Electrolaryngeal Speech Enhancement with Statistical Voice Conversion based on CLDNN

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

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

***Abstract*—An electrolarynx (EL) is a widely used device to mechanically generate excitation signals, making it possible for laryngectomees to produce EL speech without vocal fold vibrations. Although EL speech sounds relatively intelligible, is significantly less natural than normal speech owing to its mechanical excitation signals. To address this issue, a statistical voice conversion (VC) technique based on Gaussian mixture models (GMMs) has been applied to EL speech enhancement. In this technique, input EL speech is converted into target normal speech by converting spectral features of the EL speech into spectral and excitation parameters of normal speech using GMMs. Although this technique makes it possible to significantly improve the naturalness of EL speech, the enhanced EL speech is still far from the target normal speech. To improve the performance of statistical EL speech enhancement, in this paper, we propose an EL-to-speech conversion method based on CLDNNs consisting of convolutional layers, long short-term memory recurrent layers, and fully connected deep neural network layers. Three CLDNNs are trained, one to convert EL speech spectral features into spectral and band-aperiodicity parameters, one to convert them into unvoiced/voiced symbols, and one to convert them into continuous F_0 patterns. The experimental results demonstrate that the proposed method significantly outperforms the conventional method in terms of both objective evaluation metrics and subjective evaluation scores.**

***Index Terms*—electrolaryngeal speech, statistical voice conversion, speech enhancement, deep neural network**

I. Introduction

Speech communication plays an important role in human-to-human communications. Speech signals uttered by a speaker can convey not only linguistic information but also paralinguistic information, such as his/her individuality and emotions, to listeners. In a speech production mechanism, two main parts based on physical modeling of a human body are considered, an excitation generation part and a resonance part. In the excitation generation part, source excitation sounds are generated by vibrations of the vocal folds. Then, in the resonance part, they are convoluted with the acoustic resonance characteristics of a vocal tract shape determined by articulatory configurations. Therefore, speech disorders are caused if part of the speech organs is injured. A typical cause of speech disorders is a laryngectomy, which is surgery to remove the larynx including the vocal folds to treat laryngeal cancer, making a person lose the ability to produce source excitation sounds. People who have undergone this surgery are called laryngectomees.

To produce speech signals without using vocal fold vibrations, a device called an electrolarynx (EL) is widely used by laryngectomees to mechanically generate excitation signals from outside of their bodies. These excitation signals are conducted into the speakers' oral cavity and articulated to produce speech sounds. Although the produced speech called electrolaryngeal speech (EL speech) is relatively intelligible [1], there are two main problems. First, to make EL speech intelligible, the EL must generate quite loud excitation signals. As a result, the excitation signals are easily emitted outside as noise sounds, reducing the quality and intelligibility of the EL speech. The other problem is unnatural sounds in EL speech caused by the mechanically generated excitation signals. The acoustic characteristics of EL speech are very different from those of natural speech, for example, different spectral envelopes, different aperiodic components, no unvoiced sounds, and unnaturally varying F_0 patterns. 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]. The noise suppression approach focuses on reducing the noise components leaked from the excitation signals generated by the EL. Several techniques based on fundamental noise suppression methods, such as spectral subtraction [4]–[6] and Wiener filtering [7] have been proposed. Although these techniques are effective for reducing the noise components and making the EL speech more intelligible, the enhanced EL speech suffers from musical noise caused by the processing for noise suppression. Moreover, the improvements of EL speech yielded by this approach are limited because most of the acoustic characteristics of EL speech are not modified, and therefore, those of the enhanced EL speech are still very different from those of natural speech. On the other hand, the VC-based approach directly modifies these acoustic characteristics of EL speech, including spectral envelopes, aperiodicities, unvoiced/voiced information, and F_0 patterns [8], [9]. In this technique, acoustic features extracted from EL speech are converted into those of target natural speech using a conversion function trained with parallel data consisting of utterance pairs of the EL speech and the target natural speech, and converted speech signals are generated from the converted acoustic features. It has been reported that the naturalness is significantly improved by this approach [8], [9]. However, this approach suffers from various errors, such as conversion and modeling errors. Consequently, the converted speech (i.e., the

ISBN 978-90-827970-1-5 © EURASIP 2018

enhanced EL speech) is still far from natural. Therefore, it is desirable to develop a VC technique to convert EL speech to more natural-sounding speech.

Recently, statistical VC techniques have been significantly improved in the area of speaker individuality conversion [10]. As widely observed in areas of machine learning, deep learning (DL) methods are effective for improving the conversion accuracy of the statistical VC techniques [11]–[17]. In conventional EL speech enhancement, relatively traditional statistical VC techniques based on Gaussian mixture models (GMMs) and the maximum likelihood parameter generation (MLPG) method [18] are used. Therefore, it is expected that the performance of EL speech enhancement will be significantly improved by applying the state-of-the-art DL methods.

In this paper, we propose an EL speech enhancement technique based on CLDNNs [19] consisting of convolutional layers, long short-term memory recurrent layers, and fully connected layers. The convolutional layers are used to effectively extract useful information in the conversion from a spectral sequence of the EL speech. The recurrent layers are used to model the dynamic characteristics of speech parameters. The fully connected layers are used to model nonlinear mappings of speech features between the EL speech and natural speech. We conduct both objective and subjective evaluations to investigate the effectiveness of the proposed method.

II. EL speech enhancement by GMM-based VC

In the EL speech enhancement by GMM-based VC, a segmental feature vector transformed from an input EL mel-cepstrum sequence by principal component analysis (PCA) is converted into acoustic features of target normal speech such as unvoiced/voiced symbols, F_0 patterns, aperiodicities, and the mel-cepstrum using GMMs as conversion models. These conversion models are separately trained using joint feature vectors of the segmental feature vector and each target feature vector. Finally, the converted speech is generated using a vocoder with the converted acoustic features. In this section, we describe the training process and conversion process.

A. Training process

In the training process, joint probability density functions of acoustic features of the EL speech and target normal speech are modeled with GMMs using their parallel data set. As the acoustic features, we employ the segmental feature vector $\boldsymbol{X}_t$ of the EL speech and the $2D$-dimensional joint static and delta feature vector $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$ of the normal speech consisting of a D-dimensional static feature vector $\boldsymbol{y}_t$ and its dynamic feature vector $\Delta\boldsymbol{y}_t$ at frame t, where $\top$ denotes the transpose of the vector. Each joint probability density modeled by the GMM is given by

$$\begin{aligned} & P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t | \boldsymbol{\lambda}\right) \\ & = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix}\boldsymbol{X}_t \\ \boldsymbol{Y}_t\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_m^{(XX)} \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(YY)}\end{bmatrix}\right), \end{aligned} \quad (1)$$

where $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The mixture component index is m. The total number of mixture components is M. $\boldsymbol{\lambda}$ is a GMM parameter set consisting of the mixture-component weight α_m, mean vector $\boldsymbol{\mu}_m$, and covariance matrix $\boldsymbol{\Sigma}_m$ of the mth mixture component. The GMM is trained using joint vectors $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ in the parallel data set, which are automatically aligned to each other by dynamic time warping using mel-cepstrum distortion as a distance measure.

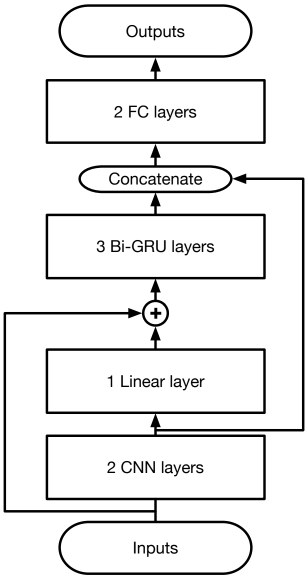

Fig. 1. Network architecture of CLDNN.

B. Conversion process

In the conversion process, the segmental feature vector of the EL speech is converted into target acoustic features based on MLPG with the GMMs [18]. In this research, GV constraint [18] for the parameter generation is applied to the estimation of mel-cepstrum as the target feature.

III. EL speech enhancement with CLDNN-based VC

To improve the naturalness and intelligibility of the enhanced speech, we employ CLDNNs as the conversion function in the EL speech enhancement. The CLDNN was originally proposed as an acoustic model in automatic speech recognition, which enabled significant improvements in speech recognition accuracy [19]. The original CLDNN consisted of convolutional neural network (CNN) layers, long short-term memory (LSTM) recurrent layers, and fully connected (FC) layers with two skipped connections. Although the effectiveness of the CLDNN has been confirmed in speech recognition, it has never been applied to statistical VC tasks including EL speech enhancement. As there are some differences between a classification task in speech recognition and a regression task in EL speech enhancement, it is worth applying the CLDNNs to EL speech enhancement and investigate its effectiveness.

A. Network architecture

Figure 1 shows the network architecture of the CLDNN for EL speech enhancement. For the inputs of the first CNN layer, a one-dimensional feature vector at frame t is transformed

ISBN 978-90-827970-1-5 © EURASIP 2018

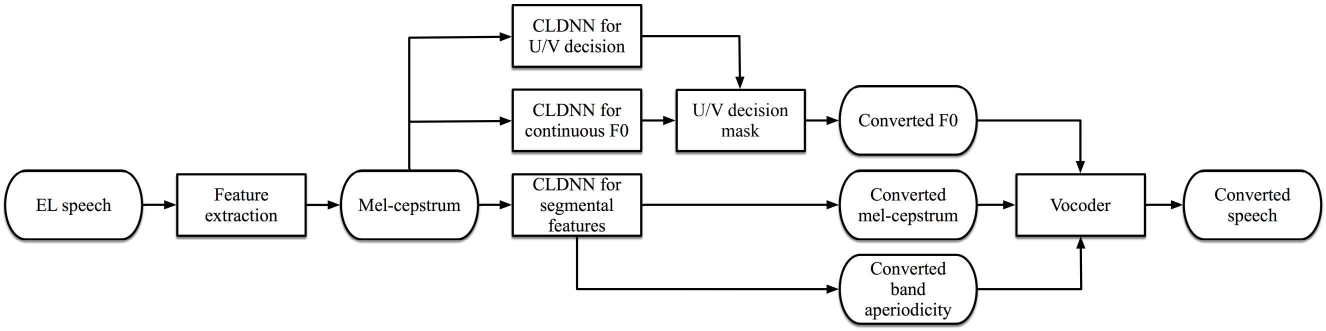

Fig. 2. Conversion process of EL speech enhancement with CLDNN-based VC.

into a two-dimensional feature matrix by concatenating several preceding and succeeding frames to capture contextual information from the spectral envelope sequence of the EL speech, which is essential to achieve the conversion from unnatural speech features into natural speech features in EL speech enhancement. Note that PCA is used in the conventional method but the CNN is used in the proposed method. Therefore, nonlinear and trainable feature extraction is achieved. 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 CNN layers. Then, the resulting outputs are fed into the recurrent layers. For the recurrent layers, we use bidirectional gated recurrent units (Bi-GRUs) to reduce the number of parameters from that in the original implementation of the LSTM. The outputs of the Bi-GRU layers are concatenated into those of the CNN layers. Finally, the resulting outputs are fed into the FC layers to be transformed into the output feature vector.

B. Training and conversion processes

In the training process, three CLDNNs are trained. The mel-cepstrum and aperiodicities, are modeled by a single CLDNN by concatenating these acoustic features. For the prosodic features, continuous F_0 and unvoiced/voiced (U/V) symbols are modeled separately because F_0, consisting of continuous values and unvoiced/voiced symbols, is difficult to model directly.

Figure 2 shows the conversion process of the proposed EL speech enhancement. In the conversion process, the mel-cepstrum extracted from the EL speech is converted into U/V symbols, a continuous F_0, the mel-cepstrum, and aperiodicities by separate 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 vocoding using these acoustic features.

IV. Experimental evaluation

In this evaluation, we compared the conventional EL speech enhancement based on GMM-based VC and the proposed EL speech enhancement based on CLDNN-based VC.

A. Experimental conditions

We used 991 Japanese sentences. One healthy male Japanese speaker uttered EL speech, which was recorded after carefully learning how to use an EL to produce EL speech. He also uttered normal speech as target natural speech by mimicking the prosodic characteristics of his recorded EL speech. The sampling frequency was set to 16 kHz. STRAIGHT [20] was used to extract spectral envelopes, which were parameterized to the 0-24th mel-cepstral coefficients as spectral features. As the source excitation features, we used a log-scaled F_0 value, the U/V symbol, and aperiodic components in five frequency bands, i.e., 0-1, 1-2, 2-4, 4-6, and 6-8 kHz, which were also extracted using STRAIGHT [21]. The frame shift was 5 ms. The MLSA filter [22] was used as the synthesis filter.

We set the number of training utterances to 64, 254, or 762, and the other 229 utterances were used for the evaluation. In the conventional method, the number of mixture components for mel-cepstrum conversion was set to 16 for 64 and 254 training utterances and 32 for 762 training utterances, and the number for aperiodicity estimation and F_0 estimation was set to 16. The other parameters in the conventional method were set to the values in [1].

For the inputs of the first CNN layer, one-dimensional feature vectors were extended by concatenating 10 preceding and 10 succeeding feature vectors to obtain 21 × 25 two-dimensional feature matrices. Two-dimensional convolutions were performed in each CNN layer using 5 × 5 and 3 × 3 kernels and 2 and 0 zero paddings with batch normalization and max pooling, respectively. The rectified linear unit was used for the activation functions of the CNN layers. In the Bi-GRU layers, the number of hidden units was set to 256 and 5% dropout was applied in each layer. For the FC layers, the number of hidden units was set to 256 and the sigmoid function was used as the activation function except for the output layer. The mean squared error was used for the loss functions except for the U/V symbols, and binary cross entropy was used for the U/V decision. For the estimated U/V symbols, we regarded predicted probabilities of over 0.5 as corresponding to the voiced frames and the other values as corresponding to unvoiced frames. All feature vectors except for the U/V symbols were transformed so as to have zero mean and unit variance in each dimension before CLDNN modeling.

The weight parameters were initialized using Xavier [23], and the biases were initialized to zero. The learning rate was set to 0.05 for mel-cepstrum and aperiodic features and to 0.0005 for prosodic features (i.e., continuous F_0 and U/V

TABLE I
OBJECTIVE EVALUATIONS OF CONVERTED F_0.

Method	Log F_0 RMSE	F_0 correlation coefficient
GMM (64)	0.067	0.63
GMM (254)	0.058	0.69
GMM (762)	0.060	0.67
CLDNN (64)	0.062	0.64
CLDNN (254)	0.053	0.68
CLDNN (762)	0.057	0.73

TABLE II
CONFUSION MATRIX OF U/V DECISION FOR THE CONVENTIONAL METHOD.

	GMM (64)		GMM (254)		GMM (762)	
	U_{est}	V_{est}	U_{est}	V_{est}	U_{est}	V_{est}
U_{tar}	0.79	0.21	0.90	0.10	0.92	0.08
V_{tar}	0.04	0.96	0.13	0.87	0.10	0.90

TABLE III
CONFUSION MATRIX OF U/V DECISION FOR THE PROPOSED METHOD.

	CLDNN (64)		CLDNN (254)		CLDNN (762)	
	U_{est}	V_{est}	U_{est}	V_{est}	U_{est}	V_{est}
U_{tar}	0.90	0.10	0.94	0.06	0.89	0.11
V_{tar}	0.08	0.92	0.10	0.90	0.03	0.97

TABLE IV
OBJECTIVE EVALUATIONS FOR SEGMENTAL FEATURES.

Method	Band-aperiodicity RMSE [dB]	Mel-CD [dB]
GMM (64)	4.03	7.40
GMM (254)	3.52	6.15
GMM (762)	3.35	5.57
CLDNN (64)	3.75	6.66
CLDNN (254)	3.27	5.66
CLDNN (762)	3.13	5.22

symbols). The mini-batch size for the segmental features (i.e., mel-cepstrum and aperiodicities) was 64 frames and that for the prosodic features was 128 frames. Stochastic gradient descent was used to optimize the network parameters. The number of epochs was set to 50. 10% of the training data was used as development data and the parameters achieving the minimum development loss were used for the evaluation.

B. Objective evaluations

As objective evaluations, we compared objective measures of the converted acoustic features based on the root mean square error (RMSE), correlation coefficients, confusion matrix, and mel-cepstrum distortion (Mel-CD). For the converted F_0, the RMSE of the logarithmic F_0, the correlation coefficient, and the confusion matrix for U/V symbols were compared. Note that the RMSE and correlation coefficients were calculated only using voiced frames. For the segmental features, the RMSE and Mel-CD were evaluated. The Mel-CD is calculated as

$$\text{Mel-CD } [dB] = \frac{10}{\ln 10}\sqrt{2\sum_{d=0}^{24}\left(mc_d^{(X)} - mc_d^{(Y)}\right)^2}, \qquad (2)$$

where $mc_d^{(X)}$ and $mc_d^{(Y)}$ represent the d^{th}-dimensional component of the converted mel-cepstrum and that of the target mel-cepstrum, respectively.

Table I gives the logarithmic F_0 RMSE and F_0 correlation coefficients between the converted and target F_0. Note that as we used normal voices with unnatural F_0 patterns similar to those of EL speech as the target natural speech, these results do not directly demonstrate the prosodic naturalness of the enhanced EL speech. However, it is still possible to compare the conversion accuracy between the two methods and examine its tendency with increasing of the amount of training data. Tables II and III show the confusion matrices of the estimated U/V symbols and target U/V symbols for the conventional and proposed methods, respectively. With increasing number of training utterances, the U/V decision accuracy tends to increase for both methods. CLDNN (762) achieves the best accuracy, and even when using few training utterances, and CLDNN (64) achieves similar accuracy to GMM (762). These results demonstrated that U/V estimation by the proposed method is better than that by the conventional method.

Table IV gives RMSEs of the converted band aperiodicities and Mel-CDs. For the segmental features, the proposed method obviously outperforms the conventional method when using not only a large number of training utterances but also a small number of training utterances. Moreover, because CLDNN (254) achieves similar performance to GMM (762), it can be concluded that the proposed method is capable of significantly reducing the amount of training data while maintaining comparable conversion performance to the conventional method.

C. Subjective evaluations

For the subjective evaluations, two preference tests were conducted. In the first test, the naturalness of the enhanced EL speech was evaluated using a mean opinion score (MOS). The enhanced speech samples generated by the conventional and proposed methods and analysis/synthesis of the target natural speech as a reference 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 80. The number of subjects was six. In the second test, the perceptual speech intelligibility of the enhanced speech was evaluated in the same manner as the naturalness. Note that this score was different from the intelligibility score determined by conducting a manual dictation test, but it still showed how easily the linguistic contents of a speech sample can be listened to.

Figure 3 shows the experimental results for the naturalness of the enhanced speech. The proposed method makes it possible to achieve significant improvements in the naturalness not only for a large amount of training data (i.e., 762 training utterances) but also a small amount of training data (i.e., 64 training utterances).

Figure 4 shows the experimental results for the perceptual speech intelligibility of the enhanced speech. Although the difference between the conventional method and proposed method is small when the number of training utterances is small, the proposed method is capable of significantly improving the perceptual speech intelligibility when a large amount of training data can be used. On the other hand, the difference

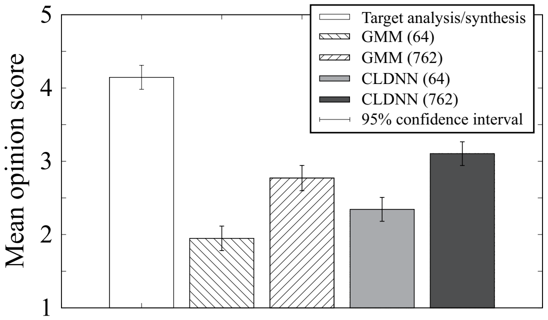

Fig. 3. Results for naturalness.

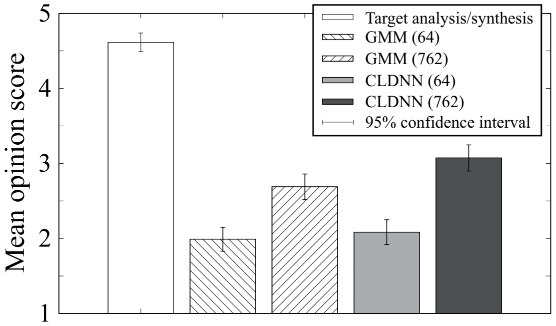

Fig. 4. Results for perceptual speech intelligibility.

in the perceptual speech intelligibility between the enhanced speech and the reference speech (i.e., analysis/synthesis) is still large compared with that in the naturalness. Therefore, it is still difficult to achieve high perceptual speech intelligibility even when using the sophisticated conversion model based on CLDNNs.

V. Conclusion

In this paper, we have proposed a speech enhancement technique for electrolaryngeal (EL) speech by converting EL speech into normal speech using CLDNNs consisting of convolutional, long short-term memory recurrent, and fully connected layers. Although the conventional EL speech enhancement technique based on Gaussian mixture models (GMMs) improves naturalness, its converted speech is still far from normal speech. To address this issue, in this paper, we have applied the CLDNNs proposed for automatic speech recognition to EL speech enhancement. The results of objective and subjective evaluations have demonstrated that the proposed method is capable of converting EL speech into normal speech with higher naturalness and perceptual speech intelligibility than those obtained by the conventional method. In future work, we will implement an EL speech enhancement technique without using traditional vocoding framework to avoid modeling errors of waveform signals by incorporating the WaveNet vocoder [24] as a waveform generator.

Acknowledgments

This work was supported in part by JSPS KAKENHI Grant-in-Aid for JSPS Research Fellow Number 16J10726, and by JST, PRESTO Grant Number JPMJPR1657.

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

[5] 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.

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

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

[8] 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.

[9] 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.

[10] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The Voice Conversion Challenge 2016," *Proc. INTERSPEECH*, Sept. 2016.

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

[12] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Trans. ASLP*, vol. 22, no. 12, pp. 1859–1872, Dec. 2014.

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

[14] 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," *Proc. APSIPA*, pp. 1–6, Dec. 2016.

[15] I. McLoughlin, J. Li, Y. Song, and H. R. Sharifzadeh, "Speech reconstruction using a deep partially supervised neural network," *IEEE Healthcare Technol. Lett.*, vol. 4, no. 4, pp. 129–133, 2017.

[16] 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.

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

[18] 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.

[19] 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.

[20] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f_0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, no. 3-4, pp. 187–207, Apr. 1999.

[21] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and system straight," *Proc. MAVEBA*, Sept. 2001.

[22] 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.

[23] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *Proc. AI & Statistics*, vol. 9, pp. 249–256, 13–15 May 2010.

[24] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," *Proc. INTERSPEECH*, pp. 1118–1122, Aug. 2017.