Speech Enhancement for Hearing Impaired Based on Bandpass Filters and a Compound Deep Denoising Autoencoder

Raghad Yaseen Lazim AL-Taai 1,2 and Xiaojun Wu 1,2,*

1 School of Computer Science, Shaanxi Normal University, Xi’an 710119, China; raghadyaseen@snnu.edu.cn
2 Key Laboratory of Modern Teaching Technology, Ministry of Education, Shaanxi Normal University, Xi’an 710119, China
* Correspondence: xjwu@snnu.edu.cn

Abstract: Deep neural networks have been applied for speech enhancements efficiently. However, for large variations of speech patterns and noisy environments, an individual neural network with a fixed number of hidden layers causes strong interference, which can lead to a slow learning process, poor generalisation in an unknown signal-to-noise ratio in new inputs, and some residual noise in the enhanced output. In this paper, we present a new approach for the hearing impaired based on combining two stages: (1) a set of bandpass filters that split up the signal into eight separate bands each performing a frequency analysis of the speech signal; (2) multiple deep denoising autoencoder networks, with each working for a small specific enhancement task and learning to handle a subset of the whole training set. To evaluate the performance of the approach, the hearing-aid speech perception index, the hearing aid sound quality index, and the perceptual evaluation of speech quality were used. Improvements in speech quality and intelligibility were evaluated using seven subjects of sensorineural hearing loss audiogram. We compared the performance of the proposed approach with individual denoising autoencoder networks with three and five hidden layers. The experimental results showed that the proposed approach yielded higher quality and was more intelligible compared with three and five layers.

Keywords: compound neural network; deep denoising autoencoder; hearing aid application; bandpass filter; deep learning

1. Introduction

Speech is a fundamental means of human communication. In most noisy situations, the speech signal is mixed with other signals transmitting energy at the same time, which can be noise or even different speech signals. Consequently, it is important to improve speech quality and the intelligibility of degraded speech. Speech enhancement (SE) techniques have been applied to high technological telecommunication systems, e.g., mobile communication or speech recognition, and hearing aids devices. The main aim of SE algorithms is to improve some perceptual aspects of speech that are corrupted by additive background noise. For hearing aids (HA), SE algorithms are used to somehow clean the noisy signal before amplification by reducing the background noise, where hearing-impaired users experience extreme difficulty communicating in environments with varying levels and types of noise (caused by the loss of temporal and spectral resolution in the auditory system of the impaired ear) [1]. In many scenarios, reducing the background noise introduces speech distortion, which reduces speech intelligibility in noisy environments [2]. Since the quality reflects the individual preferences of listeners, it is a subjective performance evaluation metric. The intelligibility is an objective measure since it offers the percentage of words that can be correctly identified by listeners. Based on these two criteria, the considerable challenge in designing an effective SE algorithm for hearing aids is to boost the overall speech quality and to increase intelligibility by suppressing noise without introducing any perceptible distortion in the signal.
In the last few years, many theories and approaches have been developed and proposed for SE [3]. Spectral subtraction (SS) algorithms were proposed in the correlation domain by Weiss et al. [4], and L. Chen et al. in [5] proposed the spectral subtraction approach in modern hearing aids in real-time speech enhancement. The approach is based on a voice activity detector (VAD) to estimate the noise spectrum when speech pauses (in silence), and subtract it from the noisy speech to estimate the clean speech [4,5]. The SS approaches frequently produce a new type of noise occurring at random frequency locations in each frame. This type of noise referred to as musical noise, and it is sometimes more disturbing not only for the human ear but also for SE systems than the original distortions. Harbach in [6] used the Wiener filter method based on prior SNR estimation to improve speech quality by using a directional microphone. However, the all-pole spectrum of the speech signal might have unnaturally sharp peaks, which in turn can result in a significant decline in speech quality. The Minimum Mean Square Error (MMSE) approach based on log-magnitude was suggested in [7,8]. The approach works to find the coefficient by minimises the mean square error (MSE) of the log-magnitude spectra. The test result of this approach showed lower levels of residual noise. Meanwhile, deep neural network (DNN) [9] approaches offered great promise and attention in addressing challenges in SE [10]. For example, L. Ding et al. [11] used a DNN model for speech denoising. The model predicts clean speech spectra when presented with noisy speech inputs, which does not necessitate RBM pre-training or complex recurrent structures. X. Lu et al. [12] presented a regression model of the denoising autoencoder (DAE) to improve the quality of speech. The model maps a noisy input to a clean signal based on the log-power spectra (LPS) feature. The study used different types of noise in the training stage to achieve an excellent ability to generalise to unseen noise environments. The results of objective evaluations showed that the approach performs better than the conventional SE approach. S. Meng et al. in [13] presented a separate deep autoencoder (SDAE) approach, which estimates the noisy and clean spectra by minimising the total reconstruction error [14] of the noisy speech spectrum by adjusting the estimated clean speech signal. Lai et al. in [15,16] suggested using a deep denoising autoencoder networks (DDAE) model in cochlear implant (CI) simulations for improving the intelligibility of vocoded speech. The method simulates speech signal processing in existing CI devices and the actual CI receiver. The study evaluated the results of different clean signals and noise for the training and testing phases, and the LPS feature was used for noise classification.

Previous studies have proven that DNN suppresses noise from noisy corrupted speech efficiently [17]. However, the experimental results of these approaches showed that, for large variations of speech patterns and noisy environments, an individual DNN network with a fixed number of hidden layers cause strong interference effects that lead to (1) a slow learning process, (2) poor generalisation of new inputs in unknown signal-to-noise ratio (SNR) [18], (3) due to DNN manner of converting the speech frame by frame, some residual noise showed in the enhanced output even though context features are used as input to the deep denoising autoencoder, and (4) poor generalisation performance in real-world environments. In this work, we propose a new method for amplification in hearing aids based on combining two stages: (1) a set of bandpass filters, in which each performs a frequency analysis of the speech signal based on the human healthy cochlea, (2) multiple DDAE networks, in which each works for a specific enhancement task and learns to handle a subset of the whole training set. The rest of this work begins in Section 2, where we describe hearing loss and speech perception. Section 3 describes the details of the proposed method. Section 4 presents our experimental setup and evaluation. Finally, Section 5 discusses the experimental results of the work.

2. Speech Perception and Hearing Loss

Speech perception refers to the ability to hear human speech, to interpret it, and to understand it (human frequency range 1–8000 Hz). The speech signal from a single source is mixed with other unwanted signals. The human ear can distinguish between
7000 different frequencies and enables the brain to locate sound sources. However, over 500 million of the population experience hearing loss (HL) [2], and 90% of hearing loss is what is known as sensorineural hearing loss (SNHL), which is caused by dysfunctions in the cochlea (the inner ear). There are tiny fine hairs in the cochlea responsible for sound transmission. The outer hair cells react to high frequencies, and the inner hair cells react to low frequencies. Together, they result in a smooth perception of the full sound range and a good separation of similar sounds. Typically, the high frequencies are affected by hearing loss first, as the respective hair cells are located at the entry of the cochlea, where every sound wave passes by (Figure 1). This usually results in difficulty in hearing and understanding high-frequency sounds such as “s” and “th”, which in normal hearing can distinguish words such as pass and path [1,19]. In the affected areas, the hair cells are no longer stimulated effectively, resulting in not enough impulses being transmitted to the brain for recognition. HA helps hearing loss by maximising a person’s remaining hearing ability by increasing the volume of speech with minimum distortion [16]. However, treating all noise environments the same would result in unsatisfying performance for different hearing impairments [1]. SNHL people have a poorer ability to hear high-frequency components than low-frequency components; consequently, adding a noise frequency classification stage based on the human cochlea is of great importance when designing an SE approach for AH.

Figure 1. The structure for the human ear. (A) The sound waves will be picked up by the outer ear, which is converted to mechanical vibration in the middle ear by the tiny bones. The vibration travels through fluid in the cochlea in the inner ear. (B) Frequency analysis in the cochlea. (I) Frequency map of the cochlea human hearing organ. (II) The basilar membrane (BM) presents the first level of frequency analysis in the cochlea because of its changing stiffness and nearly constant unit mass from (20 kHz) to the apex (20 Hz). High-frequency sounds cause more significant deflection of the basilar membrane where it is narrow and stiff, and lower frequency sounds produce more significant deflection where the basilar membrane is loose and flexible. Trends in Neurosciences, April 2018 [3].

3. Architecture of the Proposed System

This section presents the details about the approach of the two stages, which we call (HC-DDAEs). The architecture of the proposed method is presented in (Figure 2).

3.1. Bandpass Filter

A bandpass filter (BPF) passes only a certain range of frequencies without reduction. The particular band of frequency that passes by the filter is known as the passband. During this stage (Figure 3), the input signal $y$ is first pre-emphasised and then passed through a set of eight BPFs. The pre-emphasis is performed by finite impulse response (FIR) filters, which
emphasise signals in eight passbands of 40 dB, and the forward and reverse coefficients \((b_p, a_p)\) are given by the following:

\[
\begin{align*}
    b_p &= [1 - \exp(-1200 \times 2 \times \pi / F_s)] \\
    a_p &= [1 - \exp(-3000 \times 2 \times \pi / F_s)]
\end{align*}
\]  

(1)

where \(F_s\) denotes the sampling frequency.

---

**Figure 2.** The block diagram of the proposed approach (HC-DDAES) consists of: (1) a set of bandpass filters (BPF), and (2) a compound model of multiple of DDAEs networks (C-DDAEs). The C-DDAE model consists of (A) training stage and (B) testing stages.

**Figure 3.** Block diagram of bandpass filters. The incoming speech signal is decomposed into multiple frequency bands by N BPF.

Each of these passbands passes only a specific frequency band \([f_1(i) f_2(i)]\) of the entire input \([f_{low} f_{high}]\) (Table 1). where \(f_1(i)\) and \(f_2(i)\) denote the lower and upper cut-
off frequencies for the \(i\)-th BPF (Figure 4), and \(f_{\text{low}}\) and \(f_{\text{high}}\) are the lowest and highest frequencies of \(n\) passbands for the \(i\)-th BPF, which is specified as follows:

\[
f_1(i) = \left( f_{\text{low}} + 10^{C_b \times (i-1)} \right) / F_s \tag{2}
\]

\[
f_2(i) = \left( f_{\text{high}} + 10^{C_b \times (i-1)} \right) / F_s \tag{3}
\]

where \(C_b\) is the channel bandwidth, which is given by the following:

\[
C_b = \left( \log_{10} \left( f_{\text{high}} / f_{\text{low}} \right) \right) / n \tag{4}
\]

Table 1. The frequency range of each frequency band.

| Channel No. | Lower Cutoff Frequency (Hz) \(f_1(i)\) | Higher Cutoff Frequency (Hz) \(f_2(i)\) |
|-------------|--------------------------------|-----------------------------------|
| 1           | 20                              | 308                               |
| 2           | 308                             | 662                               |
| 3           | 662                             | 1157                              |
| 4           | 1157                            | 1832                              |
| 5           | 1832                            | 3321                              |
| 6           | 3321                            | 4772                              |
| 7           | 4772                            | 6741                              |
| 8           | 6741                            | 8000                              |

Figure 4. Simulation results of the speech signal in \(i\)-th BPF.

The gain in each filter could be set individually to any value between zero and one, where one corresponds to the gain provided by the original hearing aid. The gain adjustments are made on a computer keyboard and are shown graphically on the computer’s display.

The \(n\) channels in the output of the filter are then added together to produce the synthesised speech waveform using the periodicity detection method as follows:

\[
x_2 = IDFT \left( \left| DFT \left( x_{\text{low}} \right) \right|^k \right) + IDFT \left( \left| DFT \left( x_{\text{high}} \right) \right|^k \right) \tag{5}
\]
where $x_{low}$ and $x_{high}$ are the two channels—channel one (0–4 kHz) and channel two (4–8 kHz), respectively.

### 3.2. Compound DDAEs (C-DDAEs)

The output of the first stage passes through the compound approach. This approach utilises multiple networks. Each network is a multi-layer DDAE (three hidden layers for each DDAE) and has different hidden units as follows:

1. DDAE-1: 128 units for each layer, $F_{DDAE1_{128}}$. The magnitude spectrum is 513-dimensional, which works as the input and target.
2. DDAE-2: 512 units for each layer, $F_{DDAE2_{512}}$. Three frames of spectra used: $\begin{bmatrix} x_{t-1}^T, x_{t}^T, x_{t+1}^T \end{bmatrix}^T$, where the target is the single spectrum $|s_t|$.
3. DDAE3: Three hidden layers with 1024 units for each layer, $F_{DDAE3_{1024}}$. Five frames of spectra are used: $\begin{bmatrix} x_{t-2}^T, x_{t-1}^T, x_{t}^T, x_{t+1}^T, x_{t+2}^T \end{bmatrix}^T$.

Each DDAE is specific for one enhancement task rather than one individual network being used for the general enhancement task. The total output of every network is the central frame of the model (Figure 5). This stage includes two phases, namely, training and testing. In the training phase, the training set is divided into subsets, and for each DDAE, there is a specific enhancement task to learn—a subset of the whole training set. Let us consider the statistical problem where the input speech signal in the spectral domain is given by the following.

$$y(t) = x(t) + n(t)$$

where $y(t)$ and $x(t)$ are the noisy and clean versions of the signal, respectively. $n(t)$ is unwanted noise at $t - lh$ time index ($l = 0, 1, \ldots, T - 1$), which assumed to be zero-mean random white noise or coloured noise that is uncorrelated with $x(t)$.

![Figure 5](image_url)

**Figure 5.** The structure of C-DDAEs. Each DDAE is a feedforward network. The output of the last DDAE network is considered as the desired signal.

Firstly, the input vector $y$ passes through feature extraction to produce the featured version $\tilde{y} \in \mathbb{R}^K$. Then, training pairs $(\tilde{y}, \tilde{x})$ are prepared by calculating the magnitude ratio $\tilde{x} = \frac{|x|}{|x+n|}$. Next, $F_{DDAE}$ (a function) is mapped to produce an output vector $\hat{x} \in \mathbb{R}^D$ to recover the clean original signal $x(t)$:

$$\hat{x} = F_{DDAE}(\tilde{x}) = f_{DDAE_i}(Wx + b),$$

where $f_{DDAE_i}$ is a linear function.
where $W$ is a $d \times d'$ $\in \mathbb{R}^{n \times m}$ weight matrix, $b \in \mathbb{R}^m$ is a bias vector, and $f_{DDAE_i}(\cdot)$ is a nonlinear function. If the output for $DDAE_i$ is incorrect, the weight on the model changes and updated and more tasks will sign to the next DDAE network (Figure 5), where the feedforward procedure is as follows:

$$y = F_{DDAE}(\bar{y})$$  \hspace{1cm} (9)

The hidden representation of $h(Y) \in \mathbb{R}^m$ is as follows:

$$h^i(Y^E_n) = \sigma(W^i Y^E_n + b^i)$$  \hspace{1cm} (10)

where $\sigma$ is the logistic sigmoid function and $h$ is a hidden layer. $\hat{X}^E_i$ is the vector containing the logarithmic amplitudes of enhanced speech corresponding to the noisy counterpart $Y^E_n$.

The error of the network is the expected value of the squared difference between the target and the estimated vector at $t$-th layer, which participates in the feedforward procedure as follows:

$$E^c = \langle \| d^c - O^c_{DDAE} \|^2 \rangle = \sum_{DDAE_i} p^c_{DDAE_i} \| d^c - O^c_{DDAE_i} \|^2$$  \hspace{1cm} (11)

where $O^c_{DDAE_i}$ is the output vector of expert $DDAE_i$ for case $c$, $p^c_{DDAE_i} = \exp(o_{DDAE_i}) / \sum_{DDAE_i} \exp(o_{DDAE_i})$ is the proportional contribution of $DDAE_i$ to the combined output vector, and $d^c$ is the desired output vector in case $c$.

If the error of each DDAE network is less than the weighted average of the errors of the C-DDAEs, the responsibility for that case is increased, and vice versa. To take into account how well each DDAE does in comparison with other DDAEs, the error function is as follows:

$$E^c = -\log \sum_{DDAE_i} p^c_{DDAE_i} c \| d^c - O^c_{DDAE_i} \|^2$$  \hspace{1cm} (12)

To compare the error function to the output of a DDAE network, we obtain the following:

$$\frac{\partial E^c}{\partial O^c_{DDAE_i}} = -\left[ \frac{p^c_{DDAE_i} c^{-\frac{1}{2}} \| d^c - O^c_{DDAE_i} \|^2}{\sum_{DDAE_i} c^{-\frac{1}{2}} \| d^c - O^c_{DDAE_i} \|^2} \right] (d^c - O^c_{DDAE_i})$$  \hspace{1cm} (13)

To minimise the sum of errors between the target and the estimated vector at $L$ hidden layer, we obtain the following:

$$\theta^* = \arg \min_{W^{(l)}_{\cdot}, \ldots, W^{(l+1)}_{\cdot}} \sum_{l} \mathcal{E}(\tilde{x}_l \| F_{DDAE}(\bar{y}_l) \|)$$  \hspace{1cm} (14)

where $W^{(1)} \in \mathbb{R}^{K(l+1) \times (K(l)+1)}$ holds the network parameters at $l$-th layer, which participates in the feedforward procedure as follows:

$$F_{DDAE}(\bar{y}) = z^{(l+2)}, z^{(1)} = \bar{y},$$

$$z^{(l+1)} = g^{(l)} \left( W^{(l)}, \left[ \left( z^{(l)} \right)^T, 1 \right]^T \right)$$  \hspace{1cm} (15)

Note that $z^l \in \mathbb{R}^{K(l)}$ is a vector of $K(l)$ hidden unit outputs.
4. Experiments and Evaluation

This section describes our experimental setup: (A) the data used to train and test the proposed system, and (B) the comparison of spectrograms of noisy and enhanced signals.

4.1. Experimental Setup

We conducted our experiment using CMU_ARCTIC databases from Carnegie Mellon University, which include 1186 clean speech utterances spoken by US native English-speaking male (DbI) and female (SlI) speakers [20]. The database was recorded at 16-bit, 32 kHz, in a soundproof room. In this work, we divided the database into (1) 75% of the entire dataset for the training set (about 890 speech signals); (2) 20% for the validation set (about 237 speech signals), and (3) the rest (5%) of the entire dataset for the testing set (about 56 utterances). Gaussian white and pink noise were generated and added to the training set at four different SNR levels (0, 5, 10, and 15 dB). The SNR levels were selected carefully to cover a range of noise levels (from light to high) for each noise type. Additionally, noise from a train and babbling were added to the testing set, neither of which were used in the training set. The overlapping frames had a 16 ms duration with a shift of 16 ms. The C-DDAEs model was trained offline and tested online, and layer by layer pre-training was used [21] with 20 epochs. The number of epochs for fine-tuning was 40. The activation function was sigmoid. Each windowed speech segment was addressed with a 256-point FFT and then converted to an MFCC feature vector [22]. For comparison’s sake, we trained two individual DDAE networks with three and five hidden layers each. A MATLAB R2019b-based simulator was used to implement the processing for impaired subjects, which is run and tested in windows 10pro (Intel(R) Core (TM) i5-7200U CPU, 2.71 GHz with 8 GB memory). The objective used in this study was to reduce the symptoms of HFHL for subjects 1–7 (Table 2) to test the performance of the approach.

Table 2. Audiograms of the seven sensorineural hearing losses subjects.

| Audiogram                              | Frequency (kHz) in dB H.L. |
|----------------------------------------|-----------------------------|
|                                        | 0.25 | 0.5  | 1    | 2    | 4    | 8    |
| 1 Plane loss                           | 60   | 60   | 60   | 60   | 60   | 60   |
| 2 Reverse tilt loss                    | 70   | 70   | 70   | 50   | 10   | 10   |
| 3 Moderate tilt high-frequency loss    | 40   | 40   | 50   | 60   | 65   | 65   |
| 4 Steep slope high-frequency loss with standard low-frequency threshold | 0    | 0    | 0    | 60   | 80   | 90   |
| 5 Steep slope high-frequency loss with mild low-frequency hearing loss | 0    | 15   | 30   | 60   | 80   | 85   |
| 6 Mild to moderate tilt high-frequency hearing loss | 14   | 14   | 11   | 14   | 24   | 39   |
| 7                                        | 24   | 24   | 25   | 31   | 46   | 60   |

4.2. The Spectrograms Comparison

A spectrogram is a standard tool for analysing the time-varying spectral characteristics of speech signals [23]. Figure 6 shows five spectrograms: (a) a clean signal (extracted from a male voice saying, “God bless him, I hope I will go on seeing them forever.”); (b) the noisy signal at the 0 dB SNR level; (c) the signal enhanced by DDAE-3; (d) the signal enhanced by DDAE-5; and (e) the signal enhanced by HC-DDAEs. Figure 7 presents five spectrograms: (a) a clean signal (extracted from man’s voice saying, “You yellow giant thing of the frost.”); (b) the noisy signal at 10 dB; (c) the signal enhanced by DDAE-3; (d) the signal enhanced by DDAE-5; and (e) the signal enhanced by HC-DDAEs. The x-axis and y-axis of the spectrum represent the time index and the frequency, respectively. The proposed approach provided a more similar spectrogram to that of the clean speech spectrogram than the traditional DDAE-3 and DDAE-5 under the same test conditions.
Figure 6. The spectrograms result for the utterance of “arctic_a0006”. The clean utterance is for a man’s voice saying: “God bless him, I hope I will go on seeing them forever”.

Figure 7. The spectrograms result for the utterance of “arctic_b0493”. The clean utterance is for a man’s voice saying: “Your yellow giant thing of the frost”.
5. Speech Quality and Intelligibility Evaluation

This section describes the evaluation of our method that presented in (Figure 8). We used well-known metrics to evaluate the quality and intelligibility of the enhanced speech (higher scores represent better quality). We compared the enhanced signals to the desired ones of the test signals.

![Diagram](https://example.com/diagram.png)  
**Figure 8.** The evaluation measures of speech quality and intelligibility used in this study.

5.1. Speech Quality Perception Evaluation (PESQ)

The PESQ standard [23] was defined as in the ITU-T P.862 recommendation: it uses an auditory model that includes auditory filter, spectrum, and time masking. The PESQ tests the quality of the speech by comparing the enhanced signal with the clean signal. It does so by predicting the quality with a good correlation in an extensive range of conditions, which may contain distortions, noise, filtering, and errors. PESQ is a weighted sum of the average disturbance $d_{sym}$ and the average asymmetrical disturbance $d_{sym}$, which can be defined as follows:

$$ PESQ = a_0 + a_1 d_{sym} + a_2 d_{sym} $$  

where $a_0 = 4.5$, $a_1 = -0.1$, and $a_2 = -0.0309$. It produces a score from 0.5 to 4.5, with high values indicating better speech quality.

5.2. Hearing Aid Speech Quality Index (HASQI)

The HASQI is used to predict speech quality according to the hearing threshold of the hearing-impaired individual. The HASQI is based on two independent parallel pathways: (1) $Q_{nonlin}$, which catches the noise and nonlinear distortion effects, and (2) $Q_{lin}$, which catches the linear filtering and spectral changes by targeting differences in the long-term average spectrum [24]. $Q_{nonlin}$ and $Q_{lin}$ are calculated from the output of the auditory model to quantify specific changes in the clean reference signal and the enhanced signal by the following:

$$ Q = 0.336Q_{nonlin} + 0.501Q_{nonlin}^2 + 0.001Q_{lin} + 0.16Q_{lin}^2 $$

5.3. Hearing Aid Speech Perception Index (HASPI)

The HASPI predicts speech intelligibility based on an auditory model that incorporates changes due to hearing loss. The index first collects all aspects of normal and impaired auditory functions [25]. Then, it compares the correlation values ($c$) of the outputs of the auditory model for a reference signal to the outputs of the degraded signals from tests over time. The generation of the unprocessed reference input signal is as follows:

$$ c = \frac{1}{5} \sum_{j=2}^{6} r(j) $$
where \( j \) and \( r(j) \) denote the basis function number and the normalised correlation, respectively. The reference signal of the auditory model is set for normal hearing, and the test signal of the model incorporates hearing loss.

The auditory model is used to measure \( c \) of the high-level part (expressed as \( a_{\text{high}} \)) of the clean signal and the enhanced signal in each frequency band. The envelope is sensitive to the dynamic signal behaviour related to consonants, and the cross-correlation tends to retain the harmonics in stable vowels. Finally, the HASPI score is calculated according to \( c \) and \( a_{\text{high}} \). Let \( a_{\text{Low}} \) be the low-level auditory coherence value, \( a_{\text{Mid}} \) be the mid-level value and \( a_{\text{High}} \) be the high-level value. Then, HASPI intelligibility is given by the following:

\[
P = -9.047 + 14.817c + 0.0a_{\text{Low}} + 0.0a_{\text{Mid}} + 4.616a_{\text{High}}
\]  

(19)

More details of the HASPI auditory model can be found in [25].

6. Results and Discussion

In this section, we present the average scores of objective measurements of the test set for the proposed approach. Our method is compared with individual multi-layer state-of-the-art DDAE approaches (with three (DDAE-3) and five hidden (DDAE-5) layers) that are based on the study in [1] and [20]. PESQ measures for different types of noise (white, pink, babble, and train noise) and SNR conditions (0, 5, 10, and 15 dB). The results of the PESQ evaluation are listed in (Table 3) and show the average PESQ scores of noisy, DDAE-3, DDAE-5, and HC-DDAES for white, pink, babble, and train noises at four different SNR levels (0, 5, 10, and 15 dB). The difference between loudness spectra is computed and an average quality score over time and frequency to produce the prediction of subjective Mean Opinion Score (MOS). PESQ score range is from \(-0.5\) to \(4.5\) (a higher score present better speech quality). The experimental results of the listening test showed that (1) the proposed HC-DDAES approach significantly enhanced the speech quality for most types of noises than the traditional DDAE with three and five hidden layers in all the test conditions. (2) the result of the babble noise in 15 dB SNR, the DDAE-5 achieved the almost same result as the Proposed HC-DDAES. (3) Individual DDAE network with three hidden layers unable to obliterate the noise due to local minima especially in high SNR levels, and increasing the number of epochs in the training stage helps but time-consuming. (4) Increasing the hidden layers on individual DDAE to five (DDAE-5) gave better speech quality than the DDAE network with three hidden layers (DDAE-3) at most SNR levels. From these results, we found that HC-DDAES provide better speech quality for hearing loss users.

Figure 9 introduces the average score of the HASQI for (noisy, DDAE-3, DDAE-5, and HC-DDAES) at four different SNR levels (0, 5, 10, and 15 dB) for seven hearing loss audiograms (1–7) over babble noise. HASQI score range is from 0 to 0.5 (A higher score presents better speech quality). The results clearly showed that: (1) HC-DDAES provided higher speech quality than DDAE-3 and DDAE-5 in all test cases for the seven audiograms, especially in low SNR levels. HC-DDAES performance is slightly degraded in audiogram seven at 15 dB SNR. (2) DDAE-3 performance is degraded in audiogram two and three at 10 dB SNR level and in audiogram four at 10–15 dB SNR. (3) Increasing the hidden layers on individual DDAE to five (DDAE-5) provided better speech quality than DDAE-3 in almost all hearing cases and slightly same speech quality with HC-DDAES of audiogram four and audiogram seven at all SNR levels especially in SNR (5 and 10 dB).

The results of the average HASPI score are presented in Figure 10 for seven audiograms (1–7) at four different SNR levels (0, 5, 10, and 15 dB) over babble noise. HASPI score range is from 0 to 1 (A higher score present better speech quality). The experimental results show that: (1) the proposed HC-DDAES achieved a higher score for HASPI than DDAE-3 in all test conditions, gave slightly same results with DDAE-5 in audiograms 2, 3, and 5 at 0 dB SNR level, and provided better intelligibility in the rest of the test cases. (2) DDAE-3 performance is degraded in audiogram 5 at 10 dB SNR and audiogram seven at 15 dB SNR level. (3) Increasing the hidden layers on individual DDAE to five (DDAE-5) gives a slightly better HASPI score than DDAE-3 in most listening cases, while provided the
same or almost the same speech intelligibility with DDAE-3 in audiograms from 1–5 at 15 dB SNR.

Figure 9. Average results of HASQI metrics of seven audiograms of sensorineural hearing loss. The vertical axis presents the HASQI score (between 0–0.5), and the horizontal axis presents the SNR levels.

The experimental results show that: (1) the proposed HC-DDAEs achieved a higher score for...
HASPI than DDAE-3 in all test conditions, gave slightly same results with DDAE-5 in audiograms 2, 3, and 5 at 0 dB SNR level, and provided better intelligibility in the rest of the test cases. (2) DDAE-3 performance is degraded in audiogram 5 at 10 dB SNR and audiogram seven at 15 dB SNR level. (3) Increasing the hidden layers on individual DDAE to five (DDAE-5) gives a slightly better HASPI score than DDAE-3 in most listening cases, while provided the same or almost the same speech intelligibility with DDAE-3 in audiograms from 1–5 at 15 dB SNR.

**Figure 10.** Average results of HASPI metrics of seven audiograms of sensorineural hearing loss. The vertical axis presents the HASPI score (between 0–1), and the horizontal axis presents the SNR levels.
Table 3. PESQ results (average measures for the test set). A higher score indicating better speech quality.

| Noise   | Method       | SNR Level |
|---------|--------------|-----------|
|         |              | 0 dB | 5 dB | 10 dB | 15 dB |
| White   | Noisy        | 1.49 | 1.62 | 1.91  | 2.12  |
|         | DDAE-3       | 2.03 | 2.09 | 2.18  | 2.34  |
|         | DDAE-5       | 2.3  | 2.16 | 2.38  | 2.5   |
|         | HC-DDAEs     | 2.71 | 2.78 | 2.83  | 2.88  |
| Pink    | Noisy        | 1.54 | 1.71 | 1.88  | 2.12  |
|         | DDAE-3       | 2.03 | 2.11 | 2.19  | 2.41  |
|         | DDAE-5       | 2.11 | 2.44 | 2.51  | 2.72  |
|         | HC-DDAEs     | 2.22 | 2.76 | 2.76  | 2.98  |
| Babble  | Noisy        | 1.49 | 1.72 | 1.8   | 1.91  |
|         | DDAE-3       | 2.01 | 2.09 | 2.33  | 2.29  |
|         | DDAE-5       | 2.04 | 2.11 | 2.47  | 2.66  |
|         | HC-DDAEs     | 2.24 | 2.29 | 2.53  | 2.67  |
| Train   | Noisy        | 1.51 | 1.54 | 1.76  | 1.91  |
|         | DDAE-3       | 2.15 | 2.27 | 2.11  | 2.11  |
|         | DDAE-5       | 2.18 | 2.26 | 2.26  | 2.46  |
|         | HC-DDAEs     | 2.27 | 2.41 | 2.48  | 2.69  |

Based on the previous results we found that the proposed HC-DDAEs improved the quality and intelligibility of the speech. However, the benefits of the proposed HC-DDAEs approach for hearing losses seem to have degraded audiograms three, five and seven in high SNR levels (Figures 9 and 10). In other words, HC-DDAEs still have room for improvement to enhance speech for hearing-impaired users in real-world noisy environments with high SNR levels. Additionally, the study was a computer-based simulation and not tested on real hearing devices, which limited the benefit of the approach. More hearing loss information such as the disturbance level and hearing threshold must be considered for future work to improve the performance of the proposed approach.

7. Conclusions

In this work, we investigated the performance of the deep denoising autoencoder for speech enhancement. We proposed a new denoising approach for improving the quality and the intelligibility of speech for hearing applications based on bandpass filters and a compound multiple deep denoising autoencoder networks. In this first stage, a set of bandpass filters splits up the speech signal into eight separate channels between 0 and 8000 Hz based on the cochlea’s frequency responses. Then, a compound model composed of multiple DDAE networks was proposed, each network of which is specialised for a specific enhancement subtask of the whole enhancement task. To monitor the improvement of speech quality in unsuitable conditions, different speech utterances and noises were used for training and testing the model. We evaluated the speech intelligibility and quality of the proposed model based on PESQ, HASQI and HASPI, which were applied to seven HFHL audiograms. Then, we compared the results of the proposed approach to results obtained by DDAE networks with three and five hidden layers separately. Based on the experimental results in this study, we concluded that: (1) The proposed HC-DDAEs approach provided better speech quality based on PESQ and HASQI in most test conditions than DDAE-3 and DDAE-5. (2) The enhanced speech based on the DDAE-3 network was unable to totally obliterate the noise due to local minima especially in high SNR levels, and increasing the number of epochs in the training stage helps but is time-consuming. (3) Increasing the hidden layers on individual DDAE to five (DDAE-5) gave better speech quality than the DDAE network with three hidden layers (DDAE-3) at most SNR levels. Meanwhile, the proposed HC-DDAEs provided low speech intelligibility based on HASPI for the audiograms 1–7 in 0 SNR level. Several issues must be further investigated for future work. First, future real-world noises in different SNR levels will be used to examine...
the proposed approach and compare it with further state-of-the-art studies. Additionally, the study was a computer-based experiment and will be tested on real hearing aid devices for future work.

**Author Contributions:** R.Y.L.A.-T. developed the algorithm, conducted the experiments and participated in writing the paper and X.W. supervised the overall work and reviewed the paper. Both authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** [http://www.festvox.org/index.html](http://www.festvox.org/index.html).

**Acknowledgments:** This work was partially supported by the National Key Research and Development Program of China (No. 2017YFB1402102), the National Natural Science Foundation of China (No. 11772178, No. 11502133, No. 11872036, No. 61701291), the Fundamental Research Fund for the Central Universities (No. GK201801004), the China Postdoctoral Science Foundation-funded project (No. 2017M613053) and the Shaanxi Natural Science Foundation Project under the grant (No. 2018Q6089).

**Conflicts of Interest:** We wish to confirm that there are no known conflicts of interest associated with this publication and there was no significant financial support for this work that could have influenced its outcome. We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship that are not listed. We further confirm that the order of authors listed in the manuscript was approved by all of us.

**References**

1. Ying, L.H.; Wei, Z.; Shih, T.; Shih, H.; Wen, H.L.; Yu, T. Improving the performance of hearing aids in noisy environments based on deep learning technology. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 18–21 July 2018; Volume 20, pp. 404–408.

2. WHO. Deafness and Hearing Loss. Available online: [http://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss](http://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss) (accessed on 5 January 2021).

3. Lesica, N. Why Do Hearing Aids Fail to Restore Normal Auditory Perception. *Trends Neurosci.* 2018, 41, 174–185. [CrossRef] [PubMed]

4. Weiss, M.; Aschkenasy, E.; Parsons, T. *Study and Development of the INTEL Technique for Improving Speech Intelligibility; Technical Report NSC-FR/4023; Nicolet Scientific Corporation*: Northvale, NJ, USA, 1974.

5. Chen, L.; Wang, Y.; Yoho, S.E.; Wang, D.; Healy, E.W. Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises. *J. Acoust. Soc. Am.* 2016, 139, 2604–2612. [CrossRef] [PubMed]

6. Harbach, A.A.; Arora, K.; Mauger, S.J.; Dawson, P.W. Combining directional microphone and single-channel noise reduction algorithms: A clinical evaluation in difficult listening conditions with cochlear implant users. *Ear Hear.* 2012, 33, 13–23.

7. Hu, Y.; Loizou, P. Subjective Comparison of Speech Enhancement Algorithms. In Proceedings of the 2006 IEEE International Conference on Acoustics, Speech, and Signal Processing, Toulouse, France, 14–19 May 2006; Volume 1, pp. 153–156.

8. Gray, R.; Buzo, A.; Gray, A.; Matsuyama, Y. Distortion measures for speech processing. *IEEE Trans. Acoust. Speech Signal Process.* 1980, 28, 367–376. [CrossRef]

9. Aubreville, M.; Ehrensperger, K.; Maier, A.; Rosenkranz, T.; Graf, B.; Puder, H. Deep Denoising for Hearing Aid Applications. Available online: [http://arxiv.org/abs/1805.01198](http://arxiv.org/abs/1805.01198) (accessed on 5 January 2021).

10. Chen, F.; Loizou, P.C. Impact of SNR and gain-function over- and under-estimation on speech intelligibility. *Speech Commun.* 2012, 54, 81–272. [CrossRef] [PubMed]

11. Liu, D.; Smaragdis, P.; Kim, M. Experiments on Deep Learning for Speech Denoising. In Proceedings of the 15th Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014.

12. Lu, X.; Tsao, Y.; Matsuda, S.; Hori, C. Ensemble modelling of denoising autoencoder for speech spectrum restoration. In Proceedings of the 15th Conference in the annual series of Interspeech, Singapore, 14–18 September 2014; pp. 885–889.

13. Sun, M.; Zhang, X.; Hamme, H.V.; Zheng, T.F. Unseen noise estimation using separable deep autoencoder for speech enhancement. *IEEE Trans. Audio Speech Lang. Process.* 2016, 24, 93–104. [CrossRef]

14. Shifás, M.; Claudio, S.; Stylianos, Y. A fully recurrent feature extraction for single-channel speech enhancement. *arXiv* 2020, arXiv:2006.05233.
15. Lai, Y.-H.; Chen, F.; Wang, S.-S.; Lu, X.; Tsao, Y.; Lee, C.-H. A Deep Denoising Autoencoder Approach to Improving the Intelligibility of Vocoded Speech in Cochlear Implant Simulation. IEEE Trans. Biomed. Eng. 2017, 64, 1568–1578. [CrossRef] [PubMed]

16. Lai, Y.-H.; Tsao, Y.; Lu, X.; Chen, F.; Su, Y.-T.; Chen, K.-C.; Chen, Y.-H.; Chen, L.-C.; Li, L.P.-H.; Lee, C.-H. Deep Learning–Based Noise Reduction Approach to Improve Speech Intelligibility for Cochlear Implant Recipients. Ear Hear. 2018, 39, 795–809. [CrossRef] [PubMed]

17. Lai, Y.-H.; Zheng, W.-Z. Multi-objective learning based speech enhancement method to increase speech quality and intelligibility for hearing aid device users. Biomed. Signal Process. Control. 2019, 48, 35–45. [CrossRef]

18. Kim, M. Collaborative Deep Learning for Speech Enhancement: A Run-time Model Selection Method Using Autoencoders. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017.

19. Souza, P. Speech Perception and Hearing Aids. In Hearing Aids; Springer Handbook of Auditory Research, Chapter 6; Springer: Berlin/Heidelberg, Germany, 2016; pp. 151–180.

20. CMU_ARCTIC Speech Synthesis Databases. Available online: http://www.festvox.org/cmu_arctic/ (accessed on 5 January 2021).

21. Seyyedsalehi, S.; Seyyedsalehi, S. A fast and efficient pre-training method based on layer-by-layer maximum discrimination for deep neural networks. Neurocomputing 2015, 168, 669–680. [CrossRef]

22. Tsao, Y.; Lai, Y.H. Generalized maximum a posteriori spectral amplitude estimation for speech enhancement. Speech Commun. 2016, 76, 112–126. [CrossRef]

23. Beerends, J.; Hoekstra, A.; Rix, A.W.; Hollier, M.P. Perceptual evaluation of speech quality (PESQ) the new ITU standard for end-to-end speech quality assessment part 2: Psychoacoustic model. J. Audio Eng. Soc. 2002, 50, 765–778.

24. Kates, S.J.; Arehart, K. The hearing-aid speech quality index (HASQI). J. Audio Eng. Soc. 2010, 58, 363–381.

25. Kates, J.M.; Arehart, K.H. The hearing-aid speech perception index (HASPI). Speech Commun. 2014, 65, 75–93. [CrossRef]