A Deep Learning based Noise Reduction Approach to Improve Speech Intelligibility for Cochlear Implant Recipients in the Presence of Competing Speech Noise

Syu-Siang Wang[†], Yu Tsao[†], Hsiao-Lan Sharon Wang[§], Ying-Hui Lai[*], and Lieber Po-Hung Li[₪,‡]

[†]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
E-mail: {sypdbhee; yu.tsao}@citi.sinica.edu.tw

[§]Department of Special Education, National Taiwan Normal University, E-mail: hlw36@ntnu.edu.tw

[*]Department of Biomedical Engineering, National Yang-Ming University, Taiwan, E-mail: yh.lai@ym.edu.tw

[₪]Department of Otolaryngology, Cheng Hsin General Hospital, Taipei, Taiwan, E-mail: lieber.chgh@gmail.com

[‡]Faculty of Medicine, School of Medicine, National Yang-Ming University, Taipei, Taiwan

***Abstract*—This paper presents the clinical results of the application of a deep-learning-based noise reduction (NR) approach to improve speech intelligibility for cochlear implant (CI) recipients in the presence of competing speech noise. The deep denoising autoencoder (DDAE) model was used as a representative deep-learning-based NR model to reduce the noise components from the noisy input. The enhanced speech was subsequently played to six Mandarin-speaking CI recipients to perform recognition tests. All the subjects used their own clinical speech processors during testing. Two traditional NR approaches were also implemented to test the performance for a comparison. The Taiwan Mandarin version of the hearing in noise test (TMHINT) sentences were adopted and further corrupted by competing two talker speech noise at signal-to-noise ratio (SNR) levels of 0 and 5 dB. The experimental results showed that the DDAE NR approach can yield higher intelligibility scores than the two classical NR techniques in the presence of competing speech. The results of qualitative analysis further showed that the DDAE NR approach notably reduced the envelope distortions. The good results also suggest that the proposed DDAE NR approach can combine well with the existing CI processors to overcome the issue of degradation of speech perception, which is caused by competing speech noise.**

I. INTRODUCTION

In a cochlear implant (CI) device, a noise reduction (NR) unit is usually adopted to process the input speech in order to provide enhanced speech with high intelligibility and quality. In the past, various NR techniques have been proposed, such as log minimum mean squared error (logMMSE) [1], Wiener filter [2], generalized maximum a posteriori spectral amplitude [3] estimation, and Karhunen-Loéve transform (KLT) [4, 5]. Most of these NR techniques were developed by exploring the statistical properties of speech and noise signals [6]. Although these NR approaches achieved satisfactory performance in stationary noisy environments, their performances may be notably degraded in non-stationary noisy environments, where the acoustic statistics for each small time period were varied dramatically [7]. In a previous study, several NR methods were evaluated with Mandarin CI recipients [8]. Even though the results indicated that these NR approaches can provide notable benefits to CI recipients under stationary noise, there is a significant scope for the performance of NR to be prompted under challenging listening conditions.

Recently, deep-learning-based NR approaches were developed and they exhibited outstanding performance in various NR tasks [7, 9-15]. These approaches adopt multiple layers of non-linear transformation to characterize the mapping function from noisy to clean speech signals. The deep denoising autoencoder (DDAE) is a well-known deep-learning-based NR approach [10]. Previous studies have confirmed that the DDAE NR approach outperforms conventional NR approaches [e.g., minimum mean squared error (MMSE) plus a standard noise tracking approach [16] in terms of several standardized objective evaluations [10]. More recently, Lai et al. [17] tested the performance of DDAE NR with vocoded speech, which was derived to simulate the CI recipients. The results showed that the DDAE NR approach outperforms conventional NR techniques in terms of both objective evaluations and subjective listening tests with normal hearing participants under non-stationary noise conditions. In the present study, we further investigate the clinical effectiveness of the DDAE NR approach with real CI recipients under challenging noise types and signal-to-noise ratio (SNR) levels.

II. DDAE-BASED NR APPROACH

Given the speech signal, $\boldsymbol{x}$, and noise signal, $\boldsymbol{n}$, the noisy signal, $\boldsymbol{y}$, can be formulated as:

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{n}. \quad (1)$$

The goal of NR is to estimate the enhanced speech signals $\hat{\boldsymbol{x}}$ from $\boldsymbol{y}$, where $\hat{\boldsymbol{x}}$ is close to $\boldsymbol{x}$. A class of conventional NR

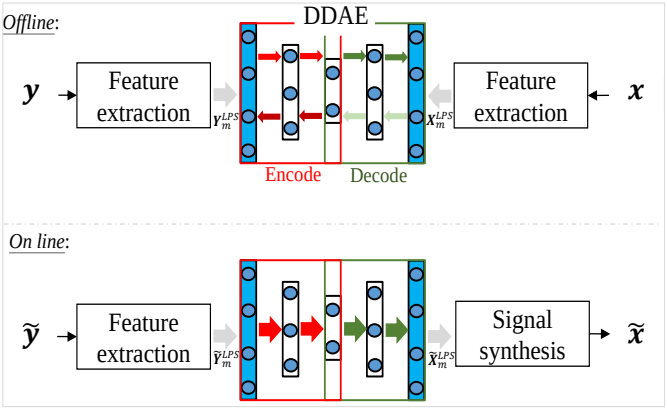

Figure 1. DDAE-based NR system.

approaches design filters with the aim to minimize a specific distortion measure between the original noise-free signal and the enhanced counterpart, such as logMMSE [1] and Wiener filter [2]. Another class of approaches divides the noisy signal into two subspaces (clean and noise) and subsequently minimizes the noise components appearing in the clean subspace. Well-known techniques include singular value decomposition [18] and KLT [4, 5]. In the present study, we conducted experiments using logMMSE and KLT for comparison with the DDAE NR approach. These approaches have been confirmed to yield satisfactory speech intelligibility improvements for Mandarin CI patients under noisy conditions effectively [8].

The overall structure of the DDAE NR model is illustrated in Fig. 1. There are two phases in the DDAE NR approach: offline and online. In the offline phase, a set of noisy-clean speech pairs is prepared. Both the noisy and clean speech signals are first converted into log power spectrum (LPS) features. The noisy ($\boldsymbol{Y}_m^{LPS}$) and clean ($\boldsymbol{X}_m^{LPS}$) LPS features are subsequently placed at the input and output sides of the DDAE model, where m denotes the frame index. For a DDAE model with L hidden layers, we obtain:

$$\begin{aligned} h^1(\boldsymbol{Y}_m^{LPS}) &= \sigma(\boldsymbol{W}^1\boldsymbol{Y}_m^{LPS} + \boldsymbol{b}^1), \\ &\vdots \\ h^L(\boldsymbol{Y}_m^{LPS}) &= \sigma(\boldsymbol{W}^{L-1}h^{L-1}(\boldsymbol{Y}_m^{LPS}) + \boldsymbol{b}^{L-1}), \\ \widehat{\boldsymbol{X}}_m^{LPS} &= \boldsymbol{W}^L h^L(\boldsymbol{Y}_m^{LPS}) + \boldsymbol{b}^L, \end{aligned} \tag{2}$$

where $\{\boldsymbol{W}^1 \ldots \boldsymbol{W}^L\}$ are the matrices of the weights; $\{\boldsymbol{b}^1 \ldots \boldsymbol{b}^L\}$ are the bias vectors; $\widehat{\boldsymbol{X}}_m^{LPS}$ is the vector that contains the LPS features of restored speech corresponding to the noisy input $\boldsymbol{Y}_m^{LPS}$; $\sigma(\cdot)$ denotes the activation function, and the logistic function is used in this study. Finally, the parameters θ (including the matrices of the weights and the bias vectors) are determined by optimizing the following objective functions:

$$\begin{aligned} \theta^* &= \arg\min_{\theta}(F(\theta) + \eta^1\|\boldsymbol{W}^1\|_2^2 + \cdots + \eta^L\|\boldsymbol{W}^L\|_2^2), \\ F(\theta) &= \tfrac{1}{M}\textstyle\sum_{m=1}^{M}\left\|\boldsymbol{X}_m^{LPS} - \widehat{\boldsymbol{X}}_m^{LPS}\right\|_2^2, \end{aligned} \tag{3}$$

where M is the total number of training samples (noisy-clean pairs). In the online phase, the DDAE transforms the noisy speech signal ($\boldsymbol{Y}_m^{LPS}$) into an enhanced speech signal ($\widehat{\boldsymbol{X}}_m^{LPS}$). More detailed introduction of the DDAE NR approach can be found in [10].

III. Experiments and Results

A. Experimental Setup

Experimental results were conducted using the Taiwan Mandarin version of the hearing in noise test (TMHINT) [19], which were recorded at the sampling rate of 16,000 Hz by a native male Taiwan speaker in a soundproof room. In TMHINT, there are 320 clean utterances with ten Chinese characters in each utterance. Among these clean utterances, we selected 120 and 200 different utterances to form the training and testing sets, respectively. The noise from two equal-level interfering female talkers (2T) was artificially added to both the clean training and testing sets under −10, −5, −3, 0, 3, 5, and 10 dB SNR noise conditions to obtain noisy data. Moreover, the noisy testing utterances derived under the SNR of 0 and 5 dB were further extracted as the final testing set for the subsequent evaluations. Therefore, 840 training noisy-clean speech pairs and 400 testing noisy utterances were generated in this task.

For the feature extraction in Fig. 1, the frame length of 16 ms and the hop size of 8 ms were selected to window the input waveform in a series of frames. Subsequently, the 129-point log spectrogram was performed by applying the 256-point fast Fourier transform to these frames in the time domain. In order to achieve the same dimensions as the 129-point log spectrum, there were 129 nodes in both the input and output layers of the DDAE NR model, which contained three hidden layers with 300 neurons in each layer. The subjects were recruited from Cheng Hsin General Hospital, Taipei, Taiwan. The study procedures were approved by the Institutional Review Board at the hospital, and informed consents were obtained from all the subjects before testing. Six native Taiwan Mandarin-speaking CI recipients participated in this study. Each recruited subject received a personalized CI unit with clinical speech processors and used it for more than 7 months. Among these subjects, two used the Advanced Bionics (AB) HiRes-120 sound coding strategy [20], and the other four used the cochlear advanced combination encoder (ACE) sound coding strategy [21]. All the subjects were required to participate in eight testing conditions (1 noise type (2T) × 2 SNR levels (0 and 5 dB) × 4 signal processing strategies (noisy, logMMSE, KLT, and DDAE)) with each condition containing 10 sentences. Notably, none of these utterances was repeated across the testing conditions. Moreover, the order of the eight conditions was also randomized for each subject. The subjects were instructed to repeat what they heard, and were allowed to hear the stimuli twice. The word correct rate (WCR) listed in eq. (4) was used as the evaluation metric, which was calculated by dividing the number of correctly identified words W_C by the total number of words under each testing condition W_T.

$$\text{WCR} = W_C/W_T \times \%. \tag{4}$$

There was a five-minute break for each subject after undergoing testing for 30 min. During the testing time, the built-in NR functions of the speech processor in the personalized CI unit of each subject were temporarily disabled. In this study, subjective tests were conducted using a double blind method [22] in a soundproof room (provided by Acoustic System, Inc.), where a notebook was equipped and connected to a GSI Audiostar Pro audiometry device (GSI, MN, USA).

B. Comparison of Spectrograms

In this section, we intend to qualitatively compare the clean, noisy, and enhanced speech using the logMMSE, KLT, and DDAE NR approaches. The spectrogram plot, which displays the spectral-temporal representations of a time-varying signal, is used for analysis. In Fig. 2, we illustrate the spectrograms of 2T noisy (0 dB SNR), clean, and enhanced sentences in (A) to (E), where (C), (D), (E) are the sentences enhanced by logMMSE, KLT, and DDAE, respectively. The sentence was recorded by a native male Mandarin speaker, saying "*There is a calligraphy competition in this semester*." From Fig. 2, when comparing (B) with (C), (D), and (E), we can observe that all the enhanced techniques can effectively suppress the noise components whereas DDAE provides the best denoising capability. The reasons for the good performance are (1) maintaining sturdier high frequency signals (less distortion) and (2) more accurately removing the speech signals from the background (sometimes logMMSE and KLT failed, owing to similar characteristics of target speech and background speech). The results suggest that DDAE has a better capability to handle non-stationary noises.

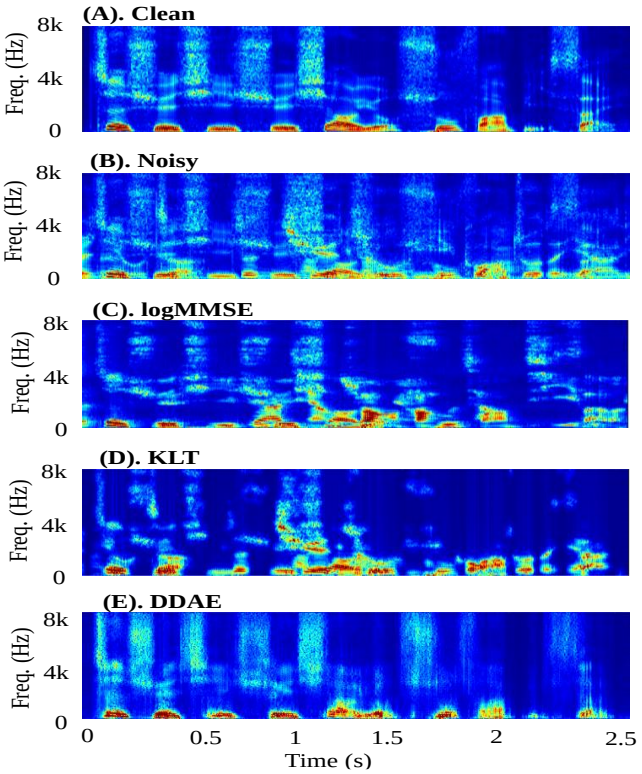

Figure 2. Spectrograms of clean, 2T noisy sentence (0 dB SNR), and enhanced sentences; (A) and (B) are clean and noisy speech, respectively; (C) to (E) show the sentences enhanced by logMMSE, KLT, and DDAE, respectively.

C. Example of the amplitude envelopes

Another useful qualitative analysis is based on the amplitude envelopes of speech signals [23]. Fig. 3 shows an example of the amplitude envelopes of speech signals—(A) clean, (B) noisy, and (C), (D), and (E) enhanced processing using the logMMSE, KLT, and DDAE approaches, respectively extracted from the third channel (with center frequency of 575 Hz), which is an important frequency band for speech intelligibility [24]. From Fig. 3, it can be observed that DDAE is more effective at suppressing noise components from noisy data and yields fewer envelop distortions [25] than the conventional NR approaches. Furthermore, when comparing with Fig. 3(A), large residual interferences can be observed in Figs. 3 (C, D) (e.g., around 0 to 0.1 s for logMMSE and 2.0 to 2.3 s for KLT). On the other hand, DDAE in Fig. 3(E) effectively removes the interferences at these segments. The results suggest that the DDAE NR approach provides superior speech intelligibility benefits for CI recipients.

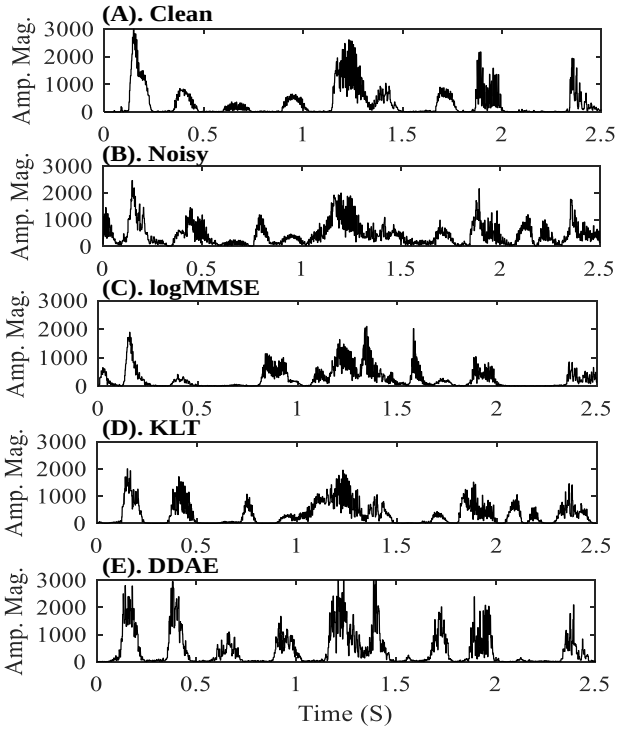

Figure 3. Amplitude envelopes extracted from the third channel (f_{center} = 575 Hz) of clean, 2T noisy sentence (0 dB SNR), and enhanced sentences; (A) and (B) are clean and noisy speech, respectively; (C) to (E) show the sentences enhanced by logMMSE, KLT, and DDAE, respectively.

D. Recognition scores by Mandarin CI recipients

Finally, we report the clinical results of listening tests performed on the six CI subjects. Fig. 4 shows the mean scores of noisy speech and the three enhancement approaches (i.e., logMMSE, KLT, and DDAE) tested on the six subjects in 2T noise at SNR levels of 0 dB and 5 dB. The performance is reported in terms of the averaged WCR. From the figure, it is evident that the DDAE NR approach achieves higher intelligibility scores than the noisy speech and enhanced speech processed by the conventional NR approaches. We further adopted the one-way analysis of variance (ANOVA) [26] and Tukey post-hoc comparisons to test the significance of the improvements. The ANOVA and Tukey post-hoc comparisons verified the significant differences for the following three group pairs at SNRs of 0 and 5 dB: DDAE and noisy; DDAE and logMMSE; DDAE and KLT.

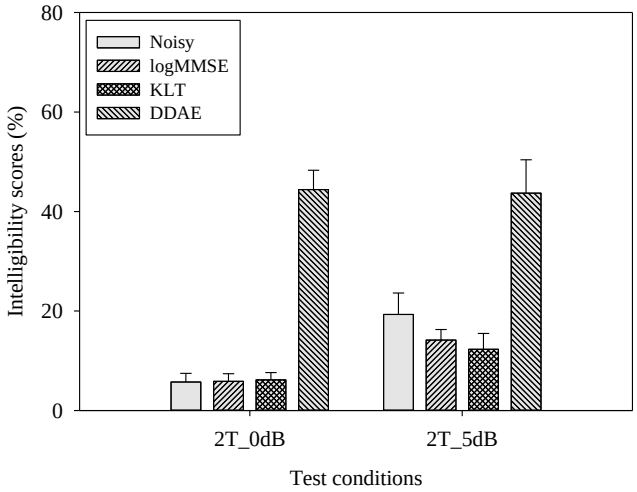

Figure 4. Mean intelligibility scores of two test conditions: 2T 0 dB SNR level (2T_0dB) and 2T 5dB SNR level (2T_5dB). The error bars denote the standard errors of mean values.

IV. Conclusions

This study investigated the performance of clinical listening tests using the DDAE NR approach for Mandarin CI recipients under competing speech noises. Two traditional methods, i.e., logMMSE and KLT, were also employed for comparison. The results show that the DDAE NR outperforms the conventional NR methods in terms of qualitative comparisons and subjective listening tests. These findings demonstrate that the DDAE, a deep learning NR approach, can be applied to CI users as a promising method for improving speech recognition performance.

V. Acknowledgement

The authors would like to thank the Ministry of Science and Technology for providing financial supports (MOST 105-2218-E-010-005-MY2, MOST 103-2420-H-003-008-MY3, MOST 105-2314-B-350-001-, and MOST 104-2221-E-001-026-MY2, MOST 106-2221-E-010-021), and thank iMediPlus Inc. for providing financial support (32T-1041126-1Q).

References

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, pp. 443-445, 1985.

[2] P. Scalart, "Speech enhancement based on a priori signal to noise estimation," *in Proc. ICASSP*, pp. 629-633, 1996.

[3] Y. Tsao and Y.-H. Lai, "Generalized maximum a posteriori spectral amplitude estimation for speech enhancement," *Speech Communication*, vol. 76, pp. 112-126, 2016.

[4] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing* , vol. 9, pp. 87-95, 2001.

[5] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 334-341, 2003.

[6] P. C. Loizou, *Speech Enhancement: Theory and Practice*: CRC press, 2013.

[7] Y. Xu, J. Du, L. R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 7-19, 2015.

[8] F. Chen, Y. Hu, and M. Yuan, "Evaluation of noise reduction methods for sentence recognition by Mandarin-speaking cochlear implant listeners," *Ear and Hearing*, vol. 36, pp. 61-71, 2015.

[9] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Ensemble modeling of denoising autoencoder for speech spectrum restoration," *in Proc. Interspeech*, pp. 885-889, 2014.

[10] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," *in Proc. Interspeech*, pp. 436-440, 2013.

[11] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 1849-1858, 2014.

[12] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," *in Proc. ICASSP*, pp. 7092-7096, 2013.

[13] S.-W. Fu, Y. Tsao, and X. Lu, "SNR-aware convolutional neural network modeling for speech enhancement," *in Proc. Interspeech*, pp. 3768-3772, 2016.

[14] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," *arXiv preprint arXiv:1703.02205*, 2017.

[15] M. Kolbæk, Z.-H. Tan and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, 2017.

[16] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 466-475, 2003.

[17] Y.-H. Lai, F. Chen, S.-S. Wang, X. Lu, Y. Tsao, and C.-H. Lee, "A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation," *IEEE Transactions on Biomedical Engineering*, vol. 64, pp. 1568-1578, 2017.

[18] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise: A regenerative approach," *Speech Communication*, vol. 10, pp. 45-57, 1991.

[19] M.-W. Huang, "Development of Taiwan Manadarin hearing in noise test," *Master Thesis, Department of speech language pathology and audiology, National Taipei University of Nursing and Health Science*, 2005.

[20] J. B. Firszt, L. K. Holden, R. M. Reeder, and M. W. Skinner, "Speech recognition in cochlear implant recipients: comparison of standard HiRes and HiRes 120 sound processing," *Otology & neurotology: official publication of the American Otological Society, American Neurotology Society [and] European Academy of Otology and Neurotology*, vol. 30, p. 146, 2009.

[21] M. W. Skinner, P. L. Arndt, and S. J. Staller, "Nucleus® 24 Advanced Encoder conversion study: Performance versus preference," *Ear and Hearing*, vol. 23, pp. 2S-17S, 2002.

[22] W. R. Wilson, F. M. Byl, and N. Laird, "The efficacy of steroids in the treatment of idiopathic sudden hearing loss: a double-blind clinical study," *Archives of Otolaryngology*, vol. 106, pp. 772-776, 1980.

[23] Y.-H. Lai, Y. Tsao, and F. Chen, "Effects of adaptation rate and noise suppression on the intelligibility of compressed-envelope based speech," *PLOS ONE*, p. 10.1371/journal.pone. 0133519, 2015.

[24] A. ANSI, "S3. 5-1997, Methods for the calculation of the speech intelligibility index," *New York: American National Standards Institute*, 1997.

[25] R. V. Shannon, F. G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, pp. 303-304, 1995.

[26] J. Dien, "Issues in the application of the average reference: Review, critiques, and recommendations," *Behavior Research Methods*, vol. 30, pp. 34-43, 1998.