INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Enhancing Intelligibility of Dysarthric Speech Using Gated Convolutional-based Voice Conversion System

Chen-Yu Chen[1], *Wei-Zhong Zheng*[1], *Syu-Siang Wang*[2], *Yu Tsao*[2] *Pei-Chun Li*[3] *and Ying-Hui Lai* [1]

[1]Department of Biomedical Engineering, National Yang-Ming University
[2] Research Center for Information Technology Innovation, Academia Sinica
[3]Department of Audiology and Speech Language Pathology, Mackay Medical College

chenchenyu231@gmail.com, s1010654@gm.ym.edu.tw,
{sypdbhee,yu.tsao}@citi.sinica.edu.tw, ankh_li@mmc.edu.tw, yh.lai@gm.ym.edu.tw

Abstract

The voice conversion (VC) system is a well-known approach to improve the communication efficiency of patients with dysarthria. In this study, we used a gated convolutional neural network (Gated CNN) with the phonetic posteriorgrams (PPGs) features to perform VC for patients with dysarthria, with WaveRNN vocoder used to synthesis converted speech. In addition, two well-known deep learning-based models, convolution neural network (CNN) and bidirectional long short-term memory (BLSTM) were used to compare with the Gated CNN in the proposed VC system. The results from the evaluation of speech intelligibility metric of Google ASR and listening test showed that the proposed system performed better than the original dysarthric speech. Meanwhile, the Gated CNN model performs better than the other models and requires fewer parameters compared to BLSTM. The results suggested that Gated CNN can be used as a communication assistive system to overcome the degradation of speech intelligibility caused by dysarthria.

Index Terms: dysarthric speech, voice conversion, deep learning, speech intelligibility, patients with dysarthria

1. Introduction

Dysarthria [1] is a speech disorder that is often caused by neurological damage. For a dysarthric speaker, it may result in phoneme loss, unstable prosody, and imprecise articulation; hence, it causes a lack of speech intelligibility and difficulty in communication. There are different augmentative and alternative communication (AAC) devices to help patients communicate with people, such as talking keyboard, eye tracking device [2], and communication board [3]. Although those devices can help dysarthric speakers express their thoughts, they are often inefficient and inconvenient due to problems such as slower communication with people and can be further improved [4]. For this, the voice conversion (VC) system can be a suitable technique for patients with dysarthria.

VC, developed more than 20 years ago, is a technique to transfer one's speech to another's while preserving linguistic information. A continuous probabilistic VC method based on Gaussian mixture model (GMM) was proposed by Stylianou et al. [5] in 1998. Toda et al. proposed a GMM-based VC using dynamic features and global variance, which improves the conversion performance in both speech quality and accuracy [6]. Besides GMM-based VC, different methods were also proposed. Desai et al. [7] introduced VC based on an artificial neural network (ANN), and the results showed that it achieved better VC performance than the GMM model.

More recently, the deep learning-based VC model was proven to achieve a more markable performance than ANN in the VC task (e.g., [8]). Following the success of deep learning technology, several VC methods with deep learning technology were continuously proposed. Chen et al. [9] proposed a VC method using a deep neural network (DNN) trained by layer-wise generative training, which improves both similarity and naturalness better than conventional methods. Lai et al. [10] present an LSTM-RNN system that combines acoustic and linguistic information while doing VC, and the result shows that the proposed method has lower speech distortion and better intelligibility than conventional LSTM-RNN. Sun et al. [11] introduced a deep bidirectional long short-term memory based on recurrent neural networks (DBLSTM-RNNs) as the VC model. Sun et al. [12] used the DBLSTM with phonetic posterior grams (PPGs) features to perform the many-to-one VC task, and the results showed that it performed well for the users. More recently, Serrano et al.[13] used the four layers of BLSTM structure with PPGs features for Esophageal speech and the results shown the proposed system can reduce the word error rate of an ASR system. Although the DBLSTM model was proven to perform well in the VC application; however, a huge number of parameters are needed to support the structure [14]. Therefore, it could increase the cost of the system on the hardware circuit to increase the cost of a VC system for users.

On the other hand, the convolution neural network (CNN) is another well-known model for the speech signal processing task. First, the CNN model applies a set of filters to analyze the features of local time-frequency structures to extract robust feature representations. Then, a fully connected layer is subsequently used to ultimately achieve good quality output speech signals. The previous study [14] indicated that the CNN model can perform similar to the BLSTM in the speech enhancement task but only needed 7% parameters compared to the BLSTM model. Hence, the CNN model can be an efficient structure for hardware implementation in the speech signal processing task.

Although the CNN model can be more sensitive to the local time-frequency structure than the DNN model and requires fewer parameters compared with BLSTM, the sequential and hierarchical structures of speech (e.g., voiced/unvoiced segments and phonemes/morphemes) could not be effectively preserved, compared to the BLSTM architectures. More recently, the Gated CNN [15] has shown that it can capture the features of long-term speech signals well. The structure of Gated CNN is similar to that of the CNN model, but the activation function of each layer of the Gated CNN can be trained based on the data. In other words, the network of

http://dx.doi.org/10.21437/Interspeech.2020-1367

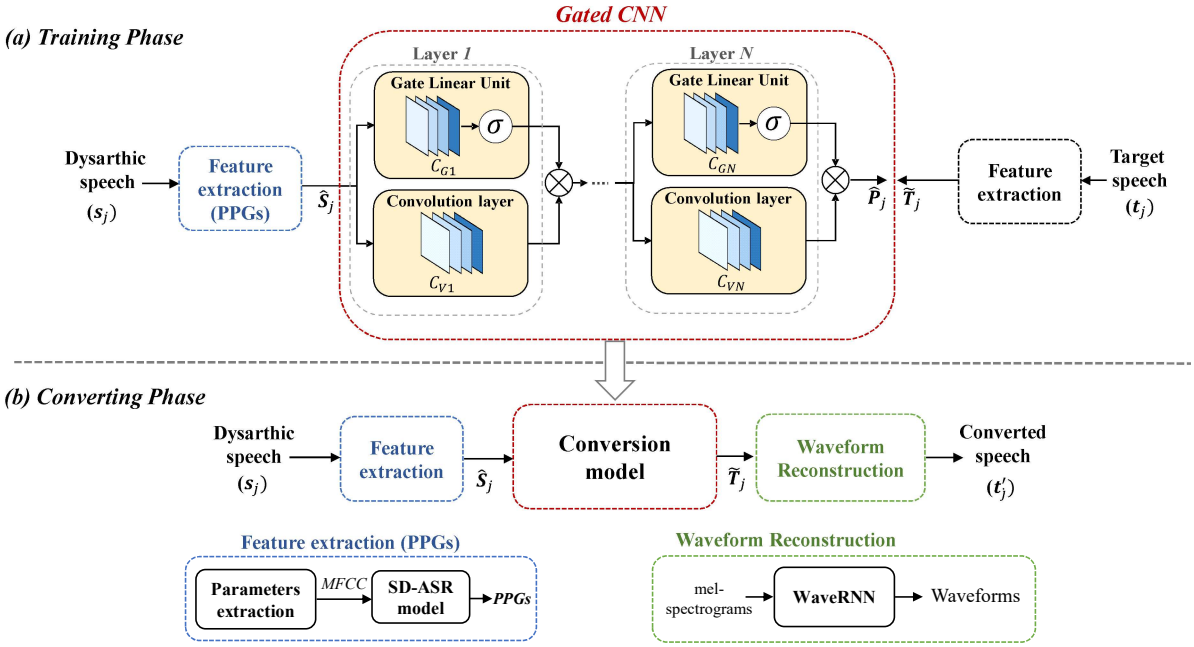

Figure 1. *Block diagram of (a) training and (b) conversion workflow of the proposed VC system*

Gated CNN can control which information should be propagated through the hierarchy of layers and achieve results similar to the RNN-based model. Hence, it could benefit to improve intelligibility performance of VC for dysarthric speakers. Moreover, previous studies indicated that the number of parameters in Gated CNN is less than in the BLSTM structure under similar circumstances[14]. Therefore, Gated CNN could be an efficient structure for the hardware implementation task, compared with the BLSTM structure. Following the success of the Gated CNN structure in the previous studies [15], this study aimed to propose a Gated CNN-based VC system with PPGs features for patients with dysarthria and to compare the speech intelligibility performance with CNN and BLSTM models in the dysarthric speech VC task.

The rest of the paper is organized as follows. A Gated CNN-based VC will be introduced in section 2. 3.Experiments and results are presented in sections 3. Finally, section 4 summarizes our findings.

2. Proposed System

The proposed system shown in Figure 1 included training and converting phases. In the training phase, parallel dysarthric speech ($\boldsymbol{s}_j$) and target speech ($\boldsymbol{t}_j$) were used to train the Gated CNN model. The $\boldsymbol{s}_j$ was processed by the unit of feature extraction (PPGs) to PPGs features. More specifically, the parameter extraction function was used to extract mel-cepstral coefficients (MCEPs) [16], which were the input of the ASR model to obtain the 74-dimension PPGs features ($\widehat{\boldsymbol{S}}_j$). The speaker dependent ASR (SD-ASR) system was used in this study. Our SD-ASR system were trained with ten patients with dysarthria and one normal speaker using Kaldi speech recognition toolkit [17], in which 3160 utterances could be used. Meanwhile, the target speech ($\boldsymbol{t}_j$) was processed by the feature extraction unit to obtain the 80-dimension mel-spectrograms ($\widetilde{\boldsymbol{T}}_j$). Finally, the $\widehat{\boldsymbol{S}}_j$ and $\widetilde{\boldsymbol{T}}_j$ were used as the input and output of the Gated CNN model to obtain the suitable parameters by training procedure.

The Gated CNN structure is illustrated in Figure 1. The input of our model is a sequence of PPGs that features $\widehat{\boldsymbol{S}}_0, \ldots, \widehat{\boldsymbol{S}}_j$. Compared with the convolution neural network (CNN) structure, Gated CNN contains a gated linear unit (GLU), which is presented as $\sigma(C_G(h_{N-1}) + b)$, where the C_G represents the convolution layer; h_{N-1} represents the output of previous layer and b is the learned bias parameter, respectively. The GLU replaces the activation function (e.g., relu, sigmoid); hence, the hidden layers $h_0, \ldots, h_N$ are computed as:

$$h_N(\widehat{\boldsymbol{S}}_j) = (C_{VN}(\widehat{\boldsymbol{S}}_j) + b_{VN}) \otimes \sigma(C_{GN}(\widehat{\boldsymbol{S}}_j) + b_{GN}) \quad (1)$$

where C_{VN}、b_{VN} are the covolution layer and bias of N layer; C_{GN}、b_{GN} are the convolution layer and bias of GLU unit in N^{th} layer. The σ is the sigmoid function and ⊗ is the element-wise product of the two matrixes. Moreover, the loss function of our training stage was defined as:

$$L = \frac{1}{N}\sum_{j=1}^{N} \left\| \widetilde{\boldsymbol{T}}_j - \hat{P}_j \right\| \quad (2)$$

wherein $\hat{P}_j$ is the predicted MCEPs of out model.

In the converting phase, the input PPGs feature $(\widehat{\boldsymbol{S}}_j)$ were extracted from dysarthric speech through the unit of feature extraction (PPGs), and was used as input for the conversion model to convert to $\widetilde{\boldsymbol{T}}_j$, the obtained Gated CNN model in the training phase was used. Next, the waveform reconstruction

based on WaveRNN vocoder [18] was used to convert waveform speech ($\hat{\mathbf{t}}_j'$) from $\widetilde{\boldsymbol{T}}_j$ to obtain the converted speech.

3. Experiments and results

3.1. Materials

The 576 (=288 utterances × 2 times) dysarthria-normal-paired utterances were used as the training set in this study, with the corpus list were adopted from Taiwan Mandarin hearing in noise test [19]. The dysarthric and target utterances were spoken by one male stroke patient and one normal male speaker, respectively. In the testing phase, we recorded 32 duplicate utterances and 32 outside utterances from the same stroke patient during the test the performance of each model. The duplicate utterances were the repetition of the sentence used in the training set, and the outside utterances were unseen content in the training set.

3.2. Procedure

This study aimed to evaluate the intelligibility benefits of the proposed VC system with the Gated CNN model (Figure 1) for patients with dysarthria. Meanwhile, two well-known models (i.e., CNN and BLSTM) were used as a comparison. More specifically, these two models were used to replace the Gated CNN model in Figure 1 to compare the intelligibility performance. In this experiment, we assessed the suitable complexity of each model; hence, we added the layer number to test the recognition rate by Google ASR system [20] and to figure out the suitable setting of each model. The detailed setting of each model in a layer is shown in Table 1, and the optimizer of Nadam is used in this study. Next, the best setting of each model (i.e., the converted speech achieved the highest accuracy of Google ASR metric) was used to represent this model for the listening test. Finally, we conducted the listening test of word correct rate (WCR) [21] and used a mean opinion score (MOS) method to test the speech naturalness. The listening tests include five listeners. The WCR evaluation is calculated by dividing the number of correctly identified words by the total number of words in each test. Each listener listened to 32 utterances (= 8 converted utterances ×3 models + 8 source dysarthric utterances) in outside test conditions. Meanwhile, the subjective listening test in the five-scale MOS test (1-bad, 2-poor, 3-fair, 4-good, 5-excellent) was used to test the speech naturalness between dysarthric and converted speeches.

3.3. Objective Evaluations of model performance

Tables 2 and 3 show the recognition rate of different layer numbers of each model by the Google ASR system in the duplicate test and the outside test conditions. From the results of duplicate test condition in Table 2, we can see that the best performance of CNN model is 82.5% accuracy rate at three layers structure; BLSTM model's best performance occurs at three layers of 75.5% accuracy rate; Gated CNN has its best performance at five and six layers reaches 85.9% accuracy rate. From the results of the outside test in Table 3, the CNN model has its highest accuracy rate of 68% at three layers; the BLSTM model's best performance is 65.7% at six layers, and Gated CNN has its best performance of 76.1% at five layers. Figure 2 shows the results of the average Google ASR recognition rate(i.e., duplicate and outside test conditions) on different layers of each model. We can find out that Gated CNN model's best performance is 81% and it has 30×10^6 parameters. BLSTM's best performance is 67%, and it has 85×10^6 parameters. CNN's best performance is 75%, and it has 12×10^6 parameters. The results lead to several conclusions: First, Gated CNN model performed the best among all of the three models, which could improve the Google ASR recognition rate from 17.1% to 81.0% in average results, it also demonstrates that Gated CNN can process these PPGs features better than the other two models. Second, the Gated CNN structure is a more efficient structure than BLSTM. For instance, consider the best performance models in average results, the best performance Gated CNN has only about 35% numbers of parameters compared to the best performance BLSTM model. From the conclusions above, we could see Gated CNN as a potential model for VC tasks to improve intelligibility performance for patients with dysarthria.

Table 1: *Details of network architectures of Gated CNN, CNN, and BLSTM models in the proposed system.*

Model	Structure description
Gated CNN	n^{th} layer = $(3 \times 3, 1024\ conv) \times n\ stacks,$ $Strides = 1$
CNN	n^{th} layer = $(3 \times 3, 1024\ conv,\ Relu) \times n\ stacks,$ $Strides = 1$
BLSTM	n^{th} layer = $74\ Dense,$ $(1024\ BLSTM \times (n-2)\ stacks),$ $80\ Dense$

Table 2: *Duplicate test of Google ASR accuracy in different layers of each model, where lr means layers.*

	Original	CNN	BLSTM	Gated CNN
3 lr	23.6%	82.5%	75.5%	85.6%
4 lr		76.2%	70.7%	83.1%
5 lr		77.8%	68.0%	85.9%
6 lr		74.5%	68.0%	85.9%

Table 3: *The outside test of Google ASR accuracy in different layers of each model, where lr means layers.*

	Original	CNN	BLSTM	Gated CNN
3 lr	10.6%	68.0%	48.8%	69.6%
4 lr		67.3%	62.0%	72.2%
5 lr		67.1%	55.2%	76.1%
6 lr		64.5%	65.7%	72.7%

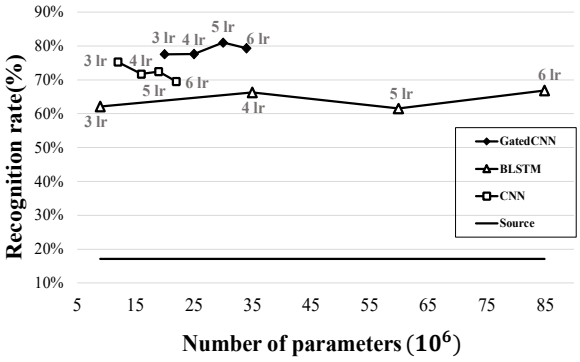

Figure 2. *The average recognition rate by Google ASR in different layers of each model*

3.4. Spectrograms analysis

The spectrogram is a well-known approach to analyze the spectral, temporal representations of a time-varying signal [22]. The same context utterances from different models are shown in Figure. 3. Noted, since the effect of the high-frequency part is similar, we only present the spectrum chart of the low-frequency area (i.e., 0 to 2k Hz) to compare the differences in Figure 3. We could see that the three models have a similar spectrogram as the target speech, which could imply that the three models had converted source to target utterances successfully. However, the Gated CNN spectrogram presents more detailed and clear formants (the red circles) than the other models. Hence, it implies that Gated CNN could provide better speech quality for listeners.

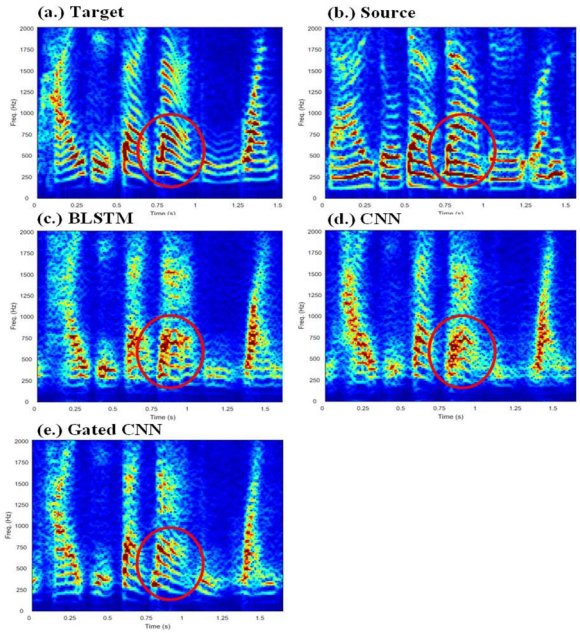

Figure 3. *The 80 dimensions spectrograms of the same context utterance from different models.*

Figures 4 and 5 are the average word correct rate (WCR) and the five-scale average MOS opinion test. The Gated CNN converted speech with an 87.8% accuracy rate, better than BLSTM (58.8%) and CNN (77.5%). The five-scale MOS comparison test shows that Gated CNN also has the best score of 3.92, better than BLSTM (2.6) and CNN (3.36). The listening tests show consistent results as the Google ASR test, and the Gated CNN still has the highest accuracy rate, and moreover, it converted speech with better naturalness than the other two models.

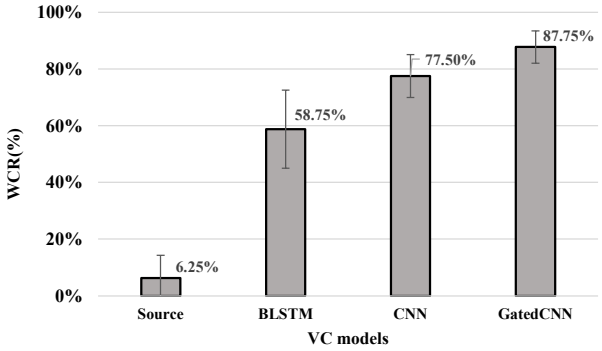

Figure 4. *Word correct rate (WCR) of average results from five listeners*

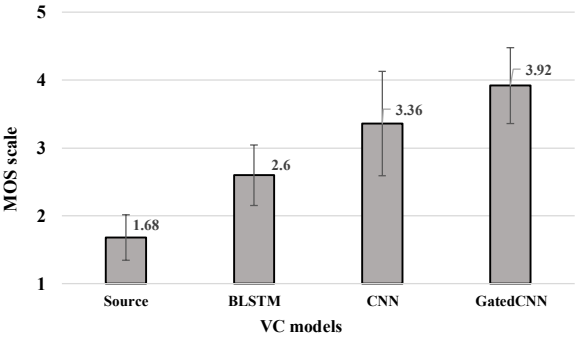

Figure 5. *MOS comparison based on naturalness of different models from five listeners.*

.

4. Conclusions

This paper proposes a Gated CNN based VC system with PPGs features to improve the speech intelligibility for patients with dysarthria, with the WaveRNN vocoder used to synthesis converted speech. The Google ASR evaluation metric showed that, on average, the Gated CNN model in the proposed VC system could improve the recognition rate from 17.1% to 81.0% (average performance in duplicate and outside test conditions in Figure 2). Meanwhile, the gated function of the Gated CNN model can further improve the performance of the CNN model in this study. Moreover, the number of parameters in the Gated CNN (i.e., 5lr, the best recognition rate of Gated CNN in this study) was roughly only 35% to that in BLSTM (i.e., 6lr, the best recognition rate of BLSTM in this study). Therefore, the cost of hardware implementation would be less than the BLSTM model for users. These findings indicate that the Gated CNN model can be a potential model for VC tasks to improve intelligibility performance for patients with dysarthria.

5. Acknowledgements

This study was supported by the Ministry of Science of Technology of Taiwan under the 108-2218-E-010 -004 project. The authors would like to thank APrevent Medical Inc. for providing the training and testing data.

6. References

[1] J. R. Duffy, *Motor speech disorders : substrates, differential diagnosis and management*. St. Louis, Missouri: Elsevier, 2013.

[2] C.-S. Lin, C.-W. Ho, W.-C. Chen, C.-C. Chiu, and M.-S. J. O. A. Yeh, "Powered wheelchair controlled by eye-tracking system," *Optica Applicata,* vol. 36, 2006.

[3] S. Calculator, C. D. A. J. J. o. s. Luchko, and H. Disorders, "Evaluating the effectiveness of a communication board training program," *Journal of speech and Hearing Disorders*, vol. 48, no. 2, pp. 185-191, 1983.

[4] U. o. Washington, "Augmentative and Alternative Communication Rate Enhancement," [Online]. Available http://depts.washington.edu/augcomm/02_features/04d_rateenhance.htm, [Accessed: 3-Feb-2020].

[5] Y. Stylianou, O. Cappé, E. J. I. T. o. s. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on speech and audio processing,* vol. 6, no. 2, pp. 131-142, 1998.

[6] T. Toda, A. W. Black, K. J. I. T. o. A. Tokuda, Speech,, and L. Processing, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222-2235, 2007.

[7] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3893-3896, 2009.

[8] T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, "Voice conversion in high-order eigen space using deep belief nets," in *Interspeech 2013*, pp. 369-372, 2013.

[9] L.-H. Chen, Z.-H. Ling, L.-J. Liu, L.-R. J. I. A. T. o. A. Dai, Speech,, and L. Processing, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1859-1872, 2014.

[10] J. Lai, B. Chen, T. Tan, S. Tong, and K. Yu, "Phone-aware LSTM-RNN for voice conversion," in *2016 IEEE 13th International Conference on Signal Processing (ICSP)*, pp. 177-182, 2016.

[11] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4869-4873, 2015.

[12] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *2016 IEEE International Conference on Multimedia and Expo (ICME*, pp. 1-6 *)*, 2016.

[13] L. Serrano, S. Raman, D. Tavarez, E. Navas, and I. J. P. I. Hernaez, "Parallel vs. Non-parallel Voice Conversion for Esophageal Speech,", in *Interspeech 2019*, pp. 4549-4553, 2019.

[14] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, H. J. I. A. T. o. A. Kawai, Speech,, and L. Processing, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1570-1584, 2018.

[15] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 933-941, 2017.

[16] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *ICASSP'83. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 8, pp. 93-96, 1983.

[17] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011.

[18] N. Kalchbrenner *et al.*, "Efficient neural audio synthesis," *arXiv,* preprint arXiv:1802.08435, 2018.

[19] L. L. Wong, S. D. Soli, S. Liu, N. Han, M.-W. J. E. Huang, and hearing, "Development of the Mandarin hearing in noise test (MHINT)," *Ear and hearing*, vol. 28, no. 2, pp. 70S-74S, 2007.

[20] *Google ASR api*. [Online]. Available: https://cloud.google.com/speech-to-text, [Accessed: 31-Jan-2020].

[21] F. Chen and P. C. J. T. J. o. t. A. S. o. A. 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.

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