distortions are introduced. Compared to UN, the total error is also reduced for all SNRs and noise. These results support CI users in testing clinical speech performance and also achieved an improvement in speech perception.

**Result and discussion**

Figures 4–7 indicates samples of consonant letters and Tamil vowel for 11 maxima containing ACE strategy. Clean speech signals electrodogram is represented for each figure (as top panel).

With the presence of two kinds of background noises (fan and music), the speech was disrupted in the second panel. DLSE processing method conditions are indicated in third panel. Ultimately, it is evident from Figure 4 that the proposed DLSE method is so beneficial for CI recipients by means of providing effective and superior intelligibility performance.

Proposed DLSE-based approach efficiency is analyzed by comparing two neural network-based approaches: (a) comparison feature set (NN_COMP), (b) auditory feature set (NN_AIM) and Wiener filtering (WF). *Comparison feature set:* Using the same set of features (depending on a complementary set of features), NN_COMP (comparison feature set) is generated. Moreover, from each 20 ms long timeframe concerning the noisy speech mixture, the mel-frequency cepstral coefficients (MFCCs), relative-spectral transform and perceptual linear prediction coefficients (RASTA-PLP), and the amplitude modulation spectrum (AMS) were extracted to produce the feature set. A dimensionality of 445 per timeframe (AMS (25 × 15) + RASTA-PLP (3 × 13) + MFCC) comprised the concatenated features. The delta–delta features for RASTA-PLP only (as described by Healy et al. and delta (consecutive frame features difference) are used for concatenating the current timeframe which was extracted from the NN_COMP. *Auditory feature set:* The auditory image model was used to the NN_AIM auditory feature set. For better comparing the two processing conditions, the statistical analysis was done individually for each noise type. Table 1 depicts the demographic data for this work. Based on the test, different probability distributions are used to compute the p values. If the selected significance level (generally 0.05) is greater than the p value then a good result is generated.
Speech intelligibility

For four SE algorithms, the SNRs of 0 and +4 dB was fixed to determine the speech intelligibility of both music and fan background noise conditions. Subsequently, the comparison is made with unimproved noise conditions. Clinical results of 12 CI users presented for the listening test are reported in this study.

**Figure 3.** Error rate analysis for two noises −4, 0, and +4 dB SNR under UN and DLSE processing conditions.

SNR: signal-to-noise ratio; DLSE: deep learning-based speech enhancement.
Results obtained for 12 CI users are tested on four SE algorithms by setting the SNR levels from 0 to +4 dB, and the mean scores of noise speech are indicated in Figures 5 and 6. Averaged Character (letter) Correct rate (CCR) is the measure used for reporting the performance and it is shown in Figure 5. CCR indicates the ratio of a total number of characters determined on each testing condition $C_t$ to the number of accurately determined character $C_c$.

Figure 4. Male speaker generating the Tamil vowel letter “ṉ (ā)” electrodogram at a level of 65 dB SPL. The clean signal is indicated in the top panel. (a) Fan background noise and (b) music background noise signals are shown in the second panel. DLSE proposed condition is depicted in the third panel. DLSE: deep learning-based speech enhancement; SPL: sound pressure level.
Figure 5. Obtained group-mean percentage of correctly recognized characters in terms of each algorithm for (a) upper panel (fan background noise) and (b) lower panel (music background noise) conditions. SNR: signal-to-noise ratio; DLSE: deep learning-based speech enhancement; WF: Wiener filtering.
Figure 6 indicates that high intelligibility scores are obtained with the proposed DLSE-based approach compared to other existing approaches.

**Speech quality**

For both SNR of 0 and +4 dB and two noise conditions, the speech quality ratings were identified in terms of each algorithm. The obtained results are indicated in Figure 7. Nonetheless, in most of the conditions, the normal distribution of data was not possible. Thus, non-parametric statistics are used and whisker and box plots are indicated. At the higher SNR, the speech quality is improved for achieving the speech intelligibility. Also, when compared to the music background noise, the proposed algorithm has shown better improvements in fan background noise.

Value of NN_AIM \( (p=0.017) \) fixed at +4 dB SNR has shown significant quality rating improvement in the music background noise conditions (i.e. quality rating gained was 0.81). When fixed, NN_AIM \( (p=0.0073) \), NN_COMP \( (p=0.017) \), and sparse coding \( (p=0.0021) \) at 0 dB SNR have shown significant improvements in the fan background conditions (i.e. improvement achieved is 0.44, 0.51, and 0.66, respectively). Also, value of NN_AIM \( (p=0.0011) \) and NN_COMP \( (p < 0.001) \) fixed at 4 dB SNR has shown significant improvements (i.e. 0.98 and 0.85, respectively).

**Effects of audibility**

For each participant and condition, their speech audibility was measured through computing the speech intelligibility index (SII; ANSI, 1997). Table 3 depicts the SII values. The shadow filtering, noise after the deployment of enhancement gain function, and speech spectra were adopted

\[
CCR = \frac{C_r}{C_s} \times \% \tag{6}
\]
for calculating SII to compensate with the enhanced conditions. However, a gain function is not used on sparse code processing but, depending on the difference between the original level and one-third octave bands and enhanced signals in 10 ms frames, a gain function was computed for calculating SII.

Furthermore, benefits on varying the speech spectrum level are determined by computing the SII value for original noise spectrum and enhanced speech spectrum. However, in response to

Figure 7. Box and whisker plots of speech quality ratings for different algorithms in (a) the fan background noise and (b) music background noise conditions. SNR: signal-to-noise ratio; DLSE: deep learning-based speech enhancement; WF: Wiener filtering.
Table 3. Each subject and conditions speech intelligibility index values.

| Subject | 0 dB SNR | 4 dB SNR |
|---------|----------|----------|
|         | UN       | NN_COMP  | NN_AIM  | DLSE  | UN       | NN_COMP  | NN_AIM  | DLSE  |
| Fan background |          |          |         |       |          |          |         |       |       |
| 1       | 0.413    | 0.593    | 0.591   | 0.601  | 0.527    | 0.667    | 0.663    | 0.668 |
| 2       | 0.427    | 0.607    | 0.605   | 0.608  | 0.548    | 0.668    | 0.665    | 0.670 |
| 3       | 0.392    | 0.540    | 0.538   | 0.548  | 0.488    | 0.599    | 0.596    | 0.600 |
| 4       | 0.449    | 0.668    | 0.667   | 0.670  | 0.584    | 0.747    | 0.743    | 0.750 |
| 5       | 0.394    | 0.562    | 0.560   | 0.566  | 0.502    | 0.633    | 0.630    | 0.635 |
| 6       | 0.389    | 0.549    | 0.546   | 0.550  | 0.487    | 0.617    | 0.614    | 0.620 |
| 7       | 0.427    | 0.642    | 0.641   | 0.645  | 0.555    | 0.729    | 0.725    | 0.730 |
| 8       | 0.402    | 0.569    | 0.567   | 0.670  | 0.509    | 0.639    | 0.635    | 0.642 |
| 9       | 0.453    | 0.665    | 0.664   | 0.668  | 0.584    | 0.733    | 0.725    | 0.736 |
| 10      | 0.409    | 0.574    | 0.573   | 0.578  | 0.516    | 0.645    | 0.641    | 0.650 |
| 11      | 0.281    | 0.351    | 0.349   | 0.355  | 0.334    | 0.346    | 0.345    | 0.350 |
| 12      | 0.412    | 0.598    | 0.595   | 0.600  | 0.550    | 0.668    | 0.665    | 0.670 |
| Music background |          |          |         |       |          |          |         |       |       |
| 1       | 0.376    | 0.525    | 0.533   | 0.535  | 0.493    | 0.622    | 0.628    | 0.630 |
| 2       | 0.376    | 0.532    | 0.545   | 0.548  | 0.503    | 0.615    | 0.625    | 0.627 |
| 3       | 0.370    | 0.482    | 0.488   | 0.490  | 0.468    | 0.561    | 0.565    | 0.567 |
| 4       | 0.376    | 0.567    | 0.575   | 0.578  | 0.510    | 0.676    | 0.684    | 0.685 |
| 5       | 0.365    | 0.508    | 0.514   | 0.520  | 0.572    | 0.596    | 0.600    | 0.602 |
| 6       | 0.360    | 0.481    | 0.490   | 0.492  | 0.462    | 0.565    | 0.572    | 0.574 |
| 7       | 0.346    | 0.532    | 0.541   | 0.543  | 0.466    | 0.642    | 0.648    | 0.650 |
| 8       | 0.372    | 0.504    | 0.511   | 0.515  | 0.480    | 0.593    | 0.598    | 0.600 |
| 9       | 0.383    | 0.569    | 0.580   | 0.582  | 0.513    | 0.680    | 0.687    | 0.690 |
| 10      | 0.373    | 0.516    | 0.522   | 0.524  | 0.489    | 0.605    | 0.609    | 0.610 |
| 11      | 0.250    | 0.311    | 0.313   | 0.315  | 0.302    | 0.322    | 0.323    | 0.325 |
| 12      | 0.373    | 0.503    | 0.533   | 0.535  | 0.492    | 0.617    | 0.625    | 0.628 |
| Babble noise |          |          |         |       |          |          |         |       |
| 1       | 0.356    | 0.424    | 0.311   | 0.422  | 0.243    | 0.322    | 0.338    | 0.430 |
| 2       | 0.382    | 0.528    | 0.509   | 0.598  | 0.503    | 0.615    | 0.625    | 0.627 |
| 3       | 0.370    | 0.482    | 0.488   | 0.490  | 0.468    | 0.561    | 0.565    | 0.567 |
| 4       | 0.376    | 0.557    | 0.575   | 0.578  | 0.510    | 0.676    | 0.684    | 0.685 |
| 5       | 0.375    | 0.500    | 0.502   | 0.545  | 0.432    | 0.525    | 0.550    | 0.595 |
| 6       | 0.361    | 0.481    | 0.490   | 0.492  | 0.462    | 0.565    | 0.572    | 0.574 |
| 7       | 0.342    | 0.530    | 0.542   | 0.543  | 0.466    | 0.642    | 0.648    | 0.652 |
| 8       | 0.372    | 0.501    | 0.510   | 0.515  | 0.481    | 0.593    | 0.598    | 0.602 |
| 9       | 0.382    | 0.565    | 0.582   | 0.583  | 0.512    | 0.682    | 0.686    | 0.692 |
| 10      | 0.366    | 0.502    | 0.525   | 0.502  | 0.488    | 0.621    | 0.624    | 0.630 |
| 11      | 0.250    | 0.311    | 0.313   | 0.315  | 0.302    | 0.322    | 0.323    | 0.325 |
| 12      | 0.365    | 0.544    | 0.525   | 0.545  | 0.492    | 0.617    | 0.625    | 0.628 |

SNR: signal-to-noise ratio; DLSE: deep learning-based speech enhancement.

unenhanced conditions there experienced only small decrease in the SII values; but the SII values slightly increase for Wiener filtering with respect to most processing and noise conditions. Here, babble noise also considered with existing two noise environments. Also, for NN_AIM and
NN_COMP, both speech-shaped noise (SSN) conditions are shown (mean values increases in speech intelligibility index (SII) of 0.0157, 0.0131, and 0.0121, respectively, at 0 dB SNR and 0.00807, 0.00424, and 0.00357 at +4 dB SNR). Table 3 presents a performance comparison of different methods for three noise environments with different SNRs on the test set.

Table 4. All methods sound quality and low processing delay performance.

| Measures          | Method |
|-------------------|--------|
|                  | DLSE   | NN_COMP | NN_AIM | WF     |
| Max delay (ms)    | 1.7    | 2       | 3.8 ± 0.3 | 4      |
| PESQ              | 3.9    | 3.2     | 3.6     | 3.2    |
| STOI              | 0.964  | 0.852   | 0.742   | 0.625  |
| User ratings      | 4.1 ± 0.4 | 3.8 ± 0.3 | 3.3 ± 0.3 | 3.5 ± 0.3 |

DLSE: deep learning-based speech enhancement; WF: Wiener filtering; PESQ: perceptual evaluation of speech quality; STOI: short-time objective intelligibility.

Perceptual evaluation of speech quality measure, short-time objective intelligibility measure, and lower processing delay-based evaluation

In the aforementioned sections, we determined the HA system components, individually. To identify the parametric values, this process supports a lot and under practical situations of these components, an acceptable performance is achieved with the identified parameter values. In this section, the performance of the proposed method is evaluated in terms of perceptual evaluation of speech quality (PESQ) measure, short-time objective intelligibility (STOI), and the low processing delay. Based on psychoacoustic principles, after temporal alignments with their respective signal, a test speech signal is analyzed using an objective measure called PESQ.$^{52}$

In the experiment, the above discussed six 3-s long input speech signals were used. To produce 4 dB SNR, the fan noise was added to the speech signals. The traditional WF method, NN_COMP, and NN_AIM were used for analyzing the performance of the proposed DLSE method. For all methods, the subjective evaluation was conducted on the sound quality output in the steady state. In the listening test, six subjects were made into participation for assessing the sound quality. A quiet environment is maintained and a pair of headphones is used for fixing them in both the ears and then the processed sounds were played. Varying the scale level from 1 to 5, each system’s subject is rated; clean speech is listened directly with the scale level 5, and poor quality signal corresponding to the scale level 1 is considered as unacceptable. More importantly, continuous loud artifacts is indicated using the rating 1, continuous low-level artifacts as rating 2, intermittent low-level noticeable artifacts as rating 3, and very good to excellent speech quality is indicated using the rating 4–5. To obtain a straight comparison of “perfect” quality, the processed output and the input could be switched by the subjects. Prior to the start of the experiment, the subjects were placed in the clean sound environment. During rating of the processed sounds, clean audio can be accessed always by the subjects.

Table 4 depicts the processing delays, STOI, and PESQ values of all methods and the user rating on 95% confidence intervals in mean values. User ratings of entire hearing methods and profiles are reported using the PESQ and STOI values. This helps HA algorithms for their parametric selection. When compared to the NN_AIM, NN_COMP, and WF method, the proposed DLSE method has gained better quality. Moreover, proposed DLSE method processing delay is
compared with other methods. It is evident from Table 4 that the proposed method has produced more artifacts than the other methods, such as NN_AIM, NN_COMP, and WF method. Ultimately, the proposed DLSE method has achieved low processing delays and high spectral resolutions than good as other methods.

**Conclusion**

SE problem in CI users was assessed using DCNNs on considering the two important factors namely, speech quality ratings and SE. Using this algorithm, frames or features were formed through decomposing the noisy speech signal. Subsequently, a DCNN is fed with decomposed noisy speech signal frames to produce frequency channel estimation. However, higher SNR information was contained in produced frequency channel estimation. Using this estimate, speech-dominated CI channels are taken to produce electrical stimulation. This process was same as that of the conventional n-of-m CI coding strategies. To determine the speech-in-noise performance of 12 CI users, the fan and music sounds applied are considered as background noises. Performance of the proposed algorithm is evaluated through considering these background noises. Low processing delay and reliable architecture are the best characteristics of DLSE algorithm; hence, this can be suitably applied for all applications of hearing devices. NN_COMP and NN_AIM are the two neural network-based methods used for comparison. Experimental outcomes indicate that the proposed DLSE method has outperformed the existing methods in terms of subjective listening tests and qualitative evaluations. These findings have proved that SE performance of CI users can be significantly improved using this promising DLSE method.

**Declaration of conflicting interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) received no financial support for the research, authorship, and/or publication of this article.

**References**

1. Lai YH and Zheng WZ. Multi-objective learning based speech enhancement method to increase speech quality and intelligibility for hearing aid device users. *Biomed Signal Pr Control* 2019; 48: 35–45.
2. Nossier SA, Rizk MRM, Moussa ND, et al. Enhanced smart hearing aid using deep neural networks. *Alex Eng J* 2019; 58(2): 539–550.
3. Martinez AMC, Gerlach L, Payá-Vayá G, et al. DNN-based performance measures for predicting error rates in automatic speech recognition and optimizing hearing aid parameters. *Speech Commun* 2019; 106: 44–56.
4. Spille C, Ewert SD, Kollmeier B, et al. Predicting speech intelligibility with deep neural networks. *Comput Speech Lang* 2018; 48: 51–66.
5. Chiea RA, Costa MH and Barrault G. New insights on the optimality of parameterized Wiener filters for speech enhancement applications. *Speech Commun* 2019; 109: 46–54.
6. Kolbæk M, Tan ZH and Jensen J. On the relationship between short-time objective intelligibility and short-time spectral-amplitude mean-square error for speech enhancement. *IEEE/ACM T Audio Speech* 2018; 27(2): 283–295.
7. Friesen LM, Shannon RV, Baskent D, et al. Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants. *J Acoust Soc Am* 2001; 110(2): 1150–1163.
8. Fu QJ, Shannon RV and Wang X. Effects of noise and spectral resolution on vowel and consonant recognition: acoustic and electric hearing. *J Acoust Soc Am* 1998; 104(6): 3586–3596.
9. Jin SH, Nie Y and Nelson P. Masking release and modulation interference in cochlear implant and simulation listeners. *Am J Audiol* 2013; 22: 135–146.
10. Tsoukalas DE, Mourjopoulos JN and Kokkinakis G. Speech enhancement based on audible noise suppression. *IEEE T Speech Aud P* 1997; 5(6): 497–514.
11. Xia B and Bao C. Speech enhancement with weighted denoising auto-encoder. In: *Proceedings of the 14th annual conference of the international speech communication association, INTERSPEECH*, Lyon, 25–29 August 2013, pp. 3444–3448.
12. Sang J, Hu H, Zheng C, et al. Speech quality evaluation of a sparse coding shrinkage noise reduction algorithm with normal hearing and hearing impaired listeners. *Hearing Res* 2015; 327: 175–185.
13. Bentler R, Wu YH, Kettel J, et al. Digital noise reduction: outcomes from laboratory and field studies. *Int J Audiol* 2008; 47(8): 447–460.
14. Zakis JA, Hau J and Blamey PJ. Environmental noise reduction configuration: effects on preferences, satisfaction, and speech understanding. *Int J Audiol* 2009; 48: 853–867.
15. Luts H, Eneman K, Wouters J, et al. Multicenter evaluation of signal enhancement algorithms for hearing aids. *J Acoust Soc Am* 2010; 127(3): 1491–1505.
16. Khaleelur Rahiman PF, Jayanthi VS and Jayanthi AN. Deep convolutional neural network-based speech enhancement to improve speech intelligibility and quality for hearing-impaired listeners. *Med Biol Eng Comput* 2019; 57: 757.
17. Goehring T, Bolner F, Monaghan JJM, et al. Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users. *Hearing Res* 2017; 344: 183–194.
18. Healy EW, Yoho SE, Chen J, et al. An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type. *J Acoust Soc Am* 2015; 138(3): 1660–1669.
19. Healy EW, Yoho SE, Wang Y, et al. An algorithm to improve speech recognition in noise for hearing-impaired listeners. *J Acoust Soc Am* 2013; 134(4): 2029–2038.
20. Bolner F, Goehring T, Monaghan JJM, et al. Speech enhancement based on neural networks applied to cochlear implant coding strategies. In: *Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP)*, Shanghai, China, 20–25 March 2016, pp. 6520–6524. New York: IEEE.
21. Acharya UR, Oh SL, Hagiwara Y, et al. A deep convolutional neural network model to classify heartbeats. *Comput Biol Med* 2017; 89: 389–396.
22. Gelderblom FB, Tronstad TV and Viggen EM. Subjective evaluation of a noise-reduced training target for deep neural network-based speech enhancement. *IEEE/ACM T Audio Speech* 2018; 27(3): 583–594.
23. Kavalekalam MS, Nielsen JK, Boldt JB, et al. Model-based speech enhancement for intelligibility improvement in binaural hearing aids. *IEEE/ACM T Audio Speech* 2019; 27(1): 99–113.
24. Hu Y and Loizou PC. Environment-specific noise suppression for improved speech intelligibility by cochlear implant users. *J Acoust Soc Am* 2010; 127(6): 3689–3695.
25. Kim T, Lu Y, Hu Y, et al. An algorithm that improves speech intelligibility in noise for normal-hearing listeners. *J Acoust Soc Am* 2009; 126(3): 1486–1494.
26. Dahl GE, D D, Deng L, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE T Audio Speech* 2011; 20(1): 30–42.
27. Geoffroy H, Li D, Dong Y, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Pr Mag* 2012; 29(6): 82–97.
28. Yang D and Mak CM. An investigation of speech intelligibility for second language students in classrooms. *Appl Acoust* 2018; 134: 54–59.
29. DiLiberto GM, Lalor EC and Millman RE. Causal cortical dynamics of a predictive enhancement of speech intelligibility. *NeuroImage* 2018; 166: 247–258.
30. Kazuhiro K and Taika K. Estimation of binaural speech intelligibility using machine learning. *Appl Acoust* 2018; 129: 408–416.
31. Wang Y and Wang DL. Towards scaling up classification-based speech separation. *IEEE T Audio Speech* 2013; 21(7): 1381–1390.
32. Wang Y, Narayanan A and Wang DL. On training targets for supervised speech separation. *IEEE/ACM T Audio Speech* 2014; 22(12): 1849–1858.
33. Huang PS, Kim M, Hasegawa-Johnson M, et al. Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM T Audio Speech* 2015; 23(12): 2136–2147.
34. Xu Y, Du J, Dai LR, et al. A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM T Audio Speech* 2015; 23(1): 7–19.
35. Sundararaj V. Optimised denoising scheme via opposition-based self-adaptive learning PSO algorithm for wavelet-based ECG signal noise reduction. *Int J Biomed Eng Technol* 2019; 31(4): 325.
36. Vinu S. An efficient threshold prediction scheme for wavelet based ECG signal noise reduction using variable step size firefly algorithm. *Int J Intell Eng Syst* 2016; 9(3): 117–126.
37. Xu Y, Du J, Dai LR, et al. Dynamic noise aware training for speech enhancement based on deep neural networks. In: *Proceedings of the 15th annual conference of the international speech communication association*, 2014, https://pdfs.semanticscholar.org/62e7/ec2f188764b3be45f04a5a5ab9f9d7185f.pdf
38. Xu Y, Du J, Huang Z, et al. Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement, 2017, arXiv:1703.07172.
39. Kim M and Smaragdis P. Adaptive denoising autoencoders: a fine-tuning scheme to learn from test mixtures. In *Proceedings of the international conference on latent variable analysis and signal separation*, Liberec, 25–28 August 2015, pp. 100–107. Cham: Springer.
40. Gao T, Du J, Xu Y, et al. Improving deep neural network based speech enhancement in low SNR environments. In *Proceedings of the international conference on latent variable analysis and signal separation*, Liberec, 25–28 August 2015, pp. 75–82. Cham: Springer.
41. Weninger F, Erdogan H, Watanabe S, et al. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In *Proceedings of the international conference on latent variable analysis and signal separation*, Liberec, 25–28 August 2015, pp. 91–99. Cham: Springer, 2015.
42. Fu SW, Tsao Y and Lu X. SNR-aware convolutional neural network modeling for speech enhancement. In *Proceedings of the Interspeech*, San Francisco, CA, 8–12 September 2016, pp. 3768–3772.
43. Tu Y, Du J, Xu Y, et al. Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers. In *Proceedings of the 9th international symposium on Chinese spoken language processing*, Singapore, 12–14 September 2014, pp. 250–254. New York: IEEE.
44. Schmidhuber J. Deep learning in neural networks: an overview. *Neural Netw* 2015; 61: 85–117.
45. Zhao B, Lu H, Chen S, et al. Convolutional neural networks for time series classification. *J Syst Eng Electron* 2017; 26(1): 162–169.
46. Yann L, Yoshua B and Hinton G. Deep learning. *Nature* 2015; 521: 436–444.
47. He K, Zhang X, Ren S, et al. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, Santiago, Chile, 7–13 December 2015, pp. 1026–1034.
48. Bouvrie J. Notes on convolutional neural networks, 2006, http://cogprints.org/5869/
49. Healy EW, Yoho SE, Wang Y, et al. Speech-cue transmission by an algorithm to increase consonant recognition in noise for hearing-impaired listeners. *J Acoust Soc Am* 2014; 136(6): 3325–3336.
50. Bleeck S, Jones T and Patterson RD. Aim-mat: the auditory image model in MATLAB. *Acta Acust United Ac* 2004; 90(4): 781–787.
51. Jürgen T and Kollmeier B. SNR estimation based on amplitude modulation analysis with applications to noise suppression. *IEEE T Speech Audio Process* 2003; 11(3): 184–192.
52. Rix AW, Beerends JG, Hollier MP, et al. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In: *Proceedings of the 2001 IEEE international conference on acoustics, speech, and signal processing* (Cat. No. 01CH37221), vol. 2, Salt Lake City, UT, 7–11 May 2001, pp. 749–752. New York: IEEE.
53. Loizou PC and Kim G. Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. *IEEE T Audio Speech* 2010; 19(1): 47–56.