2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871531

Dysarthric Speech Enhancement Based on Convolution Neural Network

Syu-Siang Wang[1], Yu Tsao[2], Wei-Zhong Zheng[3], Hsiu-Wei Yeh[3], Pei-Chun Li[4], Shih-Hau Fang[1] and Ying-Hui Lai[3,5]

***Abstract*— Generally, those patients with dysarthria utter a distorted sound and the restrained intelligibility of a speech for both human and machine. To enhance the intelligibility of dysarthric speech, we applied a deep learning-based speech enhancement (SE) system in this task. Conventional SE approaches are used for shrinking noise components from the noise-corrupted input, and thus improve the sound quality and intelligibility simultaneously. In this study, we are focusing on reconstructing the severely distorted signal from the dysarthric speech for improving intelligibility. The proposed SE system prepares a convolutional neural network (CNN) model in the training phase, which is then used to process the dysarthric speech in the testing phase. During training, paired dysarthric-normal speech utterances are required. We adopt a dynamic time warping technique to align the dysarthric–normal utterances. The gained training data are used to train a CNN-based SE model. The proposed SE system is evaluated on the Google automatic speech recognition (ASR) system and a subjective listening test. The results showed that the proposed method could notably enhance the recognition performance for more than 10% in each of ASR and human recognitions from the unprocessed dysarthric speech.**

***Clinical relevance*— This study enhances the intelligibility and ASR accuracy from a dysarthria speech to more than 10%.**

I. INTRODUCTION

Due to the damaged neuro-muscular apparatus, dysarthria patients often utter distorted sound and exert more effort on improving the sound intelligibility in communicating with both human or machine. To regain communication efficiency, speech enhancement (SE) is one of the techniques that can be applied to this issue. One primary goal of SE is to improve sound intelligibility from noise-corrupted speech. It has been used as a preprocessor in various speech-related applications including hearing aids [1], [2] and automatic speech recognition (ASR) [3], [4]. Generally, conventional SE techniques used for minimizing the noise interference from the noisy input can be broadly classified into filtering-, spectral restoration-, and speech model-based approaches [5]. The underlying idea of these unsupervised approaches involves performing regression in terms of the statistical properties of speech and distortion sources to obtain enhanced speech. Some famous methods include the Wiener filter [6], minimum mean-square-error spectral estimator [7], generalized maximum a posteriori spectral amplitude [8], harmonic model [5], and the hidden Markov model [9].

Owing to the rapid development of supervised deep-learning (DL) techniques in recent years, various studies have focused on applying DL to speech-related signal processes, including the SE task [10], [11], [12], [13], [14], [15], [16], [17]. Among these approaches, convolutional neural networks (CNNs) have been popularly used and demonstrate successful performance [18], [19], [20], [21], [22], [16]. In these SE systems, the CNN model is used to form a non-linear transformation to convert the input speech (which is with lower intelligibility) to obtain enhanced speech at the output. Due to its network architecture, the CNN model can more accurately characterize the local information than fully-connected models.

From the viewpoint of intelligibility improvement, in this study, we are going to investigate the feasibility of the application of the DL-based SE technique to the dysarthria task. To improve the speech intelligibility of dysarthric speech, we proposed to use the CNN model to build a dysarthric SE system. Instead of the de-noising operation, the CNN-based SE system is performed in a clean environment for reconstructing the normal speech from a damaged input waveform. In addition, the paired dysarthric-normal utterances are normally unavailable in the testing condition. We evaluated the intelligibility of CNN-SE in terms of subjective tests and the on-line Google ASR system. Notably, the CNN-SE processed dysarthria speech can provide more than 10% intelligibility score improvements from the system input for each of the subjective listening tests and the recognized accuracy of ASR.

The remainder of this paper is organized as follows: Section II describes the overall dysarthric SE systems. Section III introduces the design of the experiment of this study and describes the comparative systems. Section IV presents the experimental results and discussion. Finally, Section V summarizes the findings.

II. METHODOLOGY

The block diagram of the proposed SE system is presented in Fig. 1. In the figure, a short-time Fourier transform (STFT) was performed on the dysarthric speech in the feature extraction function to obtain a series of complex-valued frames in the frequency domain. In addition, the 16 ms Hamming window and 1 ms hop time were used in STFT to

This work was supported by the Ministry of Science and Technology, Taiwan, under Grant MOST 110-2218-E-A49A-501

[1]Department of Electrical Engineering, Yuan Ze University, Taoyuan, Taiwan

[2]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan

[3]Department of Biomedical Engineering, National Yang Ming Chiao Tung university, Taipei, Taiwan

[4]Department of Audiology and speech language pathology, Macky Medical College New Taipei City, Taiwan.

[5]Medical Device Innovation & Translation Center, National Yang Ming Chiao Tung University, Taipei, Taiwan

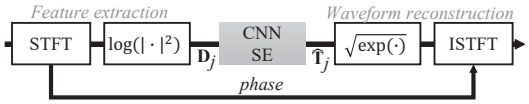

Fig. 1. Block diagram of the overall SE system.

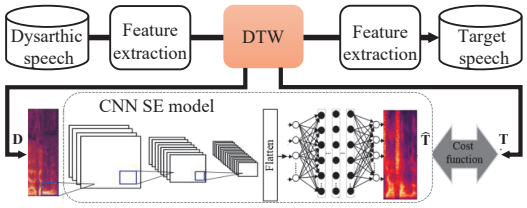

Fig. 2. Block diagram of the training procedure for the CNN-based SE approach.

provide a sequence of time frames. In the frequency domain, the logarithmic operation was then carried out to obtain the logarithmic power spectra (LPS), $\mathbf{D}_j$, after powering the magnitude components in the jth frame of the feature extraction block (Du and Huo, 2008). The SE system was applied to the input $\mathbf{D}_j$ to generate the enhanced LPS, $\mathbf{T}_j$, while keeping the phase component of the frame unaltered in the system output. In the waveform reconstruction function, this enhanced LPS was used in tandem with the preserved phase to finally reconstruct the enhanced utterances. Please note that the proposed dysarthric SE approach is prepared in a supervised training fashion.

A. CNN-based SE models

Fig. 2 shows the block diagram of the training stages of both the CNN-based SE models used in this study. As can be seen from the figure, we first prepared the dysarthric and normal speech datasets, wherein the same script was used for recording. The feature extraction operation was then performed for all utterances in both datasets to extract the LPS, $\mathbf{D}$, and $\mathbf{T}$ for the training process.

B. Dynamic time warping

To realize the proposed enhancement system, a supervised CNN model was first applied to the provided parallel data (described in the section of III-A), which can normally be easily made available for SE tasks by pairing the target speech with a noise-contaminated version of the same. However, there was no aligned target speech for the corresponding dysarthric speech of the SE task of this study. To address this issue, we performed dynamic time warping (DTW), which is a commonly used technique in voice conversion [23], [24], on the training speech corpus to align the LPS of normal speakers with those of dysarthria patients. The DTW algorithm was implemented in a frame-wise manner to measure the similarity between two temporal sequences that were varied with respect to time or speaking rates [25]. Using DTW, the parallel corpus was used to perform DL using the CNN by inputting the dysarthric speech into the model and locating the normal LPS in the output.

The DTW algorithm is used to determine the optimal alignment between $\mathbf{D}$ and $\mathbf{T}$ along the time axis caused by the explicitly mismatched frame numbers [26], [25]. There are N and M frame vectors contained in $\mathbf{D}$ and $\mathbf{T}$, respectively.

We then define the K-element effort path $\mathfrak{w}_1, \cdots, \mathfrak{w}_K$, where each element is $\mathfrak{w}_k \equiv (i, j)$ with respect to the paired frame indexes of ($\mathbf{D}$, $\mathbf{T}$).

The warping path satisfies the following conditions:

- $\mathfrak{w}_1 \equiv (1, 1)$ indicates that the warp path starts at the beginning of each time series.
- $\mathfrak{w}_K \equiv (N, M)$ indicates that the warp path ends at the end of each time series.
- If $\mathfrak{w}_k \equiv (i, j)$ and $\mathfrak{w}_{k+1} \equiv (i', j')$, then $i' \subseteq i, i+1$ and $j' \subseteq j, j+1$. This statement suggests that every index of both time series from the start to the end is monotonically increasing and is used in the warp path.

The dynamic programming principle is then employed on the accumulative Euclidean distance matrix, which is calculated from the cross-correlation between $\mathbf{D}$ and $\mathbf{T}$. The DTW is achieved by searching the minimizing-distance path on this matrix to determine the optimal effort path and the associated time alignment for both $\mathbf{D}$ and $\mathbf{T}$.

C. Training stage of the CNN

For the training phase of the CNN model, we placed $\mathbf{D}$ and $\mathbf{T}$ at the input and output of the CNN, respectively. The CNN model consisted of a series of paired convolutional layers subsequent to the flatten operation and a fully connected layer. A convolutional layer was composed of a set of filters followed by an activation function to extract several two-dimensional feature maps from the input tensor feature. However, the flatten operation and fully connected layers were used to enhance the tensor-speech representation to achieve an enhanced output LPS, $\hat{\mathbf{T}}$. The process of CNN can be briefly formulated in Eq. (1).

$$\hat{\mathbf{T}} = FCLs\{Flatten\{Convs\{\mathbf{D}\}\}\}, \tag{1}$$

where $FCL\{\cdot\}$, $Flatten\{\cdot\}$ and $Convs\{\cdot\}$ represent a fully connected layer, flatten operation and 2D-convolutional operations, respectively. For $Convs$ model, filter numbers of six hidden layers are 8, 16, 32, 64, 128, and 256 in order. In this study, for each layer, the kernel size and strides are 3×3 and 2×2, respectively. The sigmoid function is leveraged to normalize the output of each hidden layer. In addition, we apply global average pooling to perform the flatten function. The model parameter set θ was then optimized by minimizing the mean square error (MSE) between $\hat{\mathbf{T}}$ and $\mathbf{T}$. In contrast with the deep denoising autoencoder (DDAE) model [20], where every neuron in a fully connected layer was connected with all outputs of the previous layer, the applied CNN model not only reduced the number of parameters but also extracted the localized time-frequency characteristics by using several small-size filters in a convolution layer.

III. EXPERIMENTAL SETUP

A. Materials

Three hundred and twenty phrases selected from the Taiwan mandarin hearing in noise test (TMHINT) script [27]

were used to prepare the corpus for the proposed CNN-based SE model. A stroke patient and a normal speaker were asked to provide the dysarthric and normal datasets; all utterances were recorded at a 16 kHz sampling rate. Thus, there were 320 dysarthric–normal paired waveforms available for evaluation. Among them, 240 paired utterances were used as the training set, whereas the remaining 80 were used as the test set, which will henceforth be referred to as "open test." As mentioned earlier, the dysarthric–normal paired utterances in the training set were first aligned via the DTW approach [23].

In the following evaluations, it is worth noting that the closing test set (henceforth denoted as "closed test"), which comprised of 80 re-pronounced dysarthric utterances selected from the training script, was also included in the evaluation.

B. Procedure

We conducted our experiments using (1) the Google ASR system and (2) a subjective listening test to directly evaluate the intelligibility performance of the proposed method with regard to human-machine and human-human communications. For Google ASR, we used the Google-provided application programming interface for implementing ASR using a python program . In addition, character accuracy is used to evaluate the recognized results. A higher recognition score represents a better recognition performance for human-machine communication. For the subjective listening test, eight native Taiwanese Mandarin speakers aged 20 to 25 years were chosen as participants. The experiments were conducted in a soundproof booth, and the stimuli were played to the subjects through a set of M-audio BX5 D3 speakers at a comfortable listening level. The word correct rate (WCR) [28] was used to evaluate the intelligibility performance and denoted as "accuracy."

For comparison with the proposed CNN system, the DDAE and the joint dictionary learning-based non-negative matrix factorization (JD-NMF) algorithms were carried out on the TMHINT dataset for enhancing the input dysarthric speech. The DDAE model, which is composed of several fully-connected hidden layers, was applied to enhance the dysarthric LPS to normal output. The MSE cost function was employed to provide model parameters. On the other hand, the JD-NMF algorithm applied the NMF technique to jointly learn the source dysarthric and target normal dictionaries from straight vocoder features [29]. Thereafter, another dysarthric speech sample was converted to the normal speech of the target speaker using this joint dictionary. By specifying a small number of bases using the NMF technique, JD-NMF can learn a set of bases that are representative of the entire set of exemplars (estimated from the training data). In this study, we optimized JD-NMF and DDAE to achieve the most optimal performance.

IV. RESULTS AND DISCUSSION

We first qualitatively analyzed the processed utterances using amplitude envelopes in the following procedure. Two different dysarthric-normal-utterance pairs selected from the testing sets were aligned first by applying DTW. Each dysarthric speech was then processed by the JD-NMF-, DDAE-, and CNN-based SE systems and denoted as "dysarthric," "JD-NMF," "DDAE," and "CNN," respectively. All ten processed utterances were then placed individually to the input of the following vocoder system, which was consisted of the preemphasis, band-pass filter (BPF) and envelope detection. The pre-emphasis process preserved the signal components above 1200 Hz with a high-pass filter. Eight BPFs with cutoff frequencies at 80, 201, 384, 656, 1065, 1675, 2588, 3955, and 6000 Hz were applied to decompose the input speech into 8 different sub-band sequences. A low-pass filter with a 400-Hz cutoff frequency was then performed in the final step of a vocoder system on each sub-band signal to generate the frequency-band-related envelope. For each of ten vocoded speech, only the envelope at the latest frequency band (related to BPF between 3955 and 6000 Hz) was illustrated in Fig. 3 for analyzing. Notably, both normal speeches were denoted as "Target" in the figure. In addition, those envelopes depicted in the left column of the figure were the associated dysarthric-normal pair selected from the closed testing set, while those in the right-side figure were the dysarthric-normal utterances selected from open set.

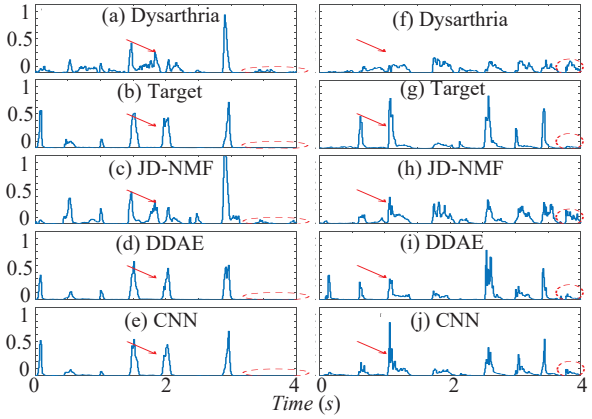

Fig. 3. The amplitude envelopes obtained from ten channels with respect to the center frequency at 5 kHz. Those envelopes were derived from (A) and (F) the dysarthric speech waveforms, (B) and (G) the target normal speech, and the enhanced (C) and (H) JD-NMF, (D) and (I) DDAE, and (E) and (J) CNN utterances. In addition, those enve-lopes in the left column were the associated dysarthric-normal speech pair selected to form the closed testing set, while those in the right-side figure were the associated dysarthric-normal pair selected form the open testing set.

In Fig. 3, the envelopes of DDAE and CNN exhibit amplitude trajectories that are more similar to those of the target when compared with the similarity between the amplitude trajectories of the envelopes of JD-NMF and the target. This observation indicates that the DDAE and CNN models are able to achieve more ac-curate transformations for a dysarthric speech on both testing sets. This comparison also suggests the good model capability of each of DDAE and CNN. However, the envelope of the CNN is more similar to that of the target in comparison with the similarity between the envelope of the DDAE and that of the target; this is especially obvious in Fig. 3 (J) in the open test set. This observation reveals that the proposed CNN-based SE system

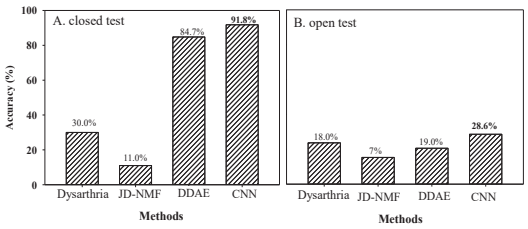

Fig. 4. The performance of speech enhancement evaluated based on the Google ASR system. The X-axis shows the approach employed, and the Y-axis shows the accuracy of each SE approach.

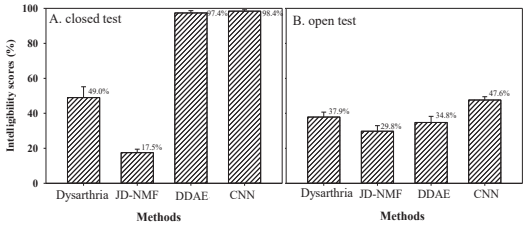

Fig. 5. The average score of intelligibility listening test by eight subjects. The X-axis is the processed method and Y-axis is the intelligibility score.

can achieve lesser sound distortion, detailed signal structures, and high-modulation depth signal infor-mation for listeners. Notably, similar observations can be made at other frequency bands. In addition, each envelope in the red dotted ellipses in Fig. 3 represents one consonant part. According to [30], for tonal Mandarin speech, the precise consonant structure of the CNN demonstrated in this figure suggests decent intelligibility for listeners.

Figure 4 shows the averaged accuracy scores obtained via the Google ASR system for the non-enhanced dysarthric samples and the samples obtained using the JD-NMF algorithm, DDAE, and CNN. From the figure, it can be observed that CNN achieved the highest accuracy when compared with those of the non-enhanced dysarthric samples, JD-NMF, and DDAE in both closed and open test conditions. For the closed test condition, the accuracy of the dysarthric samples was 30.0%, whereas those of the JD-NMF algorithm, DDAE, and CNN were 11.0%, 84.7%, and 91.8%, respectively. For the open test, the accuracy score of the dysarthric samples was 18.0%, whereas those of the JD-NMF algorithm, DDAE, and CNN were 7.0%, 19.0%, and 28.6%.

Apart from the evaluation conducted using Google ASR, we also conducted a listening test for the dysarthriac samples and for the results obtained using the JD-NMF algorithm, DDAE, and CNN to perform the intelligibility test. We list the evaluation results of the listening test in Fig. 5. In addition, the average intelligibility scores (%) and standard deviations are presented in the same figure. From Fig. 5, it can be observed that CNN achieved the highest intelligibility score when compared with the dysarthriac samples, JD-NMF algorithm, and DDAE in both closed and open tests. The intelligibility scores of the dysarthriac samples, JD-NMF algorithm, DDAE, and CNN in the close test condition are 49.0 ± 6.2, 17.5 ± 2.0, 97.4 ± 1.3, and 98.4 ± 0.8, respectively. However, the average scores achieved using the dysarthriac samples, JD-NMF algorithm, DDAE, and CNN in the open test condition are 37.9 ± 2.8, 29.8 ± 3.2, 34.8 ± 3.5, and 47.6 ± 1.9, respectively. These results indicate the effectiveness of the proposed CNN in providing good speech intelligibility for normal listeners.

From Figs. 4 and 5, it is seen that the DDAE and CNN systems, which employ a similar training procedure on the same corpus, achieve different intelligibility improvements with regard to dysarthric speech. The difference is caused by the differences in signal processing between the DDAE and CNN. In the DDAE, the input LPS was analyzed by the fully-connected model. The localized time-frequency signal structure cannot be effectively characterized by the DDAE to the extent to which CNN can. The analysis implies that the CNN model is more suitable for processing dysarthric utterances to promote effective communication.

In addition to the observations made regarding the model structure, we observe that the intelligibility scores of both the DDAE and CNN in the closed set are better than those in the open set. One possible explanation for this is that overfitting [29] may have occurred owing to the insufficiency of training data in this study. Many approaches have been proposed to address overfitting. One simple solution is to collect as many dysarthric–normal speech pairs as possible to provide a large training set to train DL-based models, thereby improving the effectiveness of a SE system with regard to processing dysarthric speech. In contrast, the application of data augmentation approaches, including [31], [32], to SE is also a feasible solution for alleviating this issue. Therefore, a relatively small training set is sufficient for training a SE model when combined with augmentation features. It should be noted that it is difficult for patients to record samples for a long time, leading to the small size of the training data set. To reduce the burden of the patients with regard to recording the samples, an assistant sound generation system [33], [34], [35], [36] could be used as an alternative means of generating various dysarthric utterances, thereby increasing the volume of the training corpus.

V. CONCLUSIONS

Herein, we proposed a CNN-based SE approach to enhance dysarthric speech and improve the speech intelligibility of dysarthric utterances. Owing to the effective signal process employed for localizing the time-frequency characteristics, the proposed method achieves a superior evaluation performance on both the closed and open testing sets. Specifically, the CNN-based model achieved a superior recognition performance in both the Google ASR system-based evaluation and the evaluation performed using the subjective listening test. In addition, analyses performed on envelopes in the frequency domain suggest that the proposed CNN-based SE system yields more detailed signal amplitude structures than those obtained via the conventional approaches. These results demonstrate that the proposed CNN-based SE system can potentially be used as an assistive system to overcome the degradation of speech intelligibility caused by dysarthria.

REFERENCES

[1] S. Doclo, M. Moonen, T. Van den Bogaert, and J. Wouters, "Reduced-bandwidth and distributed MWF-based noise reduction algorithms for binaural hearing aids," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 38–51, 2009.

[2] J. Thiemann, M. Müller, D. Marquardt, S. Doclo, and S. van de Par, "Speech enhancement for multimicrophone binaural hearing aids aiming to preserve the spatial auditory scene," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 1–11, 2016.

[3] W. Hartmann, A. Narayanan, E. Fosler-Lussier, and D. Wang, "A direct masking approach to robust asr," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 1993–2005, 2013.

[4] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, "Robust automatic speech recognition: a bridge to practical applications," 01 2015.

[5] J. Chen, J. Benesty, Y. Huang, and E. Diethorn, "Fundamentals of noise reduction in spring handbook of speech processing-chapter 43," 2008.

[6] P. Scalart *et al.*, "Speech enhancement based on a priori signal to noise estimation," in *Proc. ICASSP*, vol. 2, pp. 629–632, 1996.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

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

[9] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[10] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. AISTATS*, pp. 315–323, 2011.

[11] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Proc. APSIPA*, pp. 1–6, 2016.

[12] K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda, "Statistical voice conversion with wavenet-based waveform generation.," in *Proc. Interspeech*, pp. 1138–1142, 2017.

[13] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[14] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech Enhancement Generative Adversarial Network," in *Proc. Interspeech*, pp. 3642–3646, 2017.

[15] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Proc. ICASSP*, pp. 4869–4873, 2015.

[16] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[17] D. Yu and L. Deng, "Deep neural network-hidden markov model hybrid systems," Springer London, 2015.

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

[19] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

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

[21] K. Tan, J. Chen, and D. Wang, "Gated residual networks with dilated convolutions for supervised speech separation," in *Proc. ICASSP*, pp. 21–25, 2018.

[22] 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, no. 12, pp. 1849–1858, 2014.

[23] L. Muda, B. KM, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques," *Journal of Computing*, vol. 2, no. 3, pp. 138–143, 2010.

[24] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of straight spectrum," in *Proc. ICASSP*, pp. 841–844, 2001.

[25] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, 2007.

[26] L. Hong and J. S. Dhupia, "A time domain approach to diagnose gearbox fault based on measured vibration signals," *Journal of Sound and Vibration*, vol. 333, no. 7, pp. 2164–2180, 2014.

[27] M.-W. Huang, "Development of Taiwan Mandarin hearing in noise test," *Master thesis, Department of speech language pathology and audiology, National Taipei University of Nursing and Health science*, 2005.

[28] F. Chen and P. C. Loizou, "Predicting the intelligibility of vocoded and wideband mandarin chinese," *The Journal of the Acoustical Society of America*, vol. 129, no. 5, pp. 3281–3290, 2011.

[29] S.-W. Fu, P.-C. Li, Y.-H. Lai, C.-C. Yang, L.-C. Hsieh, and Y. Tsao, "Joint dictionary learning-based non-negative matrix factorization for voice conversion to improve speech intelligibility after oral surgery," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 11, pp. 2584–2594, 2016.

[30] F. Chen, L. L. Wong, and E. Y. Wong, "Assessing the perceptual contributions of vowels and consonants to mandarin sentence intelligibility," *The Journal of the Acoustical Society of America*, vol. 134, no. 2, pp. EL178–EL184, 2013.

[31] Y. Jiao, M. Tu, V. Berisha, and J. Liss, "Simulating dysarthric speech for training data augmentation in clinical speech applications," in *Proc. ICASSP*, pp. 6009–6013, 2018.

[32] Y. Takashima, T. Takiguchi, and Y. Ariki, "End-to-end dysarthric speech recognition using multiple databases," in *Proc. ICASSP*, pp. 6395–6399, 2019.

[33] P.-C. Li, Y.-Y. Chiang, K.-S. Tsai, and S.-T. Young, "Genetic algorithm for the efficient selection of disyllabic word lists used in mandarin speech discrimination tests," *Medical and Biological Engineering and Computing*, vol. 43, no. 5, pp. 648–657, 2005.

[34] K.-S. Tsai, L.-H. Tseng, C.-J. Wu, and S.-T. Young, "Development of a mandarin monosyllable recognition test," *Ear and hearing*, vol. 30, no. 1, pp. 90–99, 2009.

[35] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," *arXiv preprint arXiv:1711.11293*, 2017.

[36] F. Fang, J. Yamagishi, I. Echizen, and J. Lorenzo-Trueba, "High-quality nonparallel voice conversion based on cycle-consistent adversarial network," in *Proc. ICASSP*, pp. 5279–5283, 2018.