DNN-based Voice Conversion with Auxiliary Phonemic Information to Improve Intelligibility of Glossectomy Patients' Speech

Hiroki Murakami*, Sunao Hara† and Masanobu Abe‡
* Okayama University, Japan
E-mail: h_muraka@a.cs.okayama-u.ac.jp
† Okayama University, Japan
E-mail: hara@okayama-u.ac.jp
‡ Okayama University, Japan
E-mail: abe@cs.okayama-u.ac.jp

***Abstract*—In this paper, we propose using phonemic information in addition to acoustic features to improve the intelligibility of speech uttered by patients with articulation disorders caused by a wide glossectomy. Our previous studies showed that voice conversion algorithm improves the quality of glossectomy patients' speech. However, losses in acoustic features of glossectomy patients' speech are so large that the quality of the reconstructed speech is low. To solve this problem, we explored potentials of several additional information to improve speech intelligibility. One of the candidates is phonemic information, more specifically Phoneme Labels as Auxiliary input (PLA). To combine both acoustic features and PLA, we employed a DNN-based algorithm. PLA is represented by a kind of one-of-k vector, i.e., PLA has a weight value (<1.0) that gradually changes in time axis, whereas one-of-k has a binary value (0 or 1). The results showed that the proposed algorithm reduced the mel-frequency cepstral distortion for all phonemes, and almost always improved intelligibility. Notably, the intelligibility was largely improved in phonemes /s/ and /z/, mainly because the tongue is used to sustain constriction to produces these phonemes. This indicates that PLA works well to compensate the lack of a tongue.**

I. Introduction

Speech is the primary means of communication for human beings and plays a crucial role in maintaining one's quality of life in everyday life. This is also true for individuals with speech production problems. In this context, intensive studies have been performed to facilitate improvements in the speech of patients with tongue resection or tongue movement disorders [1], [2], [3].

As a new approach from a speech processing point of view, we proposed to improve speech quality uttered by glossectomy patients using voice conversion algorithms [4], [5], [6]. Voice conversion (VC) [7], [8], [9], [10] is a technique to modify one speaker's voice to another speaker while keeping its linguistic information unchanged. To improve intelligibility of glossectomy patients' speech, we recruited a glossectomy patient as a source speaker and a professional narrator or a healthy speaker as a target speaker. Our previous studies showed that acoustic features mapping based on VC improves speech intelligibility [4], [5] and direct waveform modification using spectrum differential improves the naturalness of the reconstructed speech [6]. However, the quality of the reconstructed speech was not satisfactory.

To improve the speech intelligibility, in this paper, we propose using phonemic information in addition to acoustic features. A motivation and a basic idea are as follows. Because speech uttered by glossectomy patients is quite different from that of a healthy speaker, acoustic features extracted on a frame-by-frame basis are sometimes not good enough to identify phonemes, which results in failures of finding feature correspondences using parallel corpus. To reduce the ambiguity, a possible solution is to use consecutive acoustic features or segment features. In other words, a longer period of observation enables to deal with co-articulation phenomena, which results in disambiguation.

There are several existing studies that uses a kind of phonemic information for VC. For example, in [11] and [12], phoneme labels coded by HMM are used in VC and speech coding, respectively. In [13], Phonetic Posteriorgrams (PPGs) are used in VC. These papers support the effectiveness of employing the phonemic information. Therefore, we employed the phonemic information to improve the intelligibility of glossectomy patients' speech.

In this paper, we examine performances under ideal conditions to make sure the potential of the basic idea; i.e. supposing that phonemic information is provided in advance. Moreover, we employ a DNN-based algorithm to combine both acoustic features and phonemic information. Through a subjective experiment, we show how the proposed algorithm works well to reconstruct spectrum features from perceptual point of view.

The rest of the paper is organized as follows. In Section 2, we describe the phonemic information used for training. In Section 3, we explain the algorithm that uses both acoustic features and phonemic information. In Section 4, we present our evaluation results and a discussion. Finally, in Section 5, we present our conclusions and suggest avenues for future work.

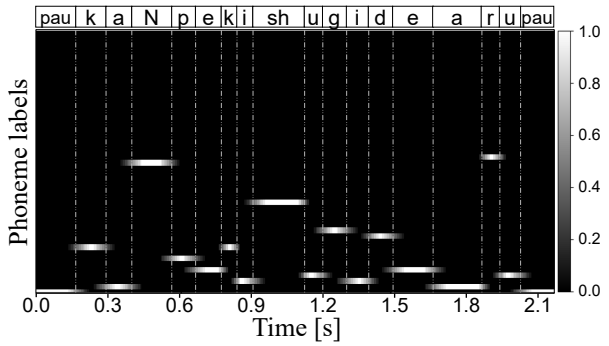

Fig. 1. Example of phoneme information. It is consisting of frame-by-frame vector elements corresponding to phoneme labels, which are represented at a top of the figure.

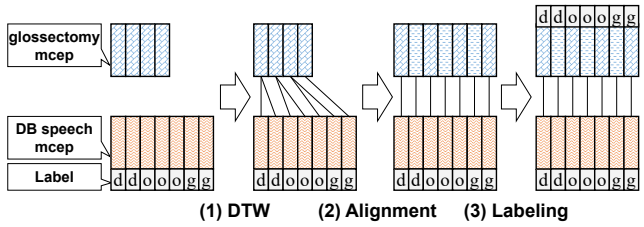

Fig. 2. Outline of assignment of phoneme by dynamic time warping (DTW).

II. Phonemic information for the voice conversion

As the phonemic information, we use 45 kinds of Phoneme Labels as Auxiliary input (PLA). The phoneme labels are as follows: vowels (i, e, a, o, u), stops (p, t, k, b, d, g), fricatives (s, z, sh, h, f, j), affricates (ch, ts), nasals (m, n, N), liquid (r), semivowels (y, w), contracted sounds (by, gy, hy, ky, my, ny, py, ry), double consonants (pp, tt, kk, dd, ss, ff, tts, cch, ssh, kky, ppy), and pause.

Figure 1 shows an example of phonemic information, or PLA, generated from phoneme labels. It is consisting of a frame-by-frame time sequence of 45 dimension vectors. The phoneme information is generated by phoneme label assignment with DTW and post processing as explained in the following subsections.

A. Phoneme label assignment with DTW

Figure 2 shows the flow of phoneme label assignment for glossectomy patient's speech whose phoneme is known but time information is unknown. Our proposed method used ATR speech database, which contains phoneme annotations with precise time information. First, time alignment is carried out by DTW between the acoustic features of the glossectomy patient's speech (glossectomy mcep) and those of the ATR speech database (DB speech mcep). Here, the length of the glossectomy patient's speech is changed to match that of the DB speech. Finally, phoneme labels are assigned to the glossectomy patient's acoustic features according to the DTW path.

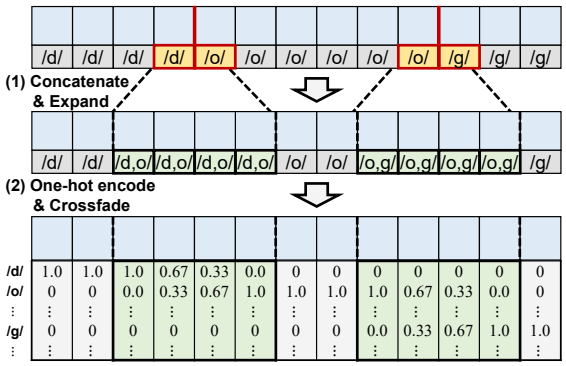

Fig. 3. Generating process of phonemic information from phoneme labels. Our phonemic information is like as one-of-k vector, but it changes gradually at around each phoneme boundaries ($N = 2$ frames).

B. Phoneme information generation with post processing

In general, acoustic features gradually change due to co-articulation, therefore, the associated phonemic information should be also changed gradually as similar as the acoustic features. Figure 3 shows an example of a generating process of the phoneme information from frame-by-frame phoneme labels estimated by previous Subsection II-A.

Around a phoneme boundary for N frames, the weight of one-of-k vector is linearly interpolated from 1 to 0, or vice versa. Note that the figure shows the example for the parameter N = 2.

III. Voice conversion using phoneme labels as auxiliary input

The voice conversion system is divided into two parts, a training part and a conversion part. This system is based on a spectral differential modification method [14] with the DNN conversion model [6].

A. Training

Figure 4 presents an outline of the training process. The process is divided into a parallel corpus generation and a training component of a conversion model.

The process of generating parallel corpus is as following. First, the acoustic features of the source speaker and target speaker are extracted by speech analysis. Here, the dynamic feature $\Delta \boldsymbol{x}_t$ is calculated from the static feature $\boldsymbol{x}_t$ of the source speaker in a frame t. The dynamic feature $\Delta \boldsymbol{y}_t$ is also calculated from $\boldsymbol{y}_t$ as same as the calculation of $\Delta \boldsymbol{x}_t$. The static features and the dynamic features are concatenated and used as a feature vector. The source speaker's acoustic feature vector is described as $\boldsymbol{X}_t = [\boldsymbol{x}_t^{\mathrm{T}}, \Delta \boldsymbol{x}_t^{\mathrm{T}}]^{\mathrm{T}}$, and target speaker's acoustic feature vector is described as $\boldsymbol{Y}_t = [\boldsymbol{y}_t^{\mathrm{T}}, \Delta \boldsymbol{y}_t^{\mathrm{T}}]^{\mathrm{T}}$. T denotes the transposition of a vector. Then, the static acoustic feature $\boldsymbol{z}_t$ of the ATR database speaker and the phoneme label $\boldsymbol{l}_t$ are described as $\boldsymbol{Z}_t = [\boldsymbol{z}_t^{\mathrm{T}}, \boldsymbol{l}_t^{\mathrm{T}}]^{\mathrm{T}}$. As described in the II-A section, phoneme labels are assigned to $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ using $\boldsymbol{Z}_t$. Finally, we obtain the source speaker's feature vector $\boldsymbol{X}'_t = [\boldsymbol{x}_t^{\mathrm{T}}, \Delta \boldsymbol{x}_t^{\mathrm{T}}, \boldsymbol{l}_t^{\mathrm{T}}]^{\mathrm{T}}$ and the target speaker's feature vector $\boldsymbol{Y}'_t =$

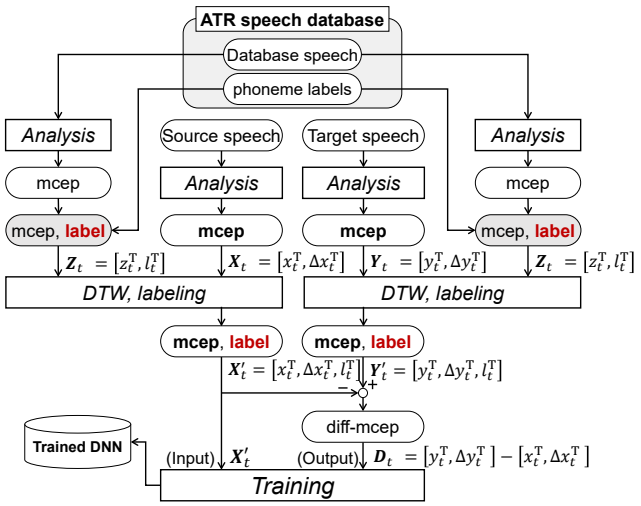

Fig. 4. Training outline.

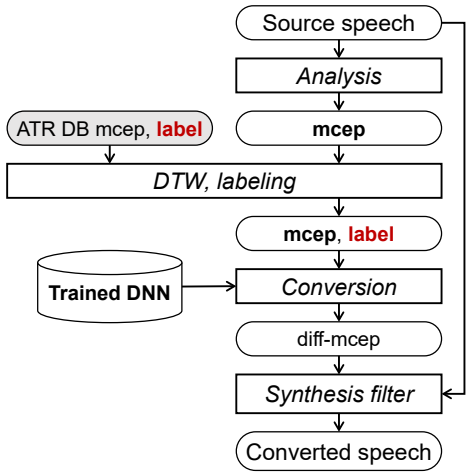

Fig. 5. Conversion outline.

$[\boldsymbol{y}_t^{\mathrm{T}}, \Delta\boldsymbol{y}_t^{\mathrm{T}}, \boldsymbol{l}_t^{\mathrm{T}}]^{\mathrm{T}}$. The source and target features are already aligned because the time axis of the ATR database speaker is fixed in DTW.

The process of training a VC model using the corpus is as following. The differential acoustic feature $\boldsymbol{D}_t = [\boldsymbol{y}_t^{\mathrm{T}}, \Delta\boldsymbol{y}_t^{\mathrm{T}}] - [\boldsymbol{x}_t^{\mathrm{T}}, \Delta\boldsymbol{x}_t^{\mathrm{T}}]$ is generated by subtraction between source feature and target feature. Finally, the function to map input feature $\boldsymbol{X}'_t$ to output feature $\boldsymbol{D}_t$ is trained by DNN [6].

B. Conversion

Figure 5 shows the conversation outline. During conversion, the source speaker's acoustic feature vector $\boldsymbol{X}_t$ is extracted by speech analysis. Next, we obtain $\boldsymbol{X}'_t$ as in the training step using ATR database speech. Here, the time axis of source speaker's feature is fixed. Next, differential acoustic feature vector $\hat{\boldsymbol{D}}_t$ convert to $\boldsymbol{X}'_t$ by trained DNN. Finally, converted speech is synthesized by directly filtering the input waveform using the $\hat{\boldsymbol{D}}_t$ by MLSA filter.

IV. Evaluation Experiments

A. Experimental conditions

For the training and validation dataset, we used 400 sentences and 50 sentences uttered phrase-by-phrase, respectively. For the evaluation dataset, we used 53 sentences uttered sentence-by-sentence. The sampling frequency was 20 kHz. The speaker is a healthy male person #1 (M1). To simulate a glossectomy patient's speech, we fabricated an intra-oral appliance that covers the lower dental arch and tongue surface to restrain tongue movements during speech [5]. The speaker 'M1' uttered speech with and without the appliance to simulate speech before and after a glossectomy. To fix notation, in the remainder of the paper speech uttered by M1 with the appliance is denoted SPM1 (Simulated Patient Male1).

To evaluate the proposed method, we compare two methods as follows:

- **DIFF-VC**: DNN-based VC using the spectral differential method [6] (baseline);
- **DIFF-PLA-VC**: DIFF-VC using Phoneme Labels as Auxiliary input (proposed).

Spectral envelopes were extracted by WORLD [15] and parameterized to the 0-25th mel-cepstral coefficients and their dynamic features. The frame shift was 5 ms. Mel log spectrum approximation (MLSA) filter [16] was used as the synthesis filter.

The PLA is a 45 dimensional vector obtained by processing introduced in Section II-B. The range of post-processing N was 5 frames. The input feature is 97 dimensional vector [mcep, Δmcep, PLA], and the output feature is 52 dimensional vector [diff_mcep, Δdiff_mcep].

Regarding DNN, we adopted multilayer perceptron (MLP) as the conversion model. In each layer, the number of units is set as [97, 1024, 1024, 1024, and 52]. The rectified linear units were used in the hidden layers, and the linear activation function was used in the output layer. The weights of the DNN were initialized randomly, and Adam was used for optimization.

B. Objective evaluation

Mel-cepstral distortion is used to objectively measure the spectral distance between converted speech and target speech. The feature used for the evaluation was extracted from the converted speech by speech analysis. Figure 6 shows the results of objective evaluation. There were 40 phonemes, excluding those not included in the evaluation data set. The proposed method (DIFF-PLA-VC) is better than the baseline method (DIFF-VC) in 36 out of 40. In particular, the proposed method has a great improvement in mel-cepstral distortion for fricatives, mainly because the tongue is used to sustain constriction for a while to produces fricatives.

C. Subjective evaluation

A dictation experiment was carried out to measure speech intelligibility. There were three types of speech in the experiment: the original simulated patient's speech (ORIG),

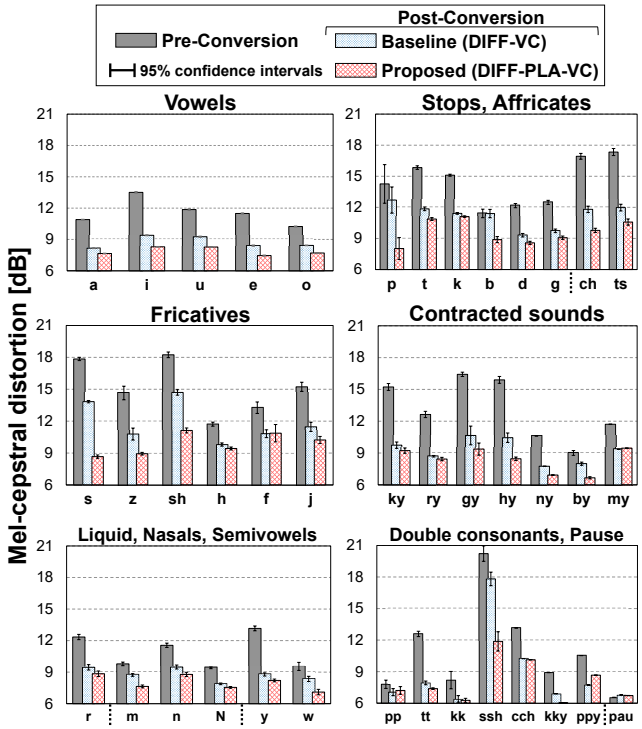

Fig. 6. Mel-cepstral distortion before and after VC.

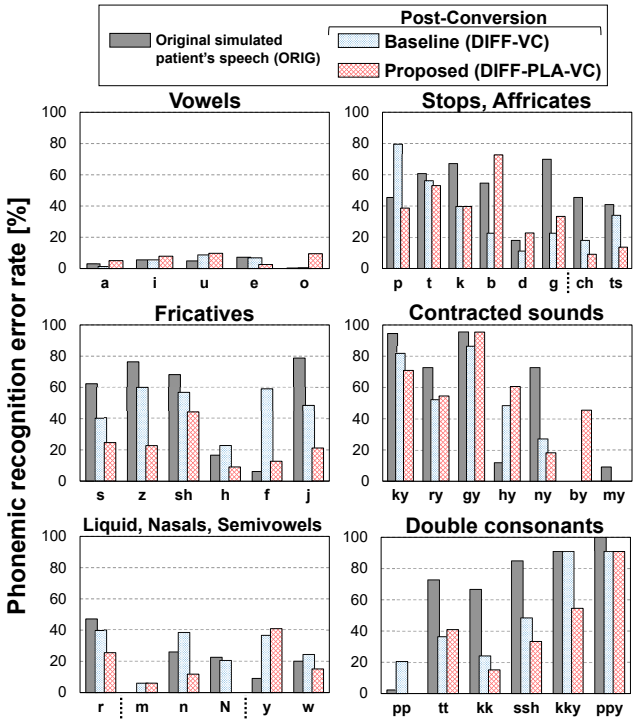

Fig. 7. Dictation experiment for speech intelligibility.

and two converted speech types: (DIFF-VC and DIFF-PLA-VC). In order to avoid guessing correct answers, healthy person's speech was not used in the subjective evaluation. For each speech type, 50 sentences were created. A total of 150 sentences were randomly shuffled. The 10 subjects listened to each speech and wrote Kana characters down as they were heard. The phonemic recognition error rate was calculated as follows. After decomposing the correct and answered sentence into syllables, the number of wrong phonemes was counted by comparing the correct syllables and the answered syllables for each phoneme. The error rate was calculated by dividing the number of wrong phonemes by the total number of phonemes. Note that if the correct and answered sentence pairs had different phoneme numbers, they were manually adjusted.

Figure 7 shows the results of subjective evaluation. In almost all phonemes, the intelligibilities were improved. Particularly in fricatives, the proposed method (DIFF-PLA-VC) is better than the baseline method (DIFF-VC). According to the results, we can say that proposed method is effective.

D. Comparison of the spectrograms

The causes of improving the phonemic intelligibility can be observed in spectrograms. Figure 8 shows spectrograms that compare the VC from SPM1 to M1 by DIFF-VC and DIFF-PLA-VC. As indicated in the regions surrounded by the red dotted lines, high-frequency components of fricative /s/ were weak in the input speech (b), however, it was reconstructed in the converted speech (c) and (d). Comparing (c) and (d), the proposed method (d) reconstructed the fricative more clearly than the baseline method (c). Thus, the proposed method, using phoneme labels as auxiliary input, works well to compensate the lack of a tongue.

V. Conclusions

We proposed an algorithm to improve intelligibility of reconstructed glossectomy patient's speech using phoneme labels as auxiliary input in DNN-based VC. During evaluation, it was found that the proposed method was better than the baseline method, especially for fricatives. This fact was clearly observed in the spectrograms, i.e., the proposed method could reconstruct stronger energy than the conventional method in a high-frequency band of fricatives.

Through the above experiments, we confirmed the potential of phoneme labels for improving intelligibility of speech uttered by the glossectomy patient. As mentioned in the introduction, we currently examine the performances under ideal conditions; phoneme labels are given in advance. As the next step, we have to estimate phoneme labels. To estimate phoneme labels as correctly as possible, we are now trying to estimate phoneme labels using not only speech signals but also lip movements and other biophysical signals.

There is another future work in our minds. As explained in the introduction, a motivation of introducing phonemic information is to use consecutive acoustic features or segment features. Because, in this paper, we confirmed the potential of segment features, we are interested in finding an appropriate way to express segment features instead of phoneme labels. The phoneme labels are so rigid to express segment features that we would like to express them more flexibly such as a set of parameters with probability density functions and so on.

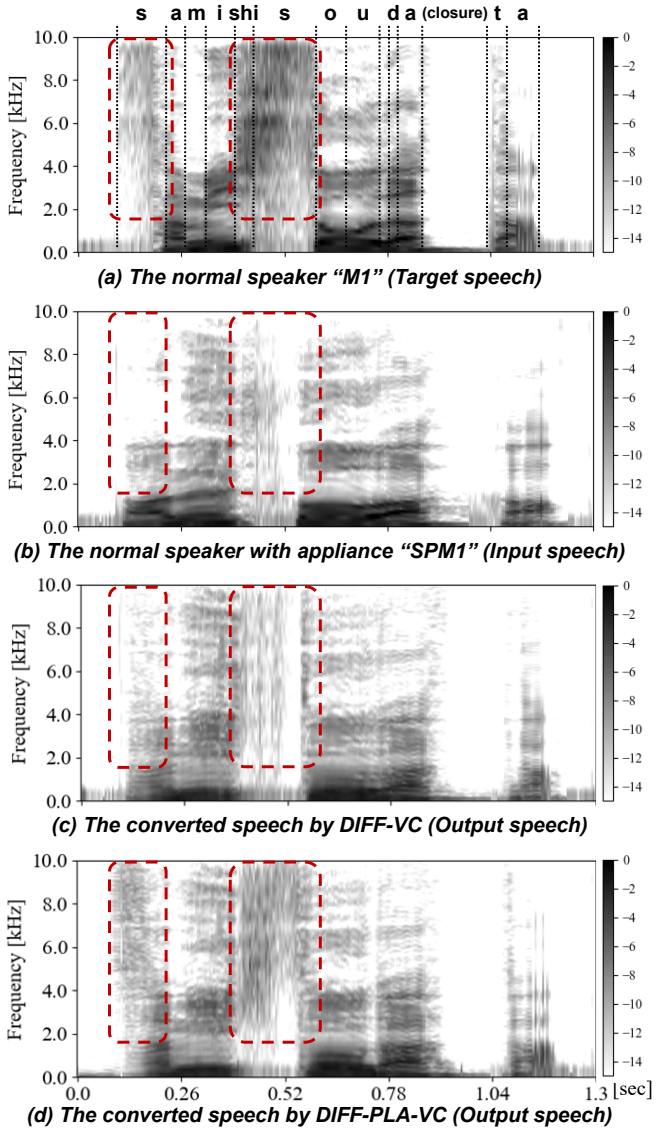

Fig. 8. Comparisons of the spectrograms.

VI. Acknowledgements

This work has been supported by JSPS KAKENHI 18K11376.

References

[1] R. Cantor, T. Curtis, T. Shipp, J. Beume, and B. Vogel, "Maxillary speech prostheses for mandibular surgical defects," *J. Prosthetic Dentistry*, vol. 22, pp. 253–260. (1969)

[2] R. Leonard, and R. Gillis, "Differential effects of speech prostheses in glossectomized patients," *J. Prosthetic Dentistry*, vol. 64, pp. 701–708. (1990)

[3] K. Kozaki, S. Kawakami, A. Gofuku, M. Abe, and S. Minagi *et al.*, "Structure of a new palatal plate and the artificial tongue for articulation disorder in a patient with subtotal glossectomy," *Acta Medica Okayama*, vol. 70, no. 3, pp. 205–211. (2016)

[4] K. Tanaka, S. Hara, M. Abe, and S. Minagi, "Enhancing a Glossectomy Patient's Speech via GMM-based Voice Conversion," *Proc. APSIPA Annual Summit and Conference*. (2016)

[5] K. Tanaka, S. Hara, M. Abe, M. Sato, and S. Minagi, "Speaker Dependent Approach for Enhancing a Glossectomy Patient's Speech via GMM-based Voice Conversion," *Proc. INTERSPEECH*, pp. 3384–3388. (2017)

[6] H. Murakami, S. Hara, M. Abe, M. Sato, and S. Minagi, "Naturalness Improvement Algorithm for Reconstructed Glossectomy Patient's Speech Using Spectral Differential Modification in Voice Conversion," *Proc. INTERSPEECH*, pp. 2464–2468. (2018)

[7] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *Proc. ICASSP*, S14.1, pp. 655–658. (1988)

[8] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142. (1998)

[9] A. Kain, and M. Macon, "Spectral voice conversion for text-to-speech synthesis," *Proc. ICASSP*, pp. 285–288. (1998)

[10] S. Desai, E.V. Raghavendra, B. Yegnanarayana, A.W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," *Proc. ICASSP*, pp. 3893–3896. (2009)

[11] M. Abe and S. Sagayama, "A segment-based approach to voice conversion," *Proc. ICASSP*, pp. 765–768. (1991)

[12] M. Abe, H. Mizuno, S. Takahashi, and S. Nakajima, "A prototype hybrid scalable text-to-speech system," *Proc. Workshop on SNHC and 3D Image*, pp. 8–11. (1997)

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

[14] K. Kobayashi, T. Toda, G. Neubig, and S. Nakamura, "Statistical singing voice conversion with direct waveform modification based on the spectrum differential," *Proc. INTERSPEECH*, pp. 2514–2518. (2014)

[15] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. on Information and Systems*, vol. 99, no. 7, pp. 1877–1884. (2016)

[16] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (mlsa) filter for speech synthesis," *Electronics and Communications in Japan (Part I: Communications)*, vol. 66, no. 2, pp. 10–18. (1983)