Deep Learning–Based Noise Reduction Approach to Improve Speech Intelligibility for Cochlear Implant Recipients

Ying-Hui Lai,[1] Yu Tsao,[2] Xugang Lu,[3] Fei Chen,[4] Yu-Ting Su,[5] Kuang-Chao Chen,[6,7] Yu-Hsuan Chen,[8] Li-Ching Chen,[7] Lieber Po-Hung Li,[7,9] Chin-Hui Lee[10]

Objective: We investigate the clinical effectiveness of a novel deep learning–based noise reduction (NR) approach under noisy conditions with challenging noise types at low signal to noise ratio (SNR) levels for Mandarin-speaking cochlear implant (CI) recipients.

Design: The deep learning–based NR approach used in this study consists of two modules: noise classifier (NC) and deep denoising autoencoder (DDAE), thus termed (NC + DDAE). In a series of comprehensive experiments, we conduct qualitative and quantitative analyses on the NC module and the overall NC + DDAE approach. Moreover, we evaluate the speech recognition performance of the NC + DDAE NR and classical single-microphone NR approaches for Mandarin-speaking CI recipients under different noisy conditions. The testing set contains Mandarin sentences corrupted by two types of maskers, two-talker babble noise, and a construction jackhammer noise, at 0 and 5 dB SNR levels. Two conventional NR techniques and the proposed deep learning–based approach are used to process the noisy utterances. We qualitatively compare the NR approaches by the amplitude envelope and spectrogram plots of the processed utterances. Quantitative objective measures include (1) normalized covariance measure to test the intelligibility of the utterances processed by each of the NR approaches; and (2) speech recognition tests conducted by nine Mandarin-speaking CI recipients. These nine CI recipients use their own clinical speech processors during testing.

Results: The experimental results of objective evaluation and listening test indicate that under challenging listening conditions, the proposed NC + DDAE NR approach yields higher intelligibility scores than the two compared classical NR techniques, under both matched and mismatched training-testing conditions.

Conclusions: When compared to the two well-known conventional NR techniques under challenging listening condition, the proposed NC + DDAE NR approach has superior noise suppression capabilities and gives less distortion for the key speech envelope information, thus, improving speech recognition more effectively for Mandarin CI recipients. The results suggest that the proposed deep learning–based NR approach can potentially be integrated into existing CI signal processors to overcome the degradation of speech perception caused by noise.

Key words: Cochlear implant, Deep denoising autoencoder, Deep learning, Noise reduction

(Ear & Hearing 2017;XX;00–00)

INTRODUCTION

A cochlear implant (CI) is an implantable electronic device that replaces the function of potentially damaged parts of the inner ear (Loizou 1999). According to a report by the Food and Drug Administration, by the end of 2012, over 324,000 people worldwide would have received a CI (NIDCD 2014). The tremendous progress of CI technologies over the past 3 decades has enabled many CI recipients to enjoy a high level of speech perception in quiet conditions. However, speech intelligibility in noisy conditions still remains a challenge (Friesen et al. 2001; Fetterman & Domico 2002; Nie et al. 2005; Loizou 2006; Chen et al. 2015).

Numerous researchers have focused on the development of effective noise reduction (NR) techniques that either preprocess noisy speech and feed the enhanced signal as an input to the processors of CI devices or somehow suppress the noise presented in the noisy envelopes (Goehring et al. 2017). These NR approaches can be broadly divided into multiple- and single-microphone frameworks. The benefit of multi-microphone NR becomes apparent when the target and noise are spatially separated (Schmidt 1986). Hamacher et al. (1997) demonstrated that the benefits obtained with the multi-microphone approaches varied from 1.1 dB for cafeteria noise conditions to 6.1 dB for meeting room environments in four CI recipients. Margo et al. (1995) evaluated a two-microphone beam-forming approach with eight Nucleus users in a take-home trial for a period of 5 to 8 weeks. Subjective reports from the CI recipients indicated that the beam-forming approach produced better sound quality and was preferred to their normal sound processing algorithms in noisy environments. Hersbach et al. (2012) tested a combination of NR approaches (i.e., an NR algorithm combined with several directional microphone algorithms available in the Cochlear CP810 sound processor) on 14 CI users. The results showed that the proposed approach could improve CI performance in noise. Later on, Hersbach et al. (2013) utilized a post-filter beam-forming technology and showed substantial improvements in intelligibility performances over conventional NR on CI recipients in some noisy conditions, where the noise is spatially separated from target speech. More recently, Buechner et al. (2014) investigated the performance for CI recipients of monaural and binaural beam-forming technologies with an additional NR approach. The results showed that for CI recipients, both the single-channel adaptive NR and adaptive binaural beam-forming techniques were significantly superior to omnidirectional microphones.

It is now believed that multi-microphone–based NR approaches can substantially increase speech intelligibility in noisy conditions, particularly in situations where there are

[1]Department of Biomedical Engineering, National Yang-Ming University, Taipei, Taiwan; [2]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan; [3]National Institute of Information and Communications Technology, Japan; [4]Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen, China; [5]Department of Mechatronic Engineering, National Taiwan Normal University, Taipei, Taiwan; [6]Department of Otolaryngology, Far Eastern Memorial Hospital, New Taipei, Taiwan; [7]Department of Otolaryngology, Cheng Hsin General Hospital, Taipei, Taiwan; [8] Department of Internal Medicine, Cheng Hsin General Hospital, Taipei, Taiwan; [9]Faculty of Medicine, School of Medicine, National Yang Ming University, Taipei, Taiwan; and [10]School of Electrical and Computer Engineering, Georgia Institute of Technology, Georgia, USA.

0196/0202/2017/XXXX-00/0 • Ear & Hearing • Copyright © 2017 Wolters Kluwer Health, Inc. All rights reserved • Printed in the U.S.A.

<zdoi; 10.1097/AUD.0000000000000537>

a few sources of interference. However, the performance of multi-microphone–based approaches might degrade in reverberant environments, and their applicability is often restricted to acoustic situations in which the speech and noise sources are spatially separated (Wouters & Vanden Berghe 2001). Compared with multi-microphone–based NR approaches, single-microphone NR methods are economically more favorable since an additional microphone increases both device expense and computational cost. Moreover, a single-microphone NR approach can be readily used as a post-filter for a multi-microphone–based NR approach. Various single-microphone NR techniques have been proposed, such as INTEL (Weiss et al. 1975; Hochberg et al. 1992), log minimum mean squared error (logMMSE) (Ephraim & Malah 1985), Wiener filter based on a priori signal to noise ratio (SNR) estimation (Wiener) (Scalart 1996), Karhunen–Loéve transform (KLT) (Rezayee & Gazor 2001; Hu & Loizou 2003; Loizou et al. 2005;), ClearVoice (Buechner et al. 2010), SNR-based (Dawson et al. 2011) NR approaches, and generalized maximum a posteriori spectral amplitude (Tsao & Lai 2016). Most of these NR techniques were developed by exploring the statistical properties of speech and noise signals (Loizou 2013). Recently, Chen et al. (2015) evaluated several single-microphone NR approaches for Mandarin CI recipients and found that these approaches performed inconsistently in various environmental conditions, although most approaches effectively improved speech recognition results in noisy conditions. Although the conventional single-microphone NR approaches can so far provide notable benefits to CI recipients under stationary noise conditions, there is still plenty of room for performance improvement under challenging listening conditions, such as when the competing signal is speech (Stickney et al. 2004) or a fast-changing noise (Xu et al. 2015).

Recently, deep learning (Hinton et al. 2006) based models, constructed with multiple hidden layers, have demonstrated outstanding performance on a wide variety of pattern classification (Hinton et al. 2012; Chen et al. 2017) and regression tasks (Lu et al. 2013, 2014; Narayanan & Wang 2013; Xia & Bao 2014; Wang et al. 2014; Weninger et al. 2015; Xu et al. 2015; Fu et al. 2016; Chen et al. 2016, 2017; Goehring et al. 2017; Kolbæk et al. 2017; Wang & Chen 2017; Xu et al. 2017). Lu et al. (2013) proposed a deep denoising autoencoder (DDAE)–based NR approach, which casts NR as a nonlinear encoder–decoder task to map noisy to clean speech features. Lu et al. found that the performance of the DDAE NR approach outperforms to that of a conventional single-microphone NR approach (i.e., MMSE plus an improved minimum controlled recursive averaging noise-tracking algorithm [Cohen 2003]) to several standardized objective evaluations in two types of noises (i.e., factory and car noise signals). More recently, Lai et al. (2017) evaluated the results of vocoded speech in CI simulation using a mismatched DDAE model (i.e., different noises and speech utterances for the training and testing phases). The results of both objective evaluations and subjective listening tests showed that under the conditions of nonstationary noise distortion, the DDAE-based NR approach yielded higher intelligibility scores than those obtained with conventional NR techniques. Although confirmed effective in CI simulations, the efficacy of the DDAE NR approach for real CI recipients remains unevaluated. It was noted that the number of just noticeable differences (JND) loudness steps was different between normal-hearing listeners and CI recipients (Zeng & Shannon 1999; Zeng et al. 2002). Thus, a compression algorithm, such as a loudness growth function (Khing et al. 2013) or an adaptive envelope compression strategy (Lai et al. 2015), is required to map the acoustic signals to currents. The result of previous studies (Chung 2004; Lai et al. 2015) showed that the benefits of the enhanced modulation depth (Shannon et al. 1995) of envelope signals by NR might be reduced when compression is imposed. Therefore, a main contribution of the present study is to test the DDAE NR approach on real recipients to further verify its effectiveness in CI signal processors.

Although a mismatched DDAE model can already provide satisfactory NR performances (Lai et al. 2017), the best performance of DDAE NR can only be achieved when working under the same noise condition as that in the training set. Therefore, a novel NR approach, referred to as NC + DDAE (NC here stands for noise classifier), is proposed in this study to further improve the benefits of DDAE-based NR for CI recipients.

The objectives of the present study are as follows. First, we propose a novel deep learning–based NR (NC + DDAE) approach for CI recipients. Second, the effectiveness of the NC and DDAE modules was individually verified. Third, we investigate the clinical effectiveness of this approach for CI recipients under noisy conditions using challenging noise types and SNR levels. Fourth, we compare the performance of speech recognition between the proposed NC + DDAE NR approach with several conventional single-microphone NR techniques.

MATERIALS AND METHODS

Classical NR Algorithms

Let $\boldsymbol{x}$ and $\boldsymbol{n}$ denote the speech and noise signals, respectively. The noisy speech signal, $\boldsymbol{y}$, is given by

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{n}. \tag{1}$$

The goal of NR is to compute the enhanced speech signals $\hat{\boldsymbol{x}}$ from $\boldsymbol{y}$. We implemented two classical NR approaches, namely, the logMMSE (Ephraim & Malah 1985) and KLT (Rezayee & Gazor 2001; Hu & Loizou 2003) because both techniques were confirmed to effectively improve speech intelligibility for Mandarin CI patients under noisy conditions in a previous study (Chen et al. 2015).

The logMMSE approach computes the spectrum of the enhanced speech signals, $\hat{\boldsymbol{X}}_i$, by filtering the spectrum of the noisy signals, $\boldsymbol{Y}_i$, through a gain function $\boldsymbol{G}_i$, where i denotes the i th frequency bin. The gain function $\boldsymbol{G}_i$ is estimated based on an MMSE criterion (Ephraim & Malah 1984) of the difference between $\hat{\boldsymbol{X}}_i^{\text{LPS}}$ and $\boldsymbol{X}_i^{\text{LPS}}$, which denote the log power spectra (LPS) (Du & Huo 2008) of $\hat{\boldsymbol{X}}_i$ and $\boldsymbol{X}_i$, respectively.

On the other hand, the KLT NR approach is a subspace method, which adopts a linear estimator, $\mathbf{H}$, to obtain the enhanced speech signal ($\hat{\boldsymbol{x}}$) by

$$\hat{\boldsymbol{x}} = \mathbf{H} \cdot \boldsymbol{x} + \mathbf{H} \cdot \boldsymbol{n}, \tag{2}$$

The error signal $\boldsymbol{\varepsilon}$ is estimated by

$$\boldsymbol{\varepsilon} = \hat{\boldsymbol{x}} - \boldsymbol{x} = \left(\mathbf{H} - \mathbf{I}\right) \cdot \boldsymbol{x} + \mathbf{H} \cdot \boldsymbol{n} = \boldsymbol{\varepsilon}_{x} + \boldsymbol{\varepsilon}_{\boldsymbol{n}}, \tag{3}$$

where $\boldsymbol{\varepsilon}$ can be divided into two components, $\boldsymbol{\varepsilon}_x$ and $\boldsymbol{\varepsilon}_n$, which represent the speech distortion and residual noise, respectively. The transformation matrix ($\mathbf{H}$) is estimated by minimizing the speech distortion with the constraint of a predetermined level of residual noise. Meanwhile, the KLT approach utilizes the voice activity detector method proposed by Mittal and Phamdo (2000) with a threshold value of 1.2.

Further details of the two unsupervised NR approaches that were tested in this study can be found in Loizou (2013), Ephraim and Van Trees (1995), and Hu et al. (2007).

The Proposed NC + DDAE NR Approach

As described in the Introduction, an additional NC module was adopted to further improve the performance of the DDAE-based NR method. We dubbed the proposed NC + DDAE approach; a block diagram of the overall system is shown in Figure 1. As shown in the figure, when given noisy speech, the NC module first determines the noise type and selects the most suitable DDAE model to perform NR. In the DDAE module, we prepared multiple noise-dependent DDAE (ND-DDAE) models and one noise-independent DDAE (NI-DDAE) model. Each ND-DDAE was trained on a specific noise type, and the NI-DDAE was trained on multiple noise types. In the following section, we detail the NC and DDAE modules individually.

Deep Neural Network–Based Noise Classification • In the proposed system, the NC module was constructed based on a deep neural network (DNN) model. A DNN model is a feed-forward artificial neural network with many hidden layers between the input and output layers. The output at the *j*th node of the *l*th layer in a DNN, $h_j^{(l)}$, is produced by Eq. (4):

$$h_j^{(l)} = F\left(\sum_i h_i^{(l-1)} W_{ij}^{(l-1)} + b_j^{(l-1)}\right) \tag{4}$$

where $h_i^{(l-1)}$ denotes the output from the *i*th node in the ($l - 1$) th layer; b_j^l is the bias of unit *j*, and W_{ij}^l is the weight between hidden unit *j* and *i*; for the input layer, we have $h_j^{(0)} = x_j$, where x_j is the *j*th element of the input vector, $\boldsymbol{x}$; $F(\cdot)$ is an activation function, and the following logistic function (Glorot et al. 2011) is used in this study:

$$F(t) = \frac{1}{1+e^{-t}}. \tag{5}$$

For the application of pattern classification, the vector of the last layer is further converted to a probability representation with another function to generate the final output, $\hat{\boldsymbol{y}}$, by $\hat{\boldsymbol{y}} = G\left(\boldsymbol{h}^{(L)}\right)$. For classification tasks, a softmax function for $G(.)$ is generally used to produce a normalized probability-based output, which is defined as (Hinton et al. 2015)

$$G(\boldsymbol{t})_r = \frac{\exp(t_r)}{\sum_{n=1}^{N}\exp(t_n)} \text{ for } r = 1,\ldots,N. \tag{6}$$

The parameter set of the DNN model, $\theta = \left\{\boldsymbol{W}^l, \boldsymbol{b}^l, l = 0,\ldots,(L-1)\right\}$, where $\boldsymbol{W}^l$ and $\boldsymbol{b}^l$, respectively, denoting the entire set of weights and biases of the *l*th layer in the DNN model, is estimated by the following cost function (Yu & Deng 2014):

$$\theta^* = \arg\min_\theta \left\{L\left(\hat{\boldsymbol{y}}, \boldsymbol{y}; \theta\right)\right\} \tag{7}$$

where $L(\cdot)$ is a loss function (Yu & Deng 2012) and $\boldsymbol{y}$ denotes the correct label. In the proposed NC + DDAE system, the label of the NC module is the noise type of the input speech signals. A back-propagation algorithm is applied to estimate the parameter set θ^* in Eq. (7). The details of the above approaches for training the DNN classifier can be found in previous studies (Bengio 2009; Mohamed et al. 2012; Chen et al. 2016).

Mel-frequency cepstral coefficients (MFCCs) (Davis & Mermelstein 1980; Rabiner & Juang 1993) were used as the acoustic feature in the NC module. The MFCC is popularly used in various acoustic pattern classification tasks, such as music classification (Räsänen et al. 2011) and automatic auscultation (Chen et al. 2016). The MFCC feature extraction process includes six steps: (1) *pre-emphasis*: to compensate for the high-frequency portion that is suppressed during the sound production mechanism in humans; (2) *windowing*: where a given signal is divided into a sequence of frames; (3) *fast Fourier transform:* to obtain the frequency response of each frame for spectral analysis. (4) *Mel-filtering:* to integrate the frequency compositions from a Mel-filter band into

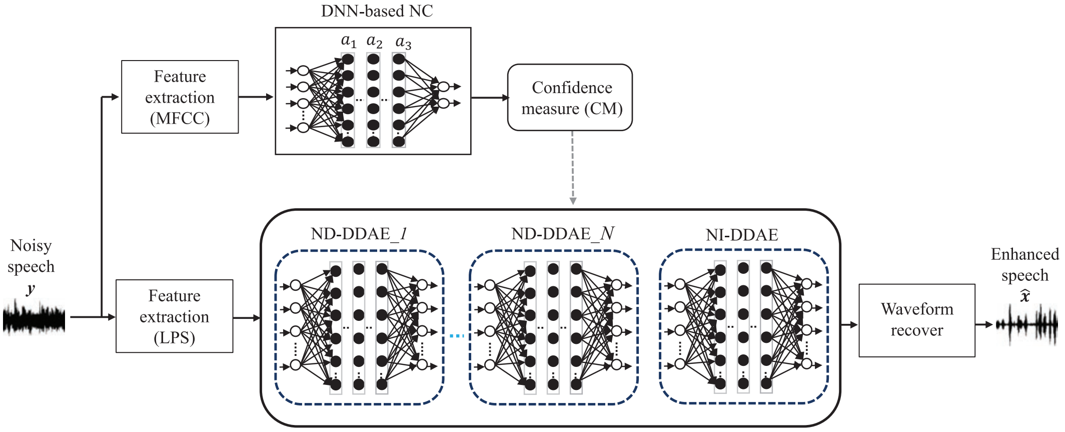

Fig. 1. The block diagram of the proposed noise classifier + deep denoising autoencoder noise reduction (NC+DDAE NR) approach. CM, confidence measure; DNN, deep neural network; LPS, log power spectra; MFCC, Mel-frequency cepstral coefficient; ND-DDAE, noise-dependent DDAE models; NI-DDAE, noise-independent DDAE model.

one-energy intensity; (5) *nonlinear transformation:* this transformation takes the logarithm of all Mel-filter band intensities; and (6) *discrete cosine transform:* to transform the logarithm of all Mel-filter band intensities into MFCCs. As reported in previous studies, 39-dimension MFCCs (13-dimension original MFCC + 13-dimension velocity + 13-dimension acceleration features) may more accurately characterize acoustic patterns and, thus, yield better recognition performance (Furui 1981; Ma et al. 2006). Therefore, the 39-dimension MFCCs were adopted as the acoustic features for the NC module in this study.

In the NC module, we further adopted the confidence measure (CM) (Jiang 2005) to evaluate the reliability of the recognition results. The CM score indicates the degree that we can trust the recognition results: a higher score denotes a higher confidence in the recognition output and vice versa. In this study, the formula for the CM score was based on that in a previous report (Mengusoglu & Ris 2001), defined as follows:

$$\text{CM score} = \frac{1}{N_{\text{seg}}}\sum_{c=1}^{N_{\text{seg}}}\log\left(p\left(q_k^c|x^c\right) / p\left(q_{\text{best}}^c|x^c\right)\right) \tag{8}$$

where q_k^c denotes a *k*th class classification result and $p\left(q_k^c|x^c\right)$ is the probability of q_k^c given the *c*th speech frame x^c. Similarly, $p\left(q_{\text{best}}^c|x^c\right)$ is the probability of the best classification result, given the *c*th speech frame. N_{seg} is the segment of frames that was used to compute the CM score. It should be noted that the classification results are determined based on voting the results of N_{seg} frames, and $p\left(q_k^c|x^c\right)$ and $p\left(q_{\text{best}}^c|x^c\right)$ are not always the same throughout these N_{seg} frames. After the calculation of the CM score, a threshold is defined to determine the confidence of the classification results. A classification result with its CM score higher than a threshold indicates that the result is trustworthy; on the other hand, a classification result with its CM score below the threshold suggests that the classifier does not have a sufficient confidence regarding the result. As introduced earlier, the goal of the NC module is to determine the noise type, which is then used to select the most suitable DDAE model to perform NR. Thus, if the CM score of the determined noise type is higher than the threshold, the corresponding ND-DDAE model is directly selected to perform NR. On the other hand, if the CM score is lower than the threshold, the NI-DDAE model is used to perform NR.

The DDAE NR Algorithm • Figure 2 illustrates the structure of the DDAE-based NR module. DDAE is a supervised NR approach that enhances speech by deriving a mapping function between noisy and clean speech signals based on the architecture of DNNs (Lu et al. 2013). There are two phases in the DDAE NR approach: training and testing. In the training phase, a set of noisy-clean speech pairs is prepared. The noisy-clean speech signals are first converted into LPS (Du & Huo 2008) features, usually used for DNN-based NR approaches (Lu et al. 2013; Xu et al. 2015). The short-time Fourier analysis is applied to the input signal, computing the discrete Fourier transform of each overlapping windowed frame. Then the LPS spectra are obtained. The LPS features of noisy ($\boldsymbol{Y}_m^{\text{LPS}}$) and clean ($\boldsymbol{X}_m^{\text{LPS}}$) speech are then placed in the input and output sides of the DDAE model; *m* denotes the frame in the short-time Fourier transform (Wang et al. 2014). For a DDAE model with *D* hidden layers, we have

$$\begin{aligned} h^1\left(\boldsymbol{Y}_m^{\text{LPS}}\right) &= \sigma\left(\boldsymbol{W}^0\boldsymbol{Y}_m^{\text{LPS}} + \boldsymbol{b}^0\right), \\ &\vdots \\ h^D\left(\boldsymbol{Y}_m^{\text{LPS}}\right) &= \sigma\left(\boldsymbol{W}^{D-1}h^{D-1}\left(\boldsymbol{Y}_m^{\text{LPS}}\right) + \boldsymbol{b}^{D-1}\right), \\ \widehat{\boldsymbol{X}}_m^{\text{LPS}} &= \boldsymbol{W}^D h^D\left(\boldsymbol{Y}_m^{\text{LPS}}\right) + \boldsymbol{b}^D, \end{aligned} \tag{9}$$

where $\{\boldsymbol{W}^0 \ldots \boldsymbol{W}^D\}$ are the matrices of the connection weights, and $\{\boldsymbol{b}^0 \ldots \boldsymbol{b}^D\}$ are the bias vectors for DDAE model, and $\hat{\boldsymbol{X}}_m^{\text{LPS}}$ is the vector containing the logarithmic amplitudes of enhanced speech corresponding to the noisy counterpart $\boldsymbol{Y}_m^{\text{LPS}}$. The logistic function defined in Eq. (5) is also used for the activation function, $\sigma(\cdot)$, in Eq. (9) in this study. Finally, the parameters are determined by optimizing the following objective functions:

$$\theta^* = \arg\min_{\theta}\left(F(\theta) + \eta^0\left\|\boldsymbol{W}^0\right\|_2^2 + \cdots + \eta^D\left\|\boldsymbol{W}^D\right\|_2^2\right), \tag{10}$$

$$F(\theta) = \frac{1}{M}\sum_{m=1}^{M}\left\|\boldsymbol{X}_m^{\text{LPS}} - \hat{\boldsymbol{X}}_m^{\text{LPS}}\right\|_2^2,$$

where *M* is the total number of training samples (noisy-clean pairs). In the testing phase, the corresponding DDAE model, selected based on NC, was used to transform the noisy speech LPS signal ($\boldsymbol{Y}_m^{\text{LPS}}$) to an enhanced speech signal ($\hat{\boldsymbol{X}}_m^{\text{LPS}}$). More detailed information of the DDAE NR approach can be found in the study by Lu et al. (2013).

As shown in Figure 1, we prepared *N* ND-DDAE models (i.e., ND-DDAE_*1* to ND-DDAE_*N*) and an NI-DDAE model. All of these (*N* + 1) models were prepared during the training phase. Notably, each of the ND-DDAE models is trained on a specific type of noise and, thus, can more accurately characterize the noisy to clean speech transformation of that particular noise type. On the other hand, the NI-DDAE model is trained on multiple noise types and, thus, may possess less characterization capability than ND-DDAE on specific noise types. Nevertheless, since the NI-DDAE model is trained on multiple noise types, it can give better NR performance for new noise types. The proposed NC + DDAE NR approach can be summarized as follows: (1) for test noises that have been encountered in the training set, the system selects the most appropriate ND-DDAE model to perform NR; (2) for test noises that are not part of the training set, the NI-DDAE, which has better generalization capability for different noises, is used to perform NR.

Subjects

A total of nine native Taiwan Mandarin-speaking CI recipients participated in this study. Six of the subjects used an Advanced Bionics HiRes-120 sound coding strategy (Firszt et al. 2009) and another three used a cochlear advanced combination encoder sound coding strategy (Skinner et al. 2002). A previous study (Holden et al. 2013) demonstrated that CI recipients can, on average, achieve an accuracy rate of 90% in a consonant–vowel nucleus–consonant test after 6.9 months of training (Peterson & Lehiste 1962). Therefore, we only recruited subjects who had received their CI at least 7 months before the experiments. Detailed individual biographical information for the nine subjects is presented in Table 1. The study procedures

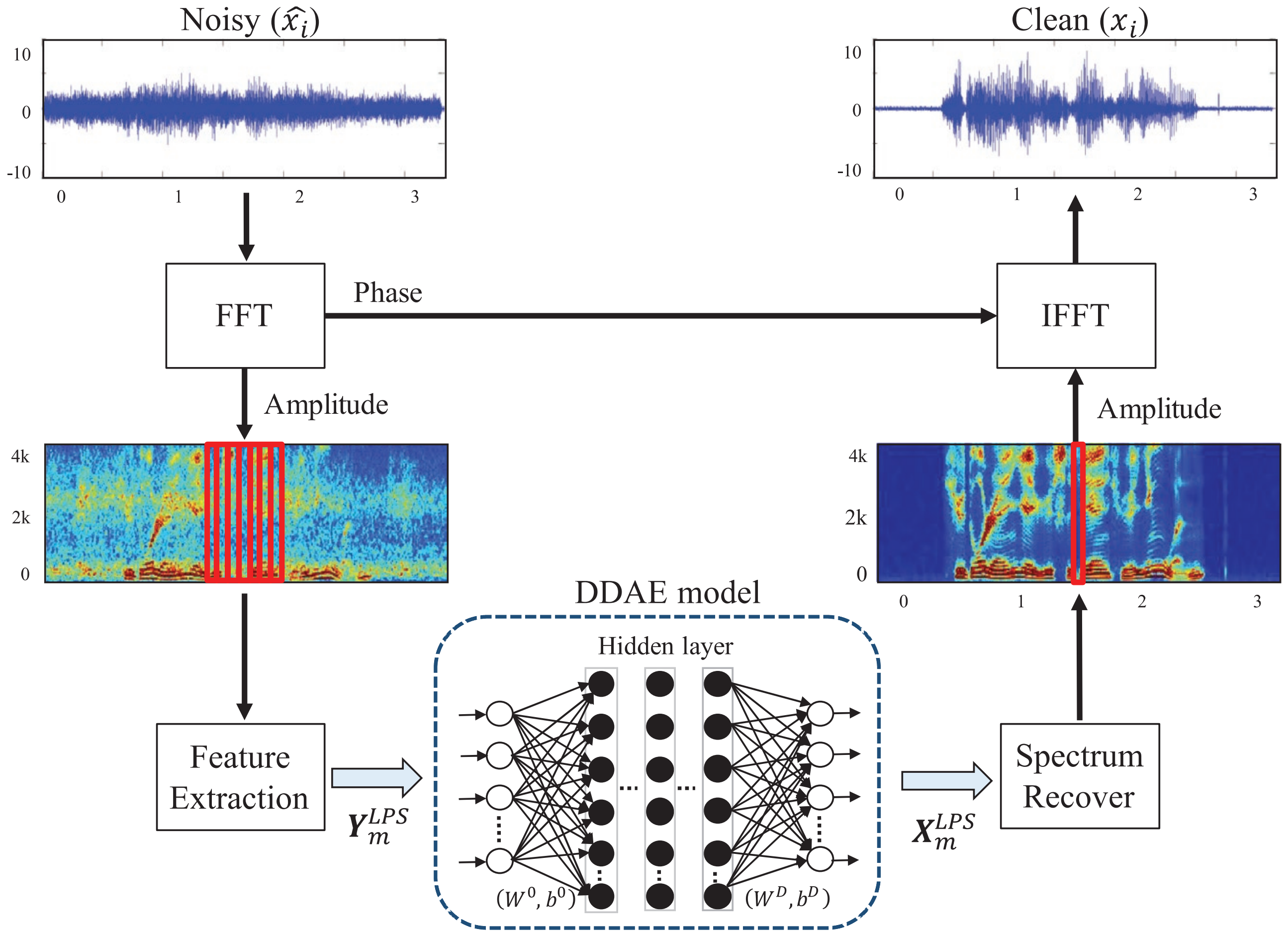

Fig. 2. Structure of a deep denoising autoencoder (DDAE)–based noise reduction (NR) system. FFT, fast Fourier transform; IFFT, inverse fast Fourier transform.

were approved by the Institutional Review Boards, and informed consent was obtained from each subject before testing.

Procedure

We conducted experiments using the Taiwan Mandarin version of the "hearing in noise test (TMHINT) sentences" (Huang 2005). In total, we recorded 320 clean utterances with ten Chinese characters in each utterance. All of these utterances were pronounced by a native male Taiwan Mandarin speaker, with a fundamental frequency ranging from 96 to 208 Hz, and were recorded at a sampling rate of 16 kHz. Among these 320 utterances, 120 utterances were used for training and the other 200 were used for testing. We adopted 12 common noise types to form the noisy training and testing data, including the inside of an airplane, engine acceleration of a car, inside of a bus, inside of a train, laughter of a crowd, applause of a crowd, cafeteria chatter, street noise, construction jackhammer, speech-shaped noise, two-talker babble, and the cry of a baby. The 12 noise types were collected from two data sets (Ideas 2002; Loizou 2013). Figure 3 shows the spectrograms of the 12 noise types. The length of each noise source was approximately 5 minutes, and 85% of the noise signals were used to establish the training set, while the remaining 15% were used to build the testing set. For each training utterance, we artificially generated a noisy version by adding noise at seven SNR levels (−10, −5, −3, 0, 3, 5, and 10 dB). These training data were used to establish the NC module and train the 12 ND-DDAE models. In addition to these 12 noise types, we collected an additional 104 noise types (Xu et al. 2015) to prepare an additional training set by corrupting the 120 training utterances with these 104 noise types at seven different SNR levels. This additional training set was used to train the NI-DDAE model. In this study, we intend to focus more on the challenging conditions and, thus, evaluated the performance using two fast-changing nonstationary noise types: two equal-level interfering male talkers (2T) (Lai et al. 2015) and construction jackhammer (CJ) noises (Loizou 2013), to corrupt the test sentences at two SNR levels of 0 and 5 dB. The 2T noise source shared similar characteristics to the target speech, and the nonstationary CJ noise had fast-changing signal characteristics. Hence, there were 1200 sentences (=60 sentences × 2 SNR levels × 2 maskers × 5 NR approaches) used to form the testing set.

The DNN-based NC model consisted of three layers, with 100 neurons in each hidden layer. The 12 ND-DDAE models and the NI-DDAE model used in this study had five hidden layers, with 500 neurons in each hidden layer. As illustrated in Figure 2, a speech framing strategy (Xu et al. 2015), with a 16-ms window and an 8-ms frame shift, was applied to each speech utterance. Each windowed speech segment was processed by a 256-point fast Fourier transform and then converted to an LPS feature vector with 129 dimensions (the other 127 components were redundant and not used). In the testing phase, given the testing utterances, the NC module performs noise classification based on the first 31 frames, that is, the first 0.256 seconds, of the entire utterance.

TABLE 1. Individual biographical data of the evaluated CI recipients

Subjects	Age (Years)	Sex	Age at HL Onset (Years)	CI Use (Years)	Etiology of Deafness	Processing Strategy
S1	10	F	2	2	LVAS	HiRes 120
S2	25	F	1	3	Congenitally deaf	HiRes 120
S3	23	M	3	4	Congenitally deaf	HiRes 120
S4	33	M	2	2	Congenitally deaf	HiRes 120
S5	22	M	1	20	Congenitally deaf	ACE
S6	45	M	33	7	Progressive hearing loss	ACE
S7	30	M	13	2	Progressive hearing loss	HiRes 120
S8	44	M	28	2	Noise-induced hearing loss	HiRes 120
S9	24	M	2	19	Fever	ACE

ACE, advanced combination encoder; CI, cochlear implant; HL, hearing loss; LVAS, large vestibular aqueduct syndrome.

This setup assumes that the first few frames of a speech utterance contains noise signals only. Then, the noise type classification was carried out by voting on the outputs of the 31 frames. Next, a CM score was computed to test the confidence of the classification result. When the CM score was higher than the predefined threshold (−0.1 in this study), then the corresponding ND-DDAE was selected based on the classification result. On the other hand, when the CM score was lower than the threshold, the NI-DDAE model was selected. Finally, the selected DDAE model was used to perform NR on the entire speech utterance.

It is important to investigate the effectiveness of the NC module with the ND-DDAE in the proposed NC + DDAE approach. Accordingly, we designed two systems: system 1, without the NC module, that is, the NI-DDAE model alone was used to perform NR; system 2, the complete NC + DDAE NR framework (including the NC module with 12 ND-DDAE and NI-DDAE models) was used to perform NR. These two systems are denoted as DDAE and NC + DDAE, respectively, in the following discussions.

To evaluate the NR performance, we adopted the normalized covariance measure (NCM) (Ma et al. 2009) to objectively evaluate the intelligibility scores of the processed speech utterances. The NCM quantifies the changes in the modulation depth between the input and output envelopes to evaluate the performance of speech intelligibility (Shannon et al. 1995; Dorman et al. 1997; Loizou 1999; Chen & Lau 2014; Lai et al. 2015; Lai et al. 2017) and has been popularly used in predicting the intelligibility of vocoded speech in CI simulations (Lai

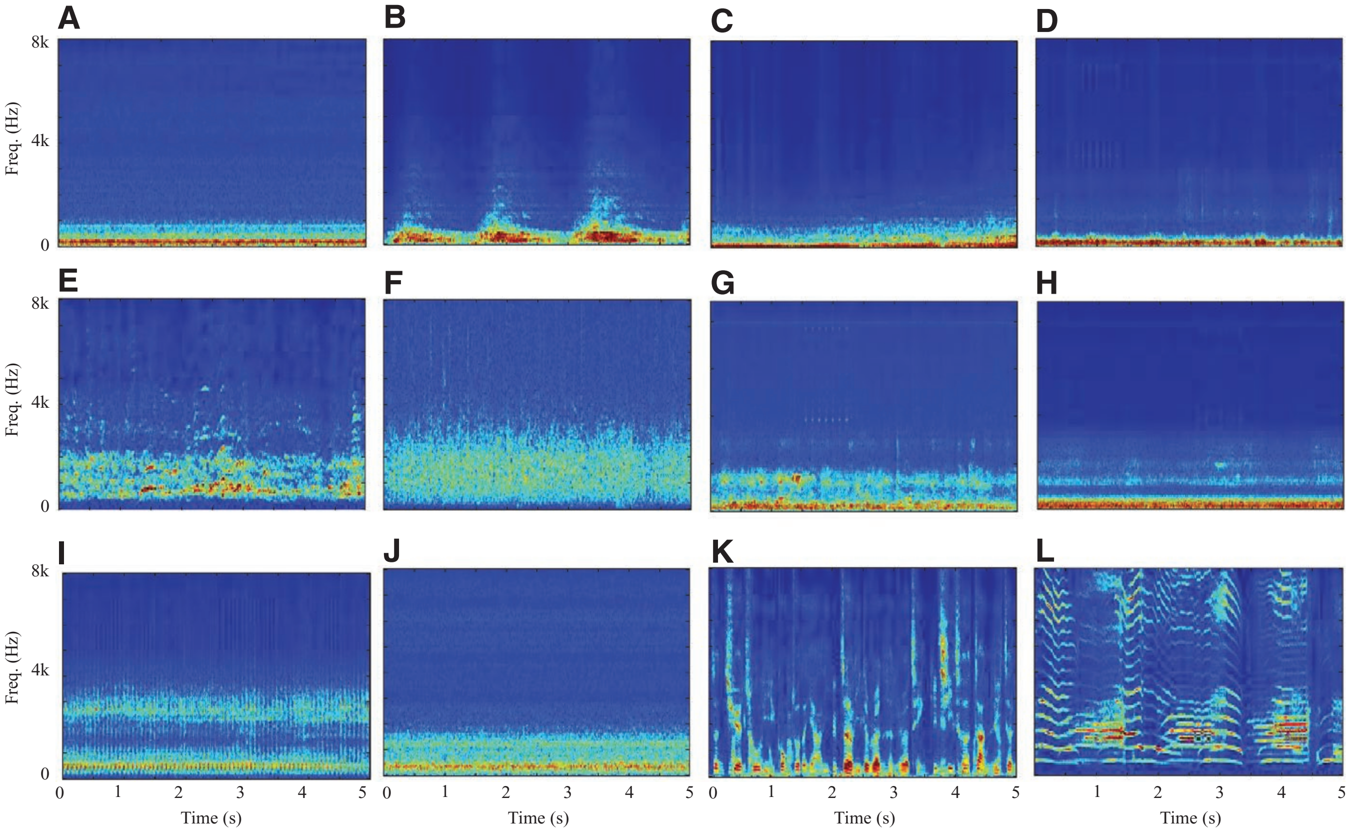

Fig. 3. Spectrograms of 12 noise signals: (A) inside of an airplane; (B) engine acceleration of a car; (C) inside of a bus; (D) inside of a train; (E) laughter of a crowd; (F) applause of a crowd; (G) cafeteria chatter; (H) street noise; (I) construction jackhammer (CJ); (J) speech-shaped noise; (K) two-talker babble (2T); and (L) the cry of a baby.

et al. 2017). More detail about the NCM measure can be found in Ma et al. (2009) and Chen (2012). To compute the NCM scores, the noise-reduced speech was first processed with an eight-channel noise vocoder (Shannon et al. 1995; Dorman et al. 1997; Lai et al. 2015) (with cutoff frequencies at 80, 221, 426, 724, 1158, 1790, 2710, 4050, and 6000 Hz) to simulate the speech heard by CI recipients. Then the vocoded speech was used to compute the NCM scores. The details of this evaluation setup can be found in Lai et al. (2017). In addition to the proposed deep learning–based models, we reported the results of noisy, KLT, and logMMSE for comparison purposes.

In addition to the objective NCM scores, clinical listening tests were carried out to assess the effectiveness of the proposed NR approach on real CI recipients. To ensure that fatigue did not affect the subjects' results, each subject only heard a total of 20 test conditions (2 noise types [2T and CJ] × 2 SNR levels [0 and 5 dB] × 5 signal processing strategies [noisy, logMMSE, KLT, DDAE, and NC + DDAE]). Each condition contained 10 sentences, and the order of the 20 conditions was randomized across subjects. None of the 10 sentences were repeated across the testing conditions. The subjects were instructed to verbally repeat what they had heard and allowed to hear the stimuli twice. Each subject was given a 5-minute break after every 30 minutes of testing.

All subjects used their own clinical speech processors. The built-in NR functions of the subjects' speech processors were temporarily disabled during the testing stage. The noisy and enhanced speech signals were preprocessed and then played at 65 dB SPL through a speaker and then processed through a CI processor to simulate the performance of each NR approach for CI users. The subjects were seated in a soundproof room (provided by Acoustic System, Inc.) equipped with a Notebook connected to a GSI Audiostar Pro audiometry device (GSI, MN). A double-blind method (Wilson et al. 1980) was used in this study. Each subject underwent the test under 20 conditions. The word correct rate (WCR) (Chen & Loizou 2011; Chen et al. 2013, 2015; Lai et al. 2015, 2017) was used as the evaluation metric, which is calculated by dividing the number of correctly identified words by the total number of words under each test condition.

RESULTS

Classification Results of the NC Module

The confusion matrix of the DNN-based NC results of the test set is listed in Table 2. The columns and rows correspond to the predicted class determined by the DNN-based NC approach and the actual class of the input feature, respectively. The diagonal of the confusion matrix represents the classification accuracy. For instance, the classification accuracy of the 11th noise type was 96.9%. Moreover, 1.1% of the actual input feature of the 11th noise type (i.e., two-talker babble) was predicted as the 5th noise type (i.e., laughter of a crowd). The average accuracy for the 12 noise types was 99.6%. Therefore, the DNN-based NC approach was able to achieve high accuracy under various noisy conditions. Notably, this efficacy was demonstrated in complex noise environments.

Next, we tested the CM scores of the NC module. In addition to the 12 noise types involved in the training set, four additional unseen noise types (i.e., these noises were not part of the training set) were used, including (1) nonstationary noise of quick squeeze toy (toy-squeeze-several.mp3), (2) stationary revs of motorcycle (motorcycl-yam-250-04-revs.mp3), (3) nonstationary noise of military battle (military-battle-medieval.mp3), and (4) nonstationary noise of boat drive (boat-outboard-9hp-ob-drive.mp3) obtained from the "Sound Ideas XV MP3 Series" database (Ideas 2002). Figure 4 shows the average CM scores of the NC module; the 1st to 12th noise types were seen noise types, and the 13th to 16th noise types were unseen noise types. The results showed that, on average, the NC module provided high CM scores (close to 0.0) and small standard deviations for seen noise types. Meanwhile, the CM scores were relatively low with large standard deviations when testing the NC module with the unseen noise types. The results suggest that the NC module can accurately identify the seen noise types with high CM scores while assigning low CM scores to the unseen noise types.

TABLE 2. The confusion matrix of the DNN-based noise classification approach

	Predicted Class											
Percent	1	2	3	4	5	6	7	8	9	10	11	12
Actual class												
1	100*	0	0	0	0	0	0	0	0	0	0	0
2	0	100*	0	0	0	0	0	0	0	0	0	0
3	0	0	99.9*	0	0	0	0.1	0	0	0	0	0
4	0	0	0	100*	0	0	0	0	0	0	0	0
5	0	0	0	0	99.9*	0.1	0	0	0	0	0	0
6	0	0	0	0	0	100*	0	0	0	0	0	0
7	0	0	0	0	0	0	100*	0	0	0	0	0
8	0	0	0	0	0	0	0	100*	0	0	0	0
9	0	0	0	0	0	0	0	0	100*	0	0	0
10	0	0	0	0.1	0	0	0	0	0	99.9*	0	0
11	0	1	1	0	1.1	0	0	0	0	0	96.9*	0
12	0	0	0	0	2	0	0	0	0	0	0	98*

Noise classes 1 to 12 are listed sequentially in the following order: the inside of an airplane, engine acceleration of a car, inside of a bus, inside of a train, laughter of a crowd, applause of a crowd, cafeteria chatter, street noise, construction jackhammer, speech-shaped noise, two-talker babble, and the cry of a baby.

**The values in the diagonal are the accuracies of DNN-based noise classification.*

DNN, deep neural network.

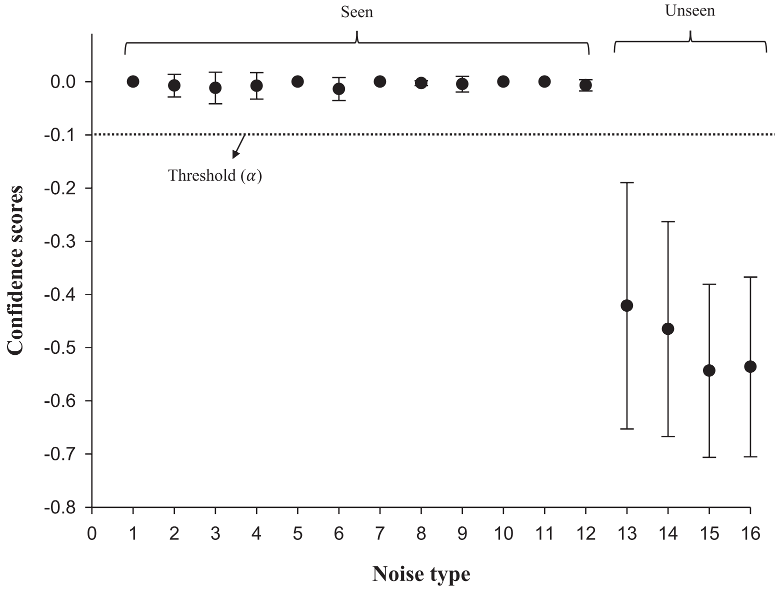

Fig. 4. The mean confidence measure (CM) scores from the deep neural network (DNN)-based noise classifier (NC). The 1st to 12th noise types are used in both training and testing sets and, thus, are denoted as "seen noise type"; The 13th to 16th noise types were only shown in the testing set and, thus, are denoted as "unseen noise type."

Objective Results of the NCM Evaluation

The NCM scores of the proposed deep learning–based and conventional NR approaches are shown in Figure 5, where the testing data were delivered under CJ and 2T noises at 0 and 5 dB SNR conditions. The results in Figure 5 indicate that both DDAE and NC + DDAE provide higher NCM scores than noisy and the two conventional NR approaches, confirming the effectiveness of the deep learning–based NR approach on the vocoded speech in CI simulations. Moreover, NC + DDAE notably outperforms DDAE, confirming that the NC module and multiple ND-DDAE models can enable vocoded speech to attain higher intelligibility. The detailed results are listed below: for the results of the 2T masker, the average scores and SEM for {noisy, logMMSE, KLT, DDAE, and NC + DDAE} are {0.222 ± 0.005, 0.171 ± 0.006, 0.157 ± 0.007, 0.253 ± 0.005, and 0.386 ± 0.004} at 0 dB SNR and {0.342 ± 0.005, 0.294 ± 0.006, 0.295 ± 0.007, 0.350 ± 0.004, and 0.393 ± 0.003} at 5 dB SNR. For the results of CJ masker, the average scores and SEM for {noisy, logMMSE, KLT, DDAE and NC + DDAE} are {0.256 ± 0.004, 0.292 ± 0.006, 0.310 ± 0.007, 0.367 ± 0.004, and 0.413 ± 0.003} at 0 dB SNR and {0.343 ± 0.004, 0.355 ± 0.006, 0.369 ± 0.006, 0.401 ± 0.003, and 0.424 ± 0.003} at 5 dB SNR.

Recognition Results of the Clinical Listening Tests

Finally, we evaluated the proposed deep learning–based approach using the mean WCR scores of the clinical listening tests on CI recipients. Figure 6 shows the WCR scores of

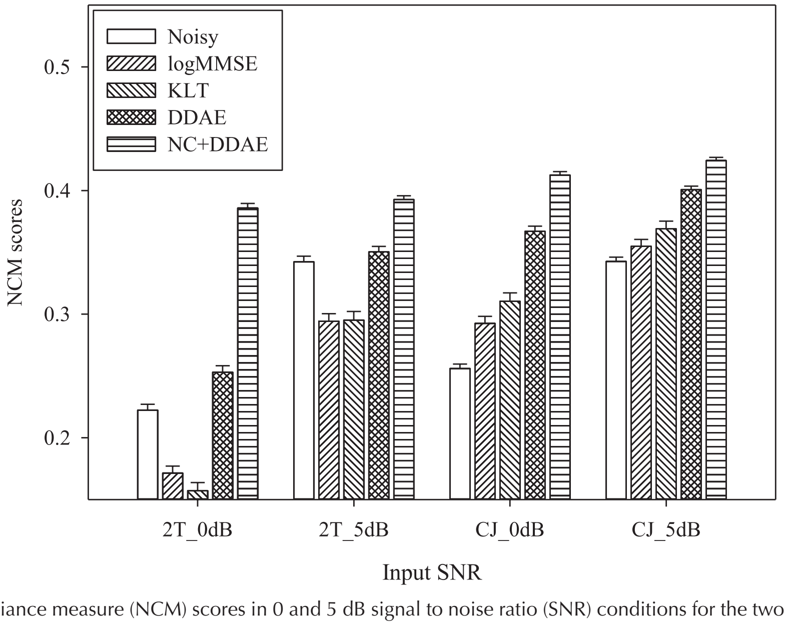

Fig. 5. Average normalized covariance measure (NCM) scores in 0 and 5 dB signal to noise ratio (SNR) conditions for the two-talker babble (2T) and construction jackhammer (CJ) maskers. Each error bar indicates one standard error of the mean (SEM). DDAE, deep denoising autoencoder; KLT, Karhunen–Loéve transform; logMMSE, log minimum mean squared error; NC, noise classifier.

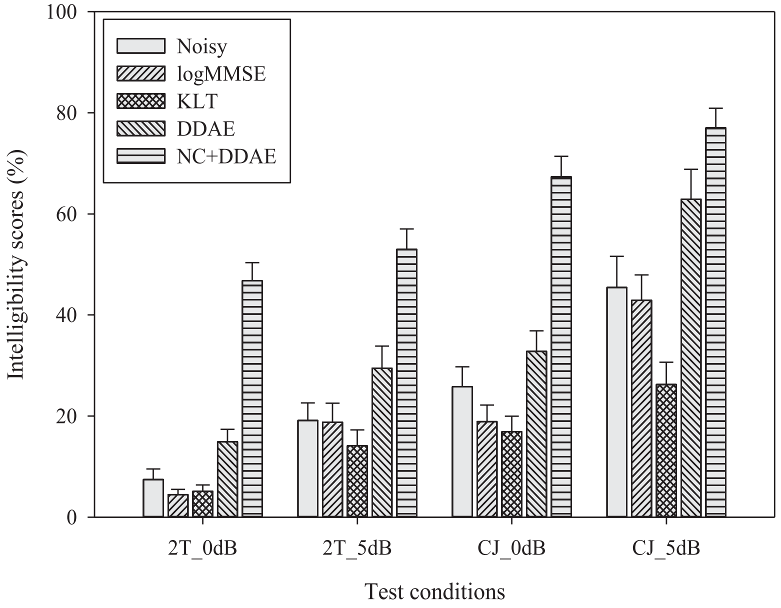

Fig. 6. Average speech recognition (WCR) scores under 0 and 5 dB SNR conditions for the two-talker babble (2T) and construction jackhammer (CJ) maskers. Each error bar indicates one standard error of the mean (SEM). DDAE, deep denoising autoencoder; KLT, Karhunen–Loéve transform; logMMSE, log minimum mean squared error; NC, noise classifier; WCR, word correct rate.

DDAE and NC + DDAE; the results of noisy, logMMSE, and KLT were also listed for comparison. Owing to the floor effect associated with conventional NR approaches, the scores were first converted to rational arcsine transform units (Studebaker 1985). From the figure, we can note that both DDAE and NC + DDAE deliver notably higher intelligibility scores than those of noisy speech and speech processed by conventional NR approaches. To further confirm the significance of improvements, one-way analysis of variance (ANOVA) (Dien 1998) and Tukey post hoc comparisons (Dien 1998) were used to analyze the results of the five NR approaches in the four test conditions. A single factor (the type of NR approach) was used in this testing. The results of these analyses are presented in Table 3, where each mean score represents the corresponding

TABLE 3. The mean scores for speech intelligibility using different strategies

Test Condition	Processed Method	*n*	Mean Score	*F*	*p*	Post Hoc Comparison* (group_i, group_j)
2T (SNR = 0 dB)				61.23	<0.001	(NC + DDAE, noisy) (NC + DDAE, logMMSE) (NC + DDAE, KLT) (NC + DDAE, DDAE) (DDAE, noisy) (DDAE, logMMSE) (DDAE, KLT)
	noisy	9	7.4			
	logMMSE	9	4.4			
	KLT	9	5.1			
	DDAE	9	14.9			
	NC + DDAE	9	46.8			
2T (SNR = 5 dB)				17.09	<0.001	(NC + DDAE, noisy) (NC + DDAE, logMMSE) (NC + DDAE, KLT) (NC + DDAE, DDAE) (DDAE, KLT)
	noisy	9	19.1			
	logMMSE	9	18.8			
	KLT	9	14.1			
	DDAE	9	29.4			
	NC + DDAE	9	53.0			
CJ (SNR = 0 dB)				30.60	<0.001	(NC + DDAE, noisy) (NC + DDAE, logMMSE) (NC + DDAE, KLT) (NC + DDAE, DDAE) (DDAE, logMMSE) (DDAE, KLT)
	noisy	9	25.8			
	logMMSE	9	18.9			
	KLT	9	16.9			
	DDAE	9	32.8			
	NC + DDAE	9	67.3			
CJ (SNR = 5 dB)				14.30	<0.001	(NC + DDAE, noisy,) (NC + DDAE, logMMSE) (NC + DDAE, KLT) (DDAE, noisy) (DDAE, logMMSE) (DDAE, KLT) (noisy, KLT)
	noisy	9	45.4			
	logMMSE	9	42.9			
	KLT	9	26.2			
	DDAE	9	62.9			
	NC + DDAE	9	77.0			

**Mean difference is significant at $\alpha = 0.05$. Each factor was included in the one-way analysis of variance and Tukey post hoc testing.*

2T, two-talker babble; CJ, construction jackhammer; DDAE, deep denoising autoencoder; KLT, Karhunen–Loéve transform; logMMSE, log minimum mean squared error; NC, noise classifier; SNR, signal to noise ratio.

intelligibility scores in Figure 6, and *n* denotes the number of subjects. For the 2T masker, the one-way ANOVA results confirmed that the intelligibility scores differed significantly among the five groups, with F = 61.23 (p < 0.001) and F = 17.09 (p < 0.001) at 0 and 5 dB SNRs, respectively. The Tukey post hoc comparisons further verified the significant differences for the following group pairs: (NC + DDAE, noisy), (NC + DDAE, logMMSE), (NC + DDAE, KLT), (NC + DDAE, DDAE), (DDAE, noisy), (DDAE, logMMSE), and (DDAE, KLT) at 0 dB SNR and (NC + DDAE, noisy), (NC + DDAE, logMMSE), (NC + DDAE, KLT), (NC + DDAE, DDAE), and (DDAE, KLT) at 5 dB SNR. Meanwhile, the one-way ANOVA results for the CJ masker confirmed that the intelligibility scores differed significantly among the five groups, with F = 30.60 (p < 0.001) and F = 14.30 (p < 0.001) at 0 and 5 dB SNRs, respectively. The Tukey post hoc comparisons verified the significant difference for the following group pairs: (NC + DDAE, noisy), (NC + DDAE, logMMSE), (NC + DDAE, KLT), (NC + DDAE, DDAE), and (DDAE, KLT) at 0 dB SNR and (NC + DDAE, noisy), (NC + DDAE, logMMSE), (NC + DDAE, KLT), (DDAE, noisy), (DDAE, logMMSE), (DDAE, KLT), and (noisy, KLT) at 5 dB SNR.

Acoustic Analysis of the Processed Speech

We qualitatively analyzed the processed speech by the NC + DDAE NR approach using amplitude envelops and spectrogram plots. Figure 7 shows an example of the amplitude envelopes of clean, noisy, and enhanced speech signals (processed by the logMMSE, KLT, DDAE, and NC + DDAE) at the 3rd channel (f_{center} = 575 Hz), which is an important frequency band for speech intelligibility (ANSI 1997), where the envelope of the 3rd channel was extracted from the third channel of a noise vocoder (Shannon et al. 1995; Dorman et al. 1997). The original utterance was recorded by a native male Mandarin speaker, saying “Secretary to help the boss write a document.” From Figure 7, we note that the DDAE- and NC + DDAE-processed envelopes give clearer noise suppression with less envelope distortion when compared to the two conventional NR approaches (please compare the parts around 1.5–2.0 seconds [marked by red arrows] in Fig. 7E, F, K, and L with those in Fig. 7C, D, I, and J). Moreover, the amplitude envelops of the DDAE- and NC + DDAE-processed speech resembled that of clean speech, as shown by Figure 7A, B. Previous studies suggested that the amplitude envelop is highly related to the speech perception of CI recipients (Chen et al. 2013; van Hoesel et al. 2005; Zeng

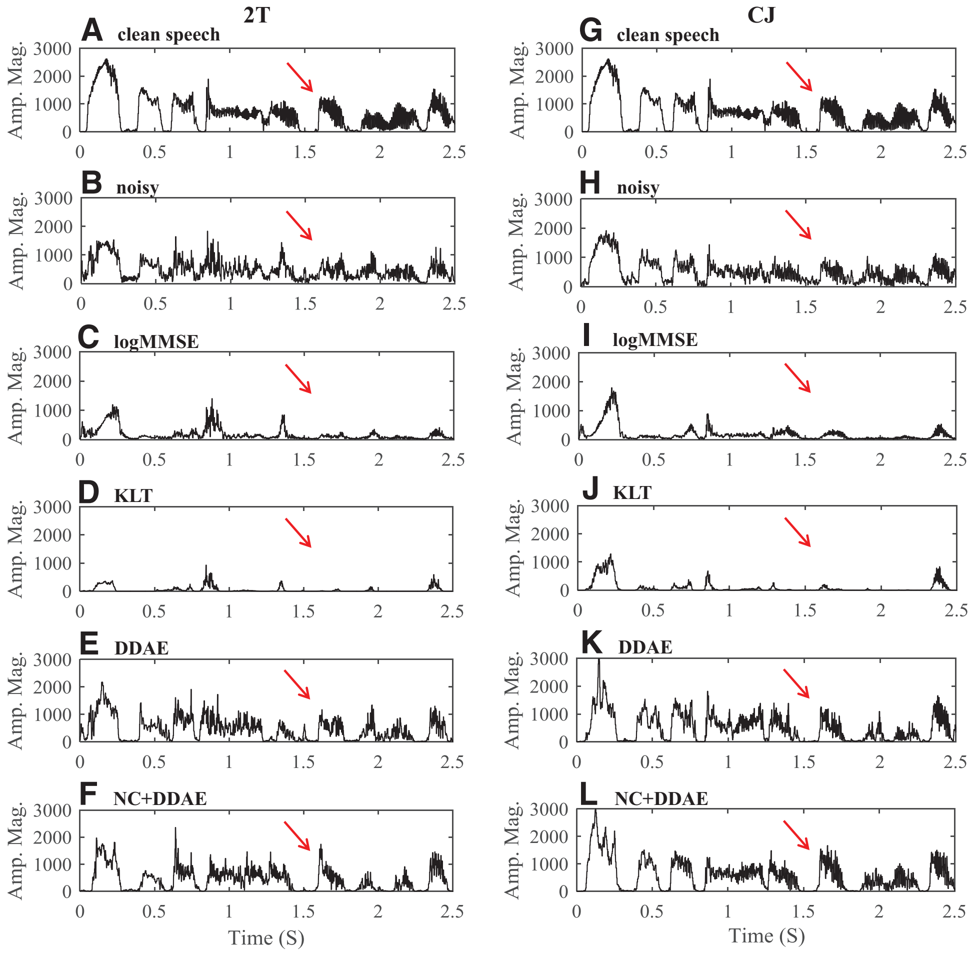

Fig. 7. Amplitude envelopes extracted from the 3rd channel (f_{center} = 575 Hz) of an utterance corrupted by two-talker babble (2T) at 0 dB signal to noise ratio (SNR) (left panels B to F) and construction jackhammer (CJ) at 0 dB SNR (right panels H to L). A and G are clean utterances; B and H are 2T and CJ noisy utterances, respectively; C to F and I to L show the sentences enhanced by logMMSE, KLT, DDAE, and NC + DDAE, respectively. Red arrows indicate the segments (approximately 1.5 to 2.0 s) where the DDAE- and NC + DDAE-processed envelopes (E, F, K, L) provide better noise suppression with less envelope distortion when compared to the two conventional NR approaches (C, D, I, J). DDAE, deep denoising autoencoder; KLT, Karhunen–Loéve transform; logMMSE, log minimum mean squared error; NC, noise classifier; NR, noise reduction.

et al. 2002). Therefore, the results in Figure 7 demonstrate the effectiveness of the proposed deep learning–based approach.

Next, we analyzed the processed utterances using spectrogram plots, which are generally adopted to show the spectral representations of a time-varying signal (Haykin 1995). Figure 8 shows the same utterances as those used in Figure 7, where Figure 8A, G are clean speech; Figure 8B and H are the 2T and CJ noisy utterances, respectively; Figure 8C–F and I–L show the utterances processed by logMMSE, KLT, DDAE, and NC + DDAE, respectively. From the results of Figure 8, we can note that first, DDAE and NC + DDAE can effectively reduce the noise components in the noisy spectra, as shown in the red boxes in Figure 8E, F, K, and L. Second, the conventional NR methods could not effectively reduce the noise components; notable residual noise signals remained in the spectrograms of the processed utterances, as shown in the red ovals in Figure 8C, D, I, and J. Third, NC + DDAE gives better NR performance than DDAE in the spectrogram of the processed utterance, as indicated by the red arrows in Figure 8E and F.

The results from the amplitude envelope and spectrogram analysis suggest that the proposed NC + DDAE system can provide improved speech signals with less noise residuals and speech distortion. Moreover, the NCM results in Figure 5 demonstrate that both DDAE and NC + DDAE outperform the other two traditional NR approaches on vocoded speech at 0 and 5 dB SNR levels. These quantitative and qualitative analysis results are consistent with that reported in Lai et al.'s (2017) study, confirming the effectiveness of deep learning–based NR in challenging listening conditions. Finally, the results in Figure 6 show that the NC + DDAE can enable CI recipients to achieve higher recognition scores. All the above findings suggest that the deep learning–based NR approach can be applied to CI signal processors as a promising method for improving speech recognition performance.

DISCUSSION AND CONCLUSION

The main contributions of the present study are fourfold: (1) we proposed a deep learning–based NR approach (NC + DDAE) for CI recipients; (2) we confirmed the effectiveness of using the NC module with multiple ND-DDAE models in the proposed NC + DDAE framework; (3) we investigated

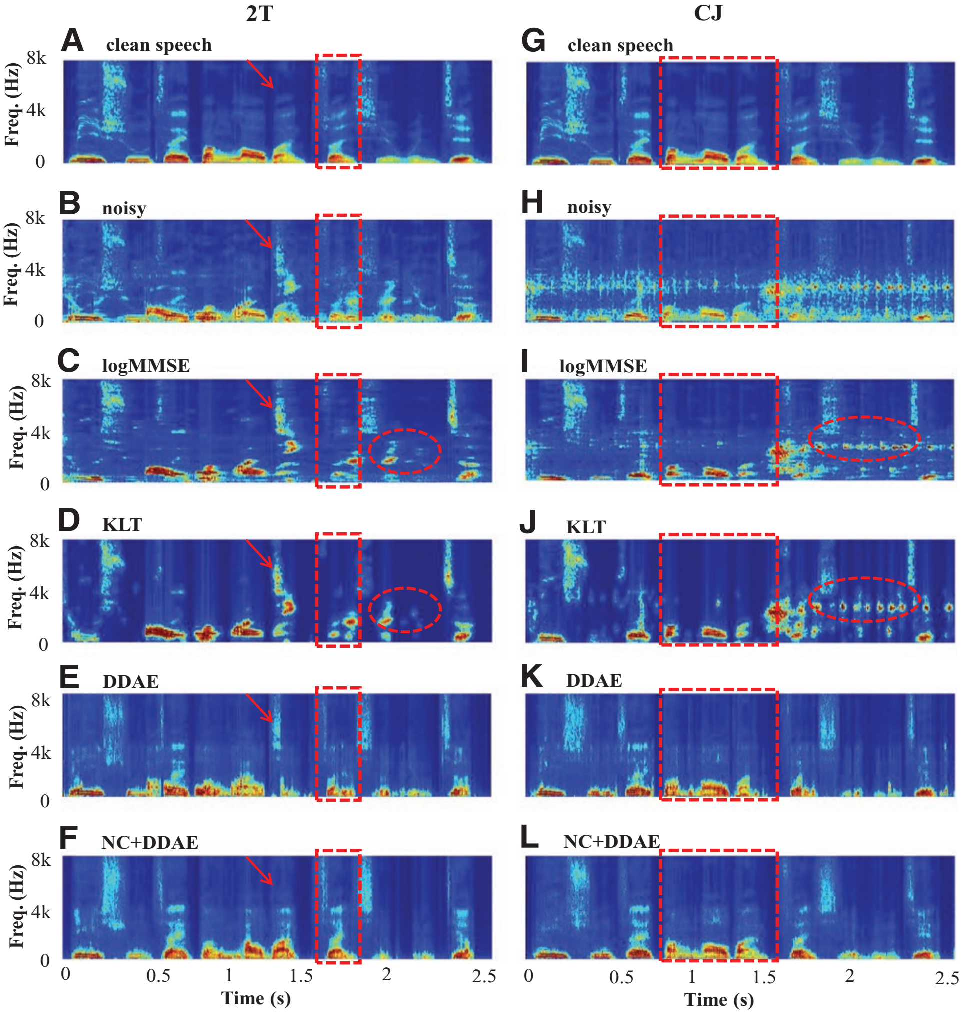

Fig. 8. Spectrograms of an utterance corrupted by two-talker babble (2T) at 0 dB signal to noise ratio (SNR) (left panels B to F) and construction jackhammer (CJ) at 0 dB SNR (right panels H to L). A and G are clean utterances; B and H are 2T and CJ noisy utterances, respectively; C to F and I to L show the sentences enhanced by logMMSE, KLT, DDAE and NC + DDAE, respectively. DDAE, deep denoising autoencoder; KLT, Karhunen–Loéve transform; logMMSE, log minimum mean squared error; NC, noise classifier. More samples available at: https://goo.gl/Gz8iLM.

the clinical effectiveness of the proposed NR approach for Mandarin-speaking CI recipients under noisy conditions with challenging noise types and SNR levels; (4) we compared the speech recognition performance of the DDAE and NC + DDAE approaches and two conventional single-microphone NR approaches under different noisy conditions. From the results shown in Table 2, the NC module of the proposed approach had a good performance for noise identification for the seen testing conditions. The average accuracy of the NC module for the 12 noise types was 99.6% in the frame-based evaluation. Moreover, from Figure 4, we can note that the CM scores for these seen testing conditions are quite high. The results suggest that the NC module can accurately classify the noise types with high CM scores when the noise has been seen in training data. On the other hand, when dealing with unseen noise types (the 13th to 16th noise types in Fig. 4), the NC module yielded low CM scores. The objective NCM results shown in Figure 5 and listening test results shown in Figure 6 confirm the effectiveness of the NC + DDAE NR approach for nonstationary noise types.

Figure 6 shows the results of listening test in this study. The intelligibility scores of the proposed NR approaches were notably higher than those of noisy, logMMSE, and KLT, which was consistent with the results of a previous study (Lai et al. 2017) based on vocoded speech in CI simulations. The one-way ANOVA results in Table 3 further confirm that NC + DDAE NR provided significantly higher recognition scores than those of noisy, logMMSE, and KLT in most testing conditions. The results suggest that the NC + DDAE NR approach could potentially be integrated into a CI processor to achieve better speech perception for CI recipients under challenging noisy conditions.

In Figure 6, the conventional NR approaches did not improve the CI recipients' speech recognition rates under challenging noisy conditions, which was consistent with previous findings (Walden et al. 2000; Ricketts et al. 2001; Ricketts & Hornsby 2005; Bentler et al. 2008). For a conventional NR approach, a noise estimation algorithm is generally required to compute the noise statistics. The noise estimation algorithm plays an important role in the quality of enhanced speech. If noise estimation is inaccurate, interfering residual noise or speech distortion may occur, which may result in a loss of speech intelligibility. Over the past two decades, numerous noise estimation algorithms have been proposed, such as voice activity detection (Sohn et al. 1999), a minima controlled recursive algorithm (Cohen & Berdugo 2002), noise power spectral density estimation (Martin 2001), and minima controlled recursive algorithm version 2 (Rangachari & Loizou 2006). Although these noise estimation algorithms provide satisfactory NR performance for stationary noise, most of them may not perform very well when dealing with challenging noise types. (e.g., babble noise) (Rangachari & Loizou 2006). Consequently, these classical noise estimation algorithms may fail to track fast-changing noise in real-world scenarios (Xu et al. 2015).

When compared to conventional NR techniques, the NC + DDAE NR approach can be regarded as a data-driven, supervised learning method. This approach involves learning from the mapping function from noisy to clean speech signals without imposing any assumptions. Subsequently, based on the mapping function of the NC + DDAE model, noisy speech is directly transformed into clean speech without using any noise estimation algorithm. Accordingly, the intelligibility performance of NC + DDAE is higher than that of conventional NR even when dealing with difficult, competing noises, or with a 0-dB SNR.

In this study, we used 12 types of noise to train the NC + DDAE models and selected the corresponding and most appropriate DDAE model for CI recipients based on the NC results. In other words, we assumed that prior information, such as the speaker's identity, was provided in the experiment scenario. The results for speech recognition, as shown in Figure 6, demonstrate the effectiveness of the NC + DDAE NR approach for various noisy conditions under different SNRs. In the future, we intend to further investigate the efficacy of the NC + DDAE NR approach when the speaker's identity is not known, similar to the study by Goehring et al. (2017). Additionally, although the benefits of the NC + DDAE NR approach outweigh those of the conventional NR approaches, there is still room for improvement. Auxiliary features, such as fundamental frequency cues (Chen et al. 2014), pitch (Chen et al. 2010, 2014), band importance function (ANSI 1997), and incorporating the dynamic range of Mandarin speech (Lai et al. 2013) to achieve better intelligibility and sound quality for enhanced speech (Xu et al. 2015) could provide further benefits. More recently, researchers have started to work on multi-tiered acoustic noises using the deep learning–based models, such as Williamson and Wang (2017). Based on the results of the present study, we will further explore the capability of the deep learning–based approach on testing conditions with multi-tiered acoustic noise in the future.

The computational complexity is currently a critical issue for DNN-based NR to be implemented in a CI device. Because of its multiple-layer structure, a DNN model requires a large quantity of memory storage and high computational costs at runtime. Therefore, it is highly demanding to simplify the architecture of the DNN model by reducing the online computations while maintaining its performance. More recently, many approaches have been developed to attain highly reconfigurable and energy-efficiency DNN-based processors to support various pattern classification and regression tasks (Bang et al. 2017; Bong et al. 2017; Desoli et al. 2017; Moons et al. 2017; Price et al. 2017; Shin et al. 2017; Whatmough et al. 2017). Meanwhile, endeavors have been made to address the issue of high computation costs. For example, the distillation approach (Hinton et al. 2015) involves transferring the knowledge from a complicated model to a simplified model that is more suitable for deployment. Another well-known approach is the binarization of the parameters in a deep learning–based model for memory size and access reductions (Courbariaux et al. 2016). With the rapid advances in deep learning algorithms and hardware, we believe that the proposed NC + DDAE NR approach can be realized in CI devices in the near future. Moreover, several systems have been developed to integrate CIs with other devices, such as smartphone, television, or MP3 player. These devices generally can provide superior computation power and storage to CIs, and thus, the issue of the high computation cost of the proposed DNN-based NR approach can be well addressed.

In conclusion, this study investigated the effectiveness of the NC + DDAE NR approach and compared its performance with two well-known single-microphone (i.e., logMMSE and KLT) strategies for the intelligibility of Mandarin speech under challenging noisy conditions. Our results showed that NC + DDAE NR provided superior noise suppression compared to that of conventional NR methods and less distortion of speech envelope information for Mandarin CI recipients. Moreover, the intelligibility of NC + DDAE processed sentences was significantly better than that of sentences processed by conventional NR approaches. In terms of applicability, the proposed NC + DDAE approach is a practical solution for a CI used in real-world noisy conditions. Because a CI is a personal device, these *N* noise types can be specified by users off-line, and the ND-DDAE models can be well trained with a sufficient amount of training data. When the testing condition is under a noise type that belongs to one of these *N* noise types, then a matched DDAE model will be selected to perform NR. On the other hand, if the testing noise is unseen, then the NI-DDAE model is directly used to perform NR. The experimental results from the present study and our previous study (Lai et al. 2017) have both confirmed that NI-DDAE can also yield better performance than conventional NR approaches, showing that the proposed NC + DDAE approach performs well even under unseen noise types and could potentially be implemented in CI speech processors to provide benefits to CI recipients.

ACKNOWLEDGMENTS

This work was supported by the Ministry of Science and Technology, Taiwan, under Grant MOST 105-2314-B-350-001, Grant MOST 106-2314-B-350-002, Grant MOST 105-2218-E-155-014-MY2, Grant MOST 106-2221-E-010-021, Grant MOST105-2218-E-001-004, and Grant CHGH 105-11.

The authors have no conflicts of interest to disclose.

This work was also supported by the National Natural Science Foundation of China (Grant No. 61571213).

Address for correspondence: Lieber Po-Hung Li, Department of Otolaryngology, Cheng Hsin General Hospital, Taipei, Taiwan, No. 45, Cheng Hsin Street, Pai-Tou, Taipei. E-mail: lieber.chgh@gmail.com

Received January 10, 2017; accepted November 1, 2017.

REFERENCES

ANSI. (1997). S3. 5-1997, Methods for the calculation of the speech intelligibility index. New York, NY: American National Standards Institute, *19*, (pp. 90–119).

Bang, S., Wang, J., Li, Z., et al. (2017). 14.7 A 288 µW programmable deep-learning processor with 270 KB on-chip weight storage using non-uniform memory hierarchy for mobile intelligence. *Proc Int Solid-State Circuits Conference* (pp. 250–251). San Francisco: IEEE.

Bengio, Y. (2009). Learning deep architectures for AI. In M. Gordan (Ed), *Foundations and Trends® in Machine Learning*, *2*, (pp. 1–127). University of California, Berkeley: James Finlay.

Bentler, R., Wu, Y. H., Kettel, J., et al. (2008). Digital noise reduction: Outcomes from laboratory and field studies. *Int J Audiol*, *47*, 447–460.

Bong, K., Choi, S., Kim, C., et al. (2017). 14.6 A 0.62 mW ultra-low-power convolutional-neural-network face-recognition processor and a CIS integrated with always-on haar-like face detector. *Proc Int Solid-State Circuits Conference* (pp. 248–249). San Francisco: IEEE.

Buechner, A., Brendel, M., Saalfeld, H., et al. (2010). Results of a pilot study with a signal enhancement algorithm for HiRes 120 cochlear implant users. *Otol Neurotol*, *31*, 1386–1390.

Buechner, A., Dyballa, K. H., Hehrmann, P., et al. (2014). Advanced beamformers for cochlear implant users: acute measurement of speech perception in challenging listening conditions. *PLoS One*, *9*, e95542.

Chen, F. (2012). Predicting the intelligibility of cochlear-implant vocoded speech from objective quality measure. *J Med Biol Eng*, *32*, 189–194.

Chen, F., & Lau, A. H. (2014). Effect of vocoder type to Mandarin speech recognition in cochlear implant simulation. *Proc ISCSLP* (pp. 551–554). Singapore, Singapore: IEEE.

Chen, F., & Loizou, P. C. (2011). Predicting the intelligibility of vocoded and wideband Mandarin Chinese. *J Acoust Soc Am*, *129*, 3281–3290.

Chen, F., Hu, Y., Yuan, M. (2015). Evaluation of noise reduction methods for sentence recognition by Mandarin-speaking cochlear implant listeners. *Ear Hear*, *36*, 61–71.

Chen, F., Wong, L. L., Hu, Y. (2014). Effects of lexical tone contour on Mandarin sentence intelligibility. *J Speech Lang Hear Res*, *57*, 338–345.

Chen, F., Wong, L. L., Qiu, J., et al. (2013). The contribution of matched envelope dynamic range to the binaural benefits in simulated bilateral electric hearing. *J Speech Lang Hear Res*, *56*, 1166–1174.

Chen, J., Wang, Y., Yoho, S. E., et al. (2016). Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises. *J Acoust Soc Am*, *139*, 2604.

Chen, J. K. C., Chuang, A. Y. C., McMahon, C., et al. (2010). Music training improves pitch perception in prelingually deafened children with cochlear implants. *Pediatrics*, *125*, 793–800.

Chen, T. E., Yang, S. I., Ho, L. T., et al. (2017). S1 and S2 heart sound recognition using deep neural networks. *IEEE Trans Biomed Eng*, *64*, 372–380.

Chung, K. (2004). Challenges and recent developments in hearing aids. Part I. Speech understanding in noise, microphone technologies and noise reduction algorithms. *Trends Amplif*, *8*, 83–124.

Cohen, I. (2003). Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. *IEEE Trans Speech Audio Process*, *11*, 466–475.

Cohen, I., & Berdugo, B. (2002). Noise estimation by minima controlled recursive averaging for robust speech enhancement. *IEEE Signal Proc Lett*, *9*, 12–15.

Courbariaux, M., Hubara, I., Soudry, D. et al. (2016). Binarized neural networks: Training deep neural networks with weights and activations constrained to+1 or-1. *arXiv preprint arXiv*, 1602.02830.

Davis, S. & Mermelstein P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans Acoust Speech Signal Process*, *28*(4), 357–366.

Dawson, P. W., Mauger, S. J., Hersbach, A. A. (2011). Clinical evaluation of signal-to-noise ratio-based noise reduction in Nucleus® cochlear implant recipients. *Ear Hear*, *32*, 382–390.

Desoli, G., Chawla, N., Boesch, T., et al. (2017). 14.1 A 2.9 TOPS/W deep convolutional neural network SoC in FD-SOI 28 nm for intelligent embedded systems. *Proc Int Solid-State Circuits Conference* (pp. 238–239).

Dien, J. (1998). Issues in the application of the average reference: Review, critiques, and recommendations. *Behav Res Meth Instrum Comput*, *30*, 34–43.

Dorman, M. F., Loizou, P. C., Rainey, D. (1997). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *J Acoust Soc Am*, *102*, 2403–2411.

Du, J., & Huo, Q. (2008). A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions. *Proc Interspeech*, 569–572.

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

Ephraim, Y., & Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans Acoustics Speech Signal Process*, *33*, 443–445.

Ephraim, Y., & Van Trees, H. L. (1995). A signal subspace approach for speech enhancement. *IEEE Trans Speech Audio Process*, *3*, 251–266.

Fetterman, B. L., & Domico, E. H. (2002). Speech recognition in background noise of cochlear implant patients. *Otolaryngol Head Neck Surg*, *126*, 257–263.

Firszt, J. B., Holden, L. K., Reeder, R. M., et al. (2009). Speech recognition in cochlear implant recipients: Comparison of standard HiRes and HiRes 120 sound processing. *Otol Neurotol*, *30*, 146–152.

Friesen, L. M., Shannon, R. V., Baskent, D., et al. (2001). Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants. *J Acoust Soc Am*, *110*, 1150–1163.

Fu, S. W., Tsao, Y., Lu, X. (2016). SNR-aware convolutional neural network modeling for speech enhancement. *Proc Interspeech*, 3768–3772.

Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Trans Acoustics Speech Signal Process*, *29*, 254–272.

Glorot, X., Bordes, A., Bengio, Y. (2011). Deep sparse rectifier neural networks. *Proc* 14th International Conference on Artificial Intelligence and Statistics (pp. 315–323).

Goehring, T., Bolner, F., Monaghan, J. J., et al. (2017). Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users. *Hear Res*, *344*, 183–194.

Hamacher, V., Doering, W. H., Mauer, G., et al. (1997). Evaluation of noise reduction systems for cochlear implant users in different acoustic environment. *Am J Otol*, *18*(6 Suppl), S46–S49.

Haykin, S. (1995). *Advances in Spectrum Analysis and Array Processing (vol. III)*. New Jersey: Prentice-Hall, Inc.

Hersbach, A. A., Arora, K., Mauger, S. J., et al. (2012). Combining directional microphone and single-channel noise reduction algorithms: a clinical evaluation in difficult listening conditions with cochlear implant users. *Ear Hear*, *33*, e13–e23.

Hersbach, A. A., Grayden, D. B., Fallon, J. B., et al. (2013). A beamformer post-filter for cochlear implant noise reduction. *J Acoust Soc Am*, *133*, 2412–2420.

Hinton, G., Deng, L., Yu, D., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *Vol. 29*, pp. 82–97.

Hinton, G., Vinyals, O., Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv*, 1503.02531.

Hinton, G. E., Osindero, S., Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Comput*, *18*, 1527–1554.

Hochberg, I., Boothroyd, A., Weiss, M., et al. (1992). Effects of noise and noise suppression on speech perception by cochlear implant users. *Ear Hear*, *13*, 263–271.

Holden, L. K., Finley, C. C., Firszt, J. B., et al. (2013). Factors affecting open-set word recognition in adults with cochlear implants. *Ear Hear*, *34*, 342–360.

Hu, Y., & Loizou, P. C. (2003). A generalized subspace approach for enhancing speech corrupted by colored noise. *IEEE Trans Speech Audio Process*, *11*, 334–341.

Hu, Y., Loizou, P. C., Li, N., et al. (2007). Use of a sigmoidal-shaped function for noise attenuation in cochlear implants. *J Acoust Soc Am*, *122*, 128–134.

Huang, M. w. (2005). Development of Taiwan Manadarin hearing in noise test. Master's Thesis in *Department of Speech Language Pathology and Audiology*. National Taipei University of Nursing and Health science,Taipei, Taiwan.

Ideas, S. (2002). Sample CD: XV MP3 Series SI-XV-MP3. Inc.

Jiang, H. (2005). Confidence measures for speech recognition: A survey. *Speech Commun*, *45*, 455–470.

Khing, P. P., Swanson, B. A., Ambikairajah, E. (2013). The effect of automatic gain control structure and release time on cochlear implant speech intelligibility. *PLoS One*, *8*, e82263.

Kolbæk, M., Tan, Z. H., Jensen, J. (2017). Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems. *IEEE/ACM Trans Audio Speech Lang Process*, *25*, 153–167.

Lai, Y. H., Chen, F., Wang, S. S., et al. (2017). A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation. *IEEE Trans Biomed Eng*, *64*, 1568–1578.

Lai, Y. H., Liu, T. C., Li, P. C., et al. (2013). Development and preliminary verification of a Mandarin-based hearing-aid fitting strategy. *PLoS One*, *8*, e80831.

Lai, Y. H., Tsao, Y., Chen, F. (2015). Effects of adaptation rate and noise suppression on the intelligibility of compressed-envelope based speech. *PLoS One*, *10*, e0133519.

Loizou, P. C. (1999). Introduction to cochlear implants. *IEEE Eng Med Biol Mag*, *18*, 32–42.

Loizou, P. C. (2006). Speech processing in vocoder-centric cochlear implants. *Adv Otorhinolaryngol*, *64*, 109–143.

Loizou, P. C. (2013). *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC press.

Loizou, P. C., Lobo, A., Hu, Y. (2005). Subspace algorithms for noise reduction in cochlear implants. *J Acoust Soc Am*, *118*, 2791–2793.

Lu, X., Tsao, Y., Matsuda, S., et al. (2013). Speech enhancement based on deep denoising autoencoder. *Proc Interspeech* (pp. 436–440). International Speech Communication Association.

Lu, X., Tsao, Y., Matsuda, S., et al. (2014). Ensemble modeling of denoising autoencoder for speech spectrum restoration. *Proc Interspeech* (pp. 885–889). International Speech Communication Association.

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

Ma, L., Milner, B., Smith, D. (2006). Acoustic environment classification. *ACM Trans Speech Lang Process*, *3*, 1–22.

Margo, V., Terry, M., Schweitzer, C., et al. (1995). Results of take home trial for a nonlinear beamformer used as a noise reduction strategy for cochlear implants. *J Acoust Soc Am*, *98*, 2984–2984.

Martin, R. (2001). Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans Speech Audio Process*, *9*, 504–512.

Mengusoglu, E., & Ris, C. (2001). Use of acoustic prior information for confidence measure in ASR applications. Proceedings of Eurospeech 2001 (pp. 2557–2560). http://research.org/publication/interspeech-2001.

Mittal, U., & Phamdo, N. (2000). Signal/noise KLT based approach for enhancing speech degraded by colored noise. *IEEE Trans Speech Audio Process*, *8*, 159–167.

Mohamed, A. R., Dahl, G. E., Hinton, G. (2012). Acoustic modeling using deep belief networks. *IEEE Trans Audio Speech Lang Process*, *20*, 14–22.

Moons, B., Uytterhoeven, R., Dehaene, W., et al. (2017). 14.5 Envision: A 0.26-to-10 TOPS/W subword-parallel dynamic-voltage-accuracy-frequency-scalable Convolutional Neural Network processor in 28nm FDSOI. *Proc Int Solid-State Circuits Conference* (pp. 246–247). San Francisco: IEEE.

Narayanan, A., & Wang, D. (2013). Ideal ratio mask estimation using deep neural networks for robust speech recognition. *Proc Int Conf Acoust Speech Signal Process*, 7092–7096. Vancouver, BC, Canada: IEEE.

NIDCD (National Institute on Deafness and other Communication Disorders), NIH (2014). Cochlear Implants. Retrieved from http://www.nidcd.nih.gov/health/hearing/pages/coch.aspx.

Nie, K., Stickney, G., Zeng, F. G. (2005). Encoding frequency modulation to improve cochlear implant performance in noise. *IEEE Trans Biomed Eng*, *52*, 64–73.

Peterson, G. E., & Lehiste, I. (1962). Revised CNC lists for auditory tests. *J Speech Hear Disord*, *27*, 62–70.

Price, M., Glass, J., Chandrakasan, A. P. (2017). 14.4 A scalable speech recognizer with deep-neural-network acoustic models and voice-activated power gating. *Proc Int Solid-State Circuits Conference* (pp. 244–245). San Francisco: IEEE.

Rabiner, L., & Juang, B. H. (1993). Fundamentals of speech recognition. *Prentice Hall Signal Processing Series*.

Rangachari, S., & Loizou, P. C. (2006). A noise-estimation algorithm for highly non-stationary environments. *Speech Commun*, *48*, 220–231.

Räsänen, O., Leppänen, J., Laine, U. K., et al. (2011). Comparison of classifiers in audio and acceleration based context classification in mobile phones. *IEEE Signal Processing Conference* (pp. 946–950). Barcelona, Spain: IEEE.

Rezayee, A., & Gazor, S. (2001). An adaptive KLT approach for speech enhancement. *IEEE Trans Speech Audio Process*, *9*, 87–95.

Ricketts, T., Lindley, G., Henry, P. (2001). Impact of compression and hearing aid style on directional hearing aid benefit and performance. *Ear Hear*, *22*, 348–361.

Ricketts, T. A., & Hornsby, B. W. (2005). Sound quality measures for speech in noise through a commercial hearing aid implementing. *J Am Acad Audiol*, *16*, 270–277.

Scalart, P. (1996). Speech enhancement based on a priori signal to noise estimation. *Proc Int Conf Acoust Speech Signal Process*, *2*, 629–633.

Schmidt, R. O. (1986). Multiple emitter location and signal parameter estimation. *IEEE Trans Antennas Propag*, *34*, 276–280.

Shannon, R. V., Zeng, F. G., Kamath, V., et al. (1995). Speech recognition with primarily temporal cues. *Science*, *270*, 303–304.

Shin, D., Lee, J., Lee, J., et al. (2017). 14.2 DNPU: An 8.1 TOPS/W reconfigurable CNN-RNN processor for general-purpose deep neural networks. *Proc Int Solid-State Circuits Conference* (pp. 240–241). San Francisco: IEEE.

Skinner, M. W., Arndt, P. L., Staller, S. J. (2002). Nucleus® 24 Advanced Encoder conversion study: Performance versus preference. *Ear Hear*, *23*, 2–17.

Sohn, J., Kim, N. S., Sung, W. (1999). A statistical model-based voice activity detection. *IEEE Signal Proc Lett*, *6*, 1–3.

Stickney, G. S., Zeng, F. G., Litovsky, R., et al. (2004). Cochlear implant speech recognition with speech maskers. *J Acoust Soc Am*, *116*, 1081–1091.

Studebaker, G. A. (1985). A "rationalized" arcsine transform. *J Speech Hear Res*, *28*, 455–462.

Tsao, Y., & Lai, Y. H. (2016). Generalized maximum a posteriori spectral amplitude estimation for speech enhancement. *Speech Commun*, *76*, 112–126.

van Hoesel, R., Böhm, M., Battmer, R. D., et al. (2005). Amplitude-mapping effects on speech intelligibility with unilateral and bilateral cochlear implants. *Ear Hear*, *26*, 381–388.

Walden, B. E., Surr, R. K., Cord, M. T., et al. (2000). Comparison of benefits provided by different hearing aid technologies. *J Am Acad Audiol*, *11*, 540–560.

Wang, D., & Chen J. (2017). Supervised speech separation based on deep learning: an overview. *arXiv preprint arXiv*, 1708.07524.

Wang, Y., Narayanan, A., Wang, D. (2014). On training targets for supervised speech separation. *IEEE/ACM Trans Audio Speech Lang Process*, *22*, 1849–1858.

Weiss, M., Aschkenasy, E., Parsons, T. (1975). Study and development of the INTEL technique for improving speech intelligibility. *Final Rep. NSC-FR/4023, Nicolet Scientific Corp., December*.

Weninger, F., Erdogan, H., Watanabe, S., et al. (2015). Speech enhancement with LSTM recurrent neural networks and its application to noise-rebust ASR. *International Conference on Latent Variable Anaysis and Signal Separation* (pp. 91–99). Springer, Cham.

Whatmough, P. N., Lee, S. K., Lee, H., et al. (2017). 14.3 A 28nm SoC with a 1.2 GHz 568nJ/prediction sparse deep-neural-network engine with> 0.1 timing error rate tolerance for IoT applications. *Proc Int Solid-State Circuits Conference*, (pp. 242–243). San Francisco: IEEE.

Williamson, D., & Wang, D. (2017). Time-frequency masking in the complex domain for speech dereverberation and denoising. *ACM Trans Speech Lang Process*, 25, 1492–1501.

Wilson, W. R., Byl, F. M., Laird, N. (1980). The efficacy of steroids in the treatment of idiopathic sudden hearing loss. A double-blind clinical study. *Arch Otolaryngol*, *106*, 772–776.

Wouters, J., & Vanden Berghe, J. (2001). Speech recognition in noise for cochlear implantees with a two-microphone monaural adaptive noise reduction system. *Ear Hear*, *22*, 420–430.

Xia, B., & Bao, C. (2014). Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification. *Speech Commun*, *60*, 13–29.

Xu, Y., Du, J., Dai, L. R., et al. (2015). A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Trans Audio Speech Lang Process*, *23*, 7–19.

Xu, Y., Du, J., Huang, Z., et al. (2017). Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement. *arXiv preprint arXiv*, 1703.07172.

Yu, D., & Deng, L. (2012). *Automatic Speech Recognition*. Springer.

Yu, D., & Deng, L. (2014). *Automatic Speech Recognition: A Deep Learning Approach*. New York: Springer.

Zeng, F. G., & Shannon, R. V. (1999). Psychophysical laws revealed by electric hearing. *Neuroreport*, *10*, 1931–1935.

Zeng, F. G., Grant, G., Niparko, J., et al. (2002). Speech dynamic range and its effect on cochlear implant performance. *J Acoust Soc Am*, *111*(1 Pt 1), 377–386.