Implementation of low-latency electrolaryngeal speech enhancement based on multi-task CLDNN

Kazuhiro Kobayashi
Information Technology Center
Nagoya University
Nagoya, Japan
kobayashi.kazuhiro@g.sp.m.is.nagoya-u.ac.jp

Tomoki Toda
Information Technology Center
Nagoya University
Nagoya, Japan
tomoki@icts.nagoya-u.ac.jp

***Abstract*—In this paper, we propose a low-latency speech enhancement technique for electrolaryngeal (EL) speech based on multi-task CLDNN. Although the EL speech can generate relatively intelligible speech, laryngectomees always suffer quality degradation of speech naturalness due to the mechanical excitation signals. To solve this problem, an EL speech enhancement technique based on CLDNN consisting of convolution, recurrent, and fully connected layers has been proposed. In this technique, an input feature vector of the EL speech is converted into several vocoder parameters such as excitation parameters and spectral parameters based on expert CLDNNs optimized for each feature. However, it is difficult to utilize speech communication because its bi-directional recurrent layers cause a large delay to wait for the end of the utterance. To address this issue, in this paper, we propose multi-task CLDNN with uni-directional recurrent layers for the low-latency EL speech enhancement. Moreover, to achieve comparable performance to the bi-directional CLDNN, we also propose the following techniques: 1) knowledge distillation, 2) data augmentation, and 3) phonetic regularization. The experimental results demonstrate that the proposed method makes it possible to achieve comparable objective results to the bi-directional CLDNN and outperform naturalness and speech intelligibility in the noisy condition.**

***Index Terms*—electrolaryngeal speech, low-latency speech enhancement, voice conversion, deep neural network**

I. Introduction

A laryngectomy is a surgery to remove the larynx including the vocal folds to treat laryngeal cancer, making a person lose the ability to produce source excitation sounds. To produce speech signals without using vocal fold vibrations, an electrolarynx (EL) is widely used by laryngectomees. Although the produced speech called electrolaryngeal speech (EL speech) is relatively intelligible [1], there are the following problems: 1) radiation of the intensive noise by the EL and 2) unnatural acoustic characteristic of the source excitation. Consequently, EL speech sounds mechanical and artificial compared with natural speech.

To address these issues, two approaches have been proposed. One approach is based on noise suppression [2] and the other is based on statistical voice conversion (VC) [3], [4]. The noise suppression approach [5]–[8] focuses on reducing the noise components leaked from the excitation signals. Although these techniques are effective for reducing the noise components, the enhanced EL speech suffers from musical noise caused by the processing of noise suppression. Moreover, the improvements of EL speech yielded by this approach are limited because most of the acoustic characteristics are not changed. On the other hand, the VC-based approach directly modifies these acoustic characteristics of EL speech [9], [10]. In this technique, acoustic features extracted from EL speech are converted into those of target natural speech based on the Gaussian mixture model (GMM). As a result, the voice generated by the converted acoustic features has relatively higher naturalness compared to EL speech. Moreover, by incorporating a low-latency conversion algorithm [11] for maximum likelihood parameter generation based on the GMM, it achieves not only utterance-by-utterance-based conversion but also real-time conversion.

Recently, statistical VC techniques have been significantly improved [12]–[19]. Several techniques incorporating these VC methods have been applied to the speech enhancement for the laryngectomees [20], [21]. In these techniques, it consists of bi-directional recurrent layers for estimating vocoder parameters such as F_0, unvoiced/voiced decision symbol, aperiodicity, and spectral feature. However, it is difficult to directly utilize these techniques for speech communication as long as using the bi-directional recurrent layers, because they require a whole utterance to consider not only forward state sequence but also backward state sequence, causing a long delay after starting to speak. Moreover, these techniques require large computational costs because they separately trained single-task conversion models for each acoustic feature.

In this paper, in order to implement a low-latency speech enhancement system for EL speech, we propose multi-task CLDNN [22] consisting of convolutional, uni-directional recurrent, and fully connected layers. By using uni-directional recurrent layers, it is not necessary to wait for the end of the utterance. And, the multi-task model makes it possible to reduce the computational costs drastically because it generates several vocoder parameters at once. As a result, the proposed method is capable of converting the input feature vectors frame-by-frame, making it possible to implement low-latency speech enhancement. Furthermore, in speech communication, the low-latency speech enhancement system needs to be used in clean conditions as well as noisy conditions which usually makes significant quality degradation of the converted voice. In order to make the proposed method more robust in

any sound environment conditions, we also propose several techniques such as knowledge distillation [23] based on born-again network [24], data augmentation by noise injection and SpecAugment [25], and phonetic regularization based on phonetic posteriorgrams.

II. Conventional EL speech enhancement based on CLDNN

CLDNN consists of convolutional layers, recurrent layers, and fully connected (FC) layers with two skipped connections. For the inputs of the first convolutional layer, a one-dimensional feature vector at frame t is transformed into a two-dimensional feature matrix by concatenating several preceding and succeeding frames to capture contextual information. To add the original input feature vector at frame t through a skipped connection, dimension reduction is performed using a linear layer with outputs from the convolutional layers. Then, the resulting outputs are fed into the recurrent layers. For the recurrent layers, bi-directional gated recurrent units (Bi-GRU) are used to reduce the number of parameters from that in the original implementation of the long-short time memory. The outputs of the Bi-GRU layers are concatenated into those of the convolutional layers. Finally, the resulting outputs are fed into the FC layers to be transformed into the output feature vector.

In the training process, three single-task CLDNNs are trained separately. The segmental features such as mel-cepstrum and aperiodicities, are modeled by an expert CLDNN by concatenating these acoustic features. For the prosodic features, continuous F_0 and unvoiced/voiced (U/V) symbols are modeled separately. In the conversion process, mel-cepstrum extracted from EL speech is converted into U/V symbols, a continuous F_0, the mel-cepstrum, and aperiodicities based on these CLDNNs. For F_0, the estimated continuous F_0 sequence is masked using the estimated U/V symbols. Finally, the enhanced speech is generated by source-filter vocoder using these acoustic features.

In the conventional speech enhancement based on the CLDNN, there are two problems to implement a low-latency speech enhancement system. First, it requires to perform inference three times to generate vocoder parameters because those features, such as F_0, U/V, and segmental feature, are modeled by three single-task CLDNNs. Second, the Bi-GRU layers require to wait till the end of an utterance in order to utilize the backward state sequence. Therefore, the conventional EL speech enhancement based on the CLDNNs only accepts utterance-by-utterance conversion.

III. EL speech enhancement based on multi-task CLDNN

Figure 1 indicates a training overview of the proposed EL speech enhancement. In the proposed method, in order to solve the problems of the conventional single-task CLDNNs (ST-CLDNN) with Bi-GRU, we modify it into multi-task

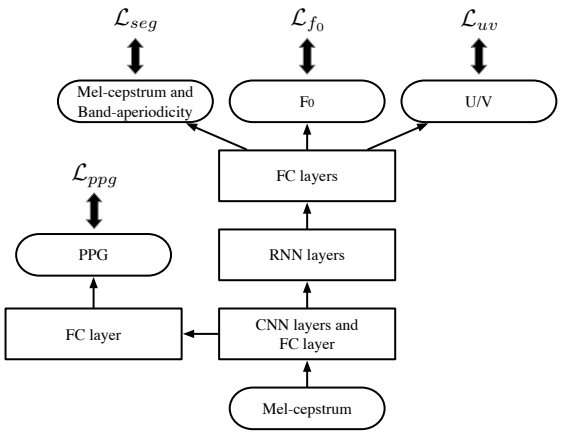

Fig. 1. Training overview of the multi-task CLDNN.

CLDNN (MT-CLDNN) with uni-directional GRU (Uni-GRU). The objective function of the MT-CLDNN follows:

$$\mathcal{L}_{obj} = \mathcal{L}_{seg} + \alpha_{pro}\left(\mathcal{L}_{f_0} + \mathcal{L}_{uv}\right), \tag{1}$$

where $\mathcal{L}_{seg}$ and $\mathcal{L}_{f_0}$ are loss functions of the mean squared error for the segmental features and continuous F_0, respectively. $\mathcal{L}_{uv}$ is loss function of binary cross-entropy for the U/V decision symbol. α_{pro} indicates a hyper-parameter to balance the optimization between the segmental features and the other prosodic features. In the conversion process, all vocoder parameters are simultaneously estimated based on the MT-CLDNN using the spectral feature of the EL speech.

A. Born-again network

Born-again network (BAN) [24] is a knowledge distillation technique to train a multi-task model by using several pre-trained single-task models based on teacher-student learning. In the end-to-end automatic speech recognition (ASR), the knowledge distillation technique from bi-directional recurrent layer into uni-directional recurrent layer to achieve online ASR [26]. In this paper, inspired by these researches, we propose a knowledge distillation technique of MT-CLDNN with Uni-GRU using pre-trained ST-CLDNNs with Bi-GRU for the low-latency speech enhancement. The loss function follows:

$$\mathcal{L}_{objban} = \alpha_{ban}\mathcal{L}_{ban} + (1 - \alpha_{ban})\mathcal{L}_{obj}, \tag{2}$$

where $\mathcal{L}_{ban}$ indicates the loss function between outputs of the ST-CLDNN and outputs of the MT-CLDNN. The hyper-parameter α_{ban} decays to control the balance between teacher-student learning and training using ground truth.

B. Data augmentation

For the low-latency EL speech enhancement, it is important that the system works in not only clean conditions but also noisy conditions. In this paper, in order to make the proposed method more robust in any sound environment, we apply two kinds of data augmentation techniques. The one

is data augmentation based on noise injection and the other is data augmentation based on SpecAugment [25]. For the noise injection, we simply impose several kinds of noises into the input EL speech. For the SpecAugment, the input feature vector is masked based on randomly selected frame and dimension lengths in each time and dimension axis.

C. Regularization by phonetic posteriorgrams

It is reported that phonetic features such as phonetic posteriorgrams (PPG) and bottleneck features of the ASR system are effective as the source feature vector for the statistical VC [15], [27]. However, it is difficult to directly apply the phonetic feature vectors to the low-latency speech enhancement system because the extraction of those features usually requires the ASR system, increasing computational cost, and delay. To address this issue, we propose a regularization technique of the network parameters by the PPG vector. It is considered that the convolutional layers mainly undertake contextual feature extraction from input feature vector sequences in the CLDNN-based speech enhancement. By regularizing these convolutional layers, it is expected that the outputs of the convolutional layers become similar to the phonetic features.

The hidden outputs of the convolutional layers are transformed into the PPG vector based on a single fully-connected layer. The objective function of the PPG regularization is calculated as follows:

$$\mathcal{L}_{objppg} = \mathcal{L}_{obj} + \alpha_{ppg}\mathcal{L}_{ppg}, \tag{3}$$

where $\mathcal{L}_{ppg}$ indicates the loss function of Kullback-Leibler divergence calculated using outputs of the FC layer and the PPG vectors extracted from target natural speech by the ASR system.

IV. Experimental evaluation

A. Experimental conditions

We used 120 Japanese sentences. One electrolarygectmee and one healthy male Japanese speaker uttered EL and normal speech, respectively. Because it is important to implement EL speech enhancement with small amount of training data for reducing the burden of EL speech recordings, we set the number of training and development utterances to 54 and 6, respectively. The frame and frameshift sizes were set to 25 ms and 5 ms, respectively. The other 60 utterances were used for the evaluation. For the inputs of the first convolutional layer, one-dimensional feature vectors were extended by concatenating 7 preceding and 3 succeeding feature vectors to obtain 11×25 two-dimensional feature matrices. Two-dimensional convolutions of 3×3 kernels were performed. Then, average pooling was applied after passing through batch normalization and activation function based on the rectified linear unit. We stacked two convolutional layers with 1 and 3 dilations for the time axis, respectively. The number of output channels for the first layer was 32 and that for the second layer was 64. In the RNN layers, the number of hidden layers was set to 2 for the Uni-GRU and 1 for the Bi-GRU to be the same parameter size, respectively. The number of hidden units was set to 256. The hyper-parameters α_{pro} and α_{ppg} were set to 0.1 and 10, respectively. α_{ban} was gradually varied from 1 to 0 by linear decay over epochs. We used stochastic gradient descent as the optimizer. The learning rate was set to 0.2. The number of epochs was set to 100. The other settings followed in [20].

For SpecAugment, we set sizes of the time and dimensional masks to 1 through 100 and 1 through 5, respectively. The mask size was randomly selected from these values based on the uniform distribution in each minibatch. For the noise injection (Noise), we prepared three kinds of environmental noises recorded in a dining room, a laboratory, and a meeting room. These noises were imposed on the training EL speech with 15, 20, and 25 dB signal-to-noise ratios. The clean EL and noise augmented EL speeches were randomly selected with an equivalent probability in each minibatch. For the PPG regularization, we used 166 dimensional PPG vector extracted by our internal implementation of the CLDNN [22] using "csj" recipe in Kaldi toolkit [28]. We did not use any context label for the PPG vector extraction.

For the evaluation, we imposed a crowd noise "N1" in [29] to the evaluation utterances with 12 dB signal-to-noise ratio. We denoted MT-CLDNN with Uni-GRU as "Uni" and MT-CLDNN with Bi-GRU as "Bi" and MT-CLDNN with Uni-GRU, SpecAugment, Noise, and PPG as "Uni + Mix".

B. Conversion latency of the proposed MT-CLDNN

Conversion latency indicates a time gap between input EL speech and output enhanced speech. In the proposed MT-CLDNN, several modules such as feature extraction, convolutional layer, waveform generation by vocoder cause algorithmic delays. For the feature extraction, to estimate the frequency spectrum of the input EL speech before parameterizing to mel-cepstrum, it requires to wait half of the frame size to store waveform samples for windowing. For the convolutional layer, we confirmed that a larger number of the succeeding frames contributes to conversion accuracy improvements in our internal evaluations. To balance the conversion accuracy and delay, we set the number of succeeding frames to 3 and it causes delays for three frames. For the vocoding process, in order to interpolate mel-cepstrums between current and next frames to perform MLSA filter [30], it causes a delay for one frame. Based on these algorithmic delays, the resulting delay of the proposed low-latency MT-CLDNN becomes 32.5 ms (12.5 ms for feature extraction, 15 ms for convolutional layer, and 5 ms for vocoder). Note that all processes from the feature extraction through waveform generation in each frame must be finished within 5 ms because the frameshift size was set to 5 ms. We ignored the latency for audio input/output in this calculation.

C. Objective evaluations

As objective evaluations, we compared objective measures of the converted acoustic features based on the root mean square error (RMSE), correlation coefficients, and mel-cepstrum distortion (Mel-CD).

TABLE I
OBJECTIVE RESULTS OF CLEAN INPUT.

Method	Mel-CD [dB]	F_0 correlation	Log F_0 RMSE	Aperiodicity RMSE
Uni	6.06	0.72	0.16	2.85
Uni + BAN	6.39	0.69	0.16	2.99
Uni + Noise	6.04	0.71	0.16	2.83
Uni + SpecAugment	6.05	0.70	0.16	2.86
Uni + PPG	5.96	**0.77**	**0.14**	**2.76**
Uni + Mix	**5.92**	0.76	0.15	**2.76**
Bi	5.91	0.78	0.14	2.76
Bi + Noise	6.00	0.79	0.14	2.87
Bi + SpecAugment	6.00	0.78	0.14	2.82
Bi + PPG	5.92	**0.81**	**0.13**	2.77
Bi + Mix	**5.83**	**0.81**	**0.13**	**2.72**

TABLE II
OBJECTIVE RESULTS OF NOISY INPUT.

Method	Mel-CD [dB]	F_0 correlation	Log F_0 RMSE	Aperiodicity RMSE
Uni	13.66	0.68	0.18	3.60
Uni + Mix	**7.47**	**0.71**	**0.16**	**3.00**
Bi	13.74	0.66	0.19	3.97
Bi + Mix	**7.67**	**0.79**	**0.14**	**3.01**

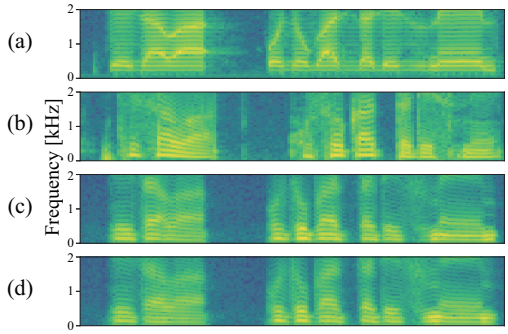

Fig. 2. Spectrograms of several converted voices. (a) EL speech, (b) normal speech, (c) converted voice by "Uni + Mix", and (d) converted voice by "Bi + Mix".

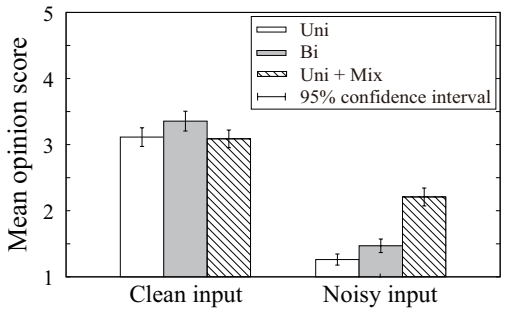

Fig. 3. Results for naturalness.

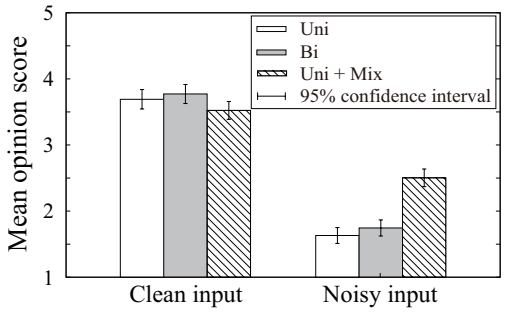

Fig. 4. Results for perceptual speech intelligibility.

Table I indicates the results of several objective measures in clean condition. We can see that there are little improvements in BAN, Noise, and SpecAugment. On the other hand, by using PPG regularization, we can see that there are large improvements in all objective measures in both uni-directional and bi-directional models. Moreover, by combining the PPG regularization with Noise and SpecAugment, it achieves further improvements in terms of Mel-CD in the uni-directional model and the performances of the "Uni + PPG" and "Uni + Mix" methods are comparable to that of "Bi". Table II indicates the results of several objective measures in noisy conditions. We can see that the "Uni + Mix" method outperforms "Uni" and "Bi".

Figure 2 indicates spectrograms of the several converted voices. You can see that the harmonics components of EL are fixed over utterances. On the other hand, the harmonics components of the proposed EL speech enhancement techniques vary gently similar to those of the normal speech.

D. Subjective evaluations

For the subjective evaluations, two preference tests were conducted. The number of subjects was 9. In the first test, the naturalness of the enhanced EL speech was evaluated using a mean opinion score (MOS). The enhanced speech samples were presented to subjects in random order. The subjects rated the naturalness of the presented speech using a five-point scale with "5" for excellent, "4" for good, "3" for fair, "2" for poor, and "1" for very poor. The number of sentences used in the evaluation for each subject was 114. In the second test, the perceptual speech intelligibility was evaluated in the same manner as the naturalness to measure easiness for speech content recognition. In this evaluation, we only evaluated "Uni", "Bi", and "Uni + Mix" in order to reduce the burden of the subjects.

Figure 3 shows the experimental results for the naturalness

of the enhanced speech. In the clean condition, we can see that there is no significant difference between all methods. On the other hand, in the noisy condition, the "Uni + Mix" method makes it possible to improve the naturalness. Figure 4 shows the experimental results for the perceptual speech intelligibility of the enhanced speech. In the clean condition, all methods are almost comparable to each other. On the other hand, the "Uni + Mix" method yields better performance compared to the others in the noisy condition.

From these results, it can be said that the "Uni + Mix" method yields better performance in the noisy condition while retaining the same parameter size compared to "Uni" though there are no improvements of the subjective result in the clean condition. It is assumed that it is still difficult to produce natural prosodic features due to less modeling capability of the uni-directional recurrent layer though the objective measures are improved.

V. Conclusion

In this paper, we have proposed a low-latency speech enhancement technique for electrolaryngeal (EL) speech based on multi-task CLDNN consisting of convolutional, uni-directional recurrent, and fully connected layers. Moreover, to improve the performance of uni-directional modeling, we proposed several techniques such as knowledge distillation, data augmentation, and phonetic regularization. The results of objective and subjective evaluations have demonstrated that the proposed method makes it possible to achieve comparable performance even using uni-directional modeling in the clean condition. Moreover, the proposed method yields better naturalness and perceptual speech intelligibility in the noisy condition. In future work, we are planning to incorporate neural vocoders into the low-latency speech enhancement for the EL speech.

Acknowledgment

This work was partly supported in part by JSPS KAKENHI Grant-in-Aid for JSPS Research Fellow Number 19K20295, and by JST, CREST Grant Number JPMJCR19A3.

References

[1] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A hybrid approach to electrolaryngeal speech enhancement based on noise reduction and statistical excitation generation," *IEICE Trans. Inf. Syst.*, vol. E97.D, no. 6, pp. 1429–1437, 2014.

[2] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[3] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *J. Acoust. Soc. Jpn. (E)*, vol. 11, no. 2, pp. 71–76, 1990.

[4] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Trans. ASLP*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.

[5] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. SAP*, vol. 27, no. 2, pp. 113–120, 1979.

[6] Y. Ephraim and H. L. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. SAP*, vol. 3, no. 4, pp. 251–266, 1995.

[7] B. L. Sim, Y. C. Tong, J. S. Chang, and C. Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Trans. ASLP*, vol. 6, no. 4, pp. 328–337, 1998.

[8] K. K. Wojcicki, B. J. Shannon, and K. K. Paliwal, "Spectral subtraction with variance reduced noise spectrum estimates," *Proc. SST*, 2006.

[9] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," *Speech Commun.*, vol. 54, no. 1, pp. 134–146, 2012.

[10] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "An evaluation of alaryngeal speech enhancement methods based on voice conversion techniques," *Proc. ICASSP*, pp. 5136–5139, May 2011.

[11] T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *Proc. INTERSPEECH*, pp. 1076–1079, Sept. 2008.

[12] J. L. Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," *Proc. Odyssey*, pp. 195–202, June 2018.

[13] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Trans. ASLP*, vol. 26, no. 1, pp. 84–96, Jan. 2018.

[14] K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda, "Statistical voice conversion with WaveNet-based waveform generation," *Proc. INTERSPEECH*, pp. 1138–1142, Aug. 2017.

[15] J.-X. Zhang, Z.-H. Ling, L.-J. Liu, Y. Jiang, and L.-R. Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *IEEE/ACM Trans. TASLP*, vol. 27, no. 3, pp. 631–644, 2019.

[16] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, "AttS2S-VC: Sequence-to-sequence voice conversion with attention and context preservation mechanisms," *Proc. ICASSP*, pp. 6805–6809, 2019.

[17] R. Arakawa, S. Takamichi, and H. Saruwatari, "Implementation of DNN-based real-time voice conversion and its improvements by audio data augmentation and mask-shaped device," *Proc. SSW10*, pp. 93–98, 2019.

[18] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T.Toda, "Voice Transformer Network: Sequence-to-Sequence Voice Conversion Using Transformer with Text-to-Speech Pretraining," *arXiv preprint arXiv:1912.06813*, 2019.

[19] J. Zhang, Z. Ling, and L.-R. Dai, "Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations," *IEEE/ACM Trans. ASLP*, vol. 28, no. 1, pp. 540–552, 2020.

[20] K. Kobayashi and T. Toda, "Electrolaryngeal speech enhancement with statistical voice conversion based on CLDNN," *Proc. EUSIPCO*, pp. 2115–2119, 2018.

[21] L. Serrano, D. Tavarez, X. Sarasola, S. Raman, I. Saratxaga, E. Navas, and H. Inma, "LSTM based voice conversion for laryngectomees," *Proc. IberSPEECH*, pp. 122–126.

[22] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," *Proc. ICASSP*, pp. 4580–4584, Apr. 2015.

[23] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *NIPS Deep Learning and Representation Learning Workshop*, pp. 1–9, 2014.

[24] T. Furlanello, Z. C. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, "Born again neural networks," *Proc. ICML*, pp. 1602–1611, 2018.

[25] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. Cubuk, and Q. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *Proc. INTERSPEECH*, pp. 2613–2617, 2019.

[26] G. Kurata and K. Audhkhasi, "Improved knowledge distillation from bi-directional to uni-directional LSTM CTC for end-to-end speech recognition," *Proc. SLT*, pp. 411–417, 2018.

[27] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," *Proc. ICME*, pp. 1–6, 2016.

[28] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," *IEEE workshop on automatic speech recognition and understanding*, 2011.

[29] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans on ASLP*, vol. 18, no. 8, pp. 2067–2079, 2010.

[30] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Electron. Commun. Jpn. (Part I: Commun.)*, vol. 66, no. 2, pp. 10–18, 1983.