Interspeech 2018
2-6 September 2018, Hyderabad

Naturalness Improvement Algorithm for Reconstructed Glossectomy Patient's Speech Using Spectral Differential Modification in Voice Conversion

Hiroki Murakami[1], *Sunao Hara*[1], *Masanobu Abe*[1], *Masaaki Sato*[2], *Shogo Minagi*[2]

[1]Graduate School of Natural Science and Technology, Okayama University, Japan
[2]Graduate School of Medicine Dentistry and Pharmaceutical Sciences, Okayama University, Japan

h_muraka@a.cs.okayama-u.ac.jp, {abe, hara}@cs.okayama-u.ac.jp,
sato.masaaki@s.okayama-u.ac.jp, minagi@md.okayama-u.ac.jp

Abstract

In this paper, we propose an algorithm to improve the naturalness of the reconstructed glossectomy patient's speech that is generated by voice conversion to enhance the intelligibility of speech uttered by patients with a wide glossectomy. While existing VC algorithms make it possible to improve intelligibility and naturalness, the result is still not satisfying. To solve the continuing problems, we propose to directly modify the speech waveforms using a spectrum differential. The motivation is that glossectomy patients mainly have problems in their vocal tract, not in their vocal cords. The proposed algorithm requires no source parameter extractions for speech synthesis, so there are no errors in source parameter extractions and we are able to make the best use of the original source characteristics. In terms of spectrum conversion, we evaluate with both GMM and DNN. Subjective evaluations show that our algorithm can synthesize more natural speech than the vocoder-based method. Judging from observations of the spectrogram, power in high-frequency bands of fricatives and stops is reconstructed to be similar to that of natural speech.

Index Terms: voice conversion, speech intelligibility, glossectomy, spectral differential, neural network

1. Introduction

Speech is the primary means of communication for human beings and plays a crucial role in maintaining one's quality of life (QoL). Speech is no less important for individuals with speech production problems. Intensive studies have been performed to facilitate improvements in the speech of patients with tongue movement disorders or tongue resection [1, 2, 3]. As a new approach from a speech processing point of view, we proposed to use voice conversion (VC) algorithms to improve speech quality uttered by glossectomy patients.

VC [4] is a technique to modify one speaker's voice to the voice of another speaker while keeping the linguistic information unchanged. A number of VC studies employed statistical approaches for mapping features of a source speaker to those of a target one. Recently, algorithms based on Gaussian Mixture Models (GMM) [5, 6] or neural networks (NNs) [7] have been widely investigated. To improve intelligibility of glossectomy patients' speech, by Tanaka *et al.* [8] we adopted a glossectomy patient as a source speaker and a professional narrator as a target speaker. We made this choice because professional narrators utter speech not only with high intelligibility but also with high consistency. We confirmed that the proposed algorithm worked well for some phonemes under some phoneme contexts, but it did not always work well. The discrepancy was caused mainly by a speaker factor, namely that the source speaker and the target speaker are different. To improve the performance, of our algorithm, by Tanaka *et al.* [9] we proposed a speaker-dependent approach of collecting speech uttered by a patient before and after the glossectomy. Experimental results showed that, in terms of intelligibility of the reconstructed speech, the speaker-dependent approach obviously outperformed speaker-independent ones. However, in the speaker-dependent cases, insufficient naturalness of the reconstructed speech is relatively noticeable. To solve this problem, we propose here to directly modify speech waveforms using a spectrum differential. The motivation for our idea is that glossectomy patients primarily have problems in the vocal tract, not in the vocal cords.

The algorithm discussed in this paper that directly modifies waveforms using spectrum differential was originally proposed to improve the quality of singing voice conversion (SVC) [9]. In the case of SVC, direct modification for the waveform is feasible because musical intervals do not need to be changed. In other words, the algorithm cannot change source parameters as VOCODER can do. An advantage of the algorithm is that it does not require source parameter extractions for speech synthesis and is thus free from errors in source parameter extractions. In the case of the application discussed in this paper, we expect three advantages to improve naturalness of the reconstructed speech. (1) Because glossectomy patients do not have problems in the vocal cords, glottal waveforms are properly generated. The direct modification for the waveform allows the use of characteristics contained in original glottal waveforms. (2) In the cases of a wide glossectomy, it is difficult for glossectomy patients to generate turbulent airflow by constructing a constriction in the vocal tract, and they therefore have weak excitation power but preserve appropriate power distribution in frequency. The direct modification for the waveform enables the use of the preserved power distribution in frequency. (3) From a viewpoint of computational cost, the direct modification for the waveform is inexpensive and can run in real time. This is an important feature to provide applications for glossectomy patients.

The remainder of the paper is organized as follows. In Section 2 we describe the speech data that is to be used for evaluations. In Section 3 we explain the algorithm that directly modifies the waveform using a spectrum differential [10] and how it is applied to our task. In Section 4 provide the results of our evaluation and a discussion. Finally, in Section 5, we present our conclusions and suggest avenues for future research.

2. Speech data for the evaluation

In order to simulate glossectomy patients, we fabricated an intra-oral appliance that covers the lower dental arch and tongue surface to restrain tongue movements during speech [9]. The appliance is made of a pressure-thermoforming resin plate. Normal speakers uttered speech with and without the appliance to simulate speech before and after a glossectomy. In order to ensure proper fit, an appliance was developed for each individual speaker using a plaster model of individual's teeth. Because the appliance permits the tongue to move only above a certain level,

10.21437/Interspeech.2018-1239

Table 1: *Recorded speech uttered by normal speakers*

session	sentences	speaking style	appliance
1	100 sentence	phrase-by-phrase	with
2	100 sentence	phrase-by-phrase	without
3	53 sentence	sentence-by-sentence	with
4	53 sentence	sentence-by-sentence	without

normal speakers with the appliance cannot correctly pronounce some phonemes, such as /t/, /d/, /k/ and /g/. The result is speech that imitates that of glossectomy patients to some extent.

Table 1 shows the recorded speech uttered by four normal speakers (three males and a female) with and without the appliance. Phrase-by-phrase utterances in Sessions 1 and 2 were read by speakers with pauses between phrases, whereas sentence-by-sentence utterances in the Sessions 3 and 4 were read without pauses between phrases.

The purpose of the phrase-by-phrase utterances was to generate mapping functions for VC in part to the burden on patients. The shorter the text, the less likely it is that patients will mispronounce portions of the text. If the text is longer, it is more likely that patients would need to repeat sentences several times in order to record correct utterances, which can place a large burden on patients. Furthermore, short utterances can reduce the chances of failures to find correspondences by dynamic time warping (DTW).

The sentence-by-sentence utterances are used for evaluations because in terms of the number of pauses, sentence-by-sentence utterances are more similar to the speech of everyday life than phrase-by-phrase utterances.

3. Voice conversion with spectral differential modification

3.1. Training process

Let $\boldsymbol{x}_t$ and $\Delta\boldsymbol{x}_t$ denote static and dynamic acoustic features of the source speaker, respectively, and let $\boldsymbol{y}_t$ and $\Delta\boldsymbol{y}_t$ be those of the target speaker. Joined static and dynamic feature vectors are defined as $\boldsymbol{X}_t = [\boldsymbol{x}_t^{\mathrm{T}}, \Delta\boldsymbol{x}_t^{\mathrm{T}}]^{\mathrm{T}}$ and $\boldsymbol{Y}_t = [\boldsymbol{y}_t^{\mathrm{T}}, \Delta\boldsymbol{y}_t^{\mathrm{T}}]^{\mathrm{T}}$, where T denotes the transposition of a vector and t denotes time. In the training process we model a function that converts $\boldsymbol{X}_t$ to $\boldsymbol{D}_t$, where $\boldsymbol{D}_t = [\boldsymbol{d}_t^{\mathrm{T}}, \Delta\boldsymbol{d}_t^{\mathrm{T}}]^{\mathrm{T}}$ and $\boldsymbol{d}_t = \boldsymbol{y}_t - \boldsymbol{x}_t$. In other words, $\boldsymbol{D}_t$ is the matrix of differential spectral features between the source and target speaker. The results reported in this paper include both DNN (Deep NN) and GMM, using a parallel data set.

3.1.1. GMM-based modeling

The joint probability density of the feature vectors is modeled by a GMM as follows [10]:

$$P(\boldsymbol{z}_t \mid \boldsymbol{\lambda}^{(\boldsymbol{z})}) = \sum_{m=1}^{M} w_m N\left(\boldsymbol{z}_t; \boldsymbol{\mu}_m^{(\boldsymbol{z})}, \boldsymbol{\Sigma}_m^{(\boldsymbol{z})}\right), \quad (1)$$

where $\boldsymbol{z}_t$ is the joint vector $[\boldsymbol{X}_t^{\mathrm{T}}, \boldsymbol{D}_t^{\mathrm{T}}]^{\mathrm{T}}$, m is the mixture component index, M is the total number of mixture components, and w_m is the weight of the m^{th} mixture component. Further, the normal distribution with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is denoted as $N(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$. A parameter set of the GMM is $\boldsymbol{\lambda}^{(\boldsymbol{z})}$, which consists of weights, mean vectors, and the covariance matrices for individual mixture components. Joint vectors $\boldsymbol{z}_t$ $(t = 1, 2, \cdots, N)$ are generated by DTW using a parallel speech corpus in which source and target speakers utter the same sentences. Finally, N is the total frame number of training data for the given speech corpus.

The mean vector $\boldsymbol{\mu}_m^{(\boldsymbol{z})}$ and the covariance matrix $\boldsymbol{\Sigma}_m^{(\boldsymbol{z})}$ of the m^{th} mixture component are written as

$$\boldsymbol{\mu}_m^{(\boldsymbol{z})} = \begin{bmatrix} \boldsymbol{\mu}_m^{(\boldsymbol{x})} \\ \boldsymbol{\mu}_m^{(\boldsymbol{d})} \end{bmatrix}, \quad \boldsymbol{\Sigma}_m^{(\boldsymbol{z})} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(\boldsymbol{xx})} & \boldsymbol{\Sigma}_m^{(\boldsymbol{xd})} \\ \boldsymbol{\Sigma}_m^{(\boldsymbol{dx})} & \boldsymbol{\Sigma}_m^{(\boldsymbol{dd})} \end{bmatrix}, \quad (2)$$

where $\boldsymbol{\mu}_m^{(\boldsymbol{x})}$ and $\boldsymbol{\mu}_m^{(\boldsymbol{d})}$ are the mean vectors of the m^{th} mixture component for the source and differential spectral features, respectively. Matrices $\boldsymbol{\Sigma}_m^{(\boldsymbol{xx})}$ and $\boldsymbol{\Sigma}_m^{(\boldsymbol{dd})}$ are the covariance matrices of the m^{th} mixture component for the source and differential spectral features, respectively. Matrices $\boldsymbol{\Sigma}_m^{(\boldsymbol{xd})}$ and $\boldsymbol{\Sigma}_m^{(\boldsymbol{dx})}$ are the cross-covariance matrices of the m^{th} mixture component for the source and differential spectral features, respectively. The GMM is trained with an expectation-maximization (EM) algorithm using the joint vectors.

The conditional probability density of $\boldsymbol{d}_t$, given $\boldsymbol{x}_t$, is also represented as a GMM as

$$\begin{aligned} &P\left(\boldsymbol{d}_t \mid \boldsymbol{x}_t, \boldsymbol{\lambda}^{(\boldsymbol{z})}\right) \\ &= \sum_{m=1}^{M} P\left(m \mid \boldsymbol{x}_t, \boldsymbol{\lambda}^{(\boldsymbol{z})}\right) P\left(\boldsymbol{d}_t \mid \boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(\boldsymbol{z})}\right), \end{aligned} \quad (3)$$

where

$$P\left(m \mid \boldsymbol{x}_t, \boldsymbol{\lambda}^{(\boldsymbol{z})}\right) = \frac{w_m N\left(\boldsymbol{x}_t; \boldsymbol{\mu}_m^{(\boldsymbol{x})}, \boldsymbol{\Sigma}_m^{(\boldsymbol{xx})}\right)}{\sum_{n=1}^{M} w_n N\left(\boldsymbol{x}_t; \boldsymbol{\mu}_n^{(\boldsymbol{x})}, \boldsymbol{\Sigma}_n^{(\boldsymbol{xx})}\right)} \quad (4)$$

and

$$P\left(\boldsymbol{d}_t \mid \boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(\boldsymbol{z})}\right) = w_m N\left(\boldsymbol{d}_t; \boldsymbol{E}_{m,t}^{(\boldsymbol{d})}, \boldsymbol{D}_m^{(\boldsymbol{d})}\right). \quad (5)$$

Mean vector $\boldsymbol{E}_{m,t}^{(\boldsymbol{d})}$ and covariance matrix $\boldsymbol{D}_m^{(\boldsymbol{d})}$ of the m^{th} conditional probability distribution are written as

$$\boldsymbol{E}_{m,t}^{(\boldsymbol{d})} = \boldsymbol{\mu}_m^{(\boldsymbol{d})} + \boldsymbol{\Sigma}_m^{(\boldsymbol{dx})} \boldsymbol{\Sigma}_m^{(\boldsymbol{xx})^{-1}} \left(\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(\boldsymbol{x})}\right) \quad (6)$$

and

$$\boldsymbol{D}_m^{(\boldsymbol{d})} = \boldsymbol{\Sigma}_m^{(\boldsymbol{dd})} - \boldsymbol{\Sigma}_m^{(\boldsymbol{dx})} \boldsymbol{\Sigma}_m^{(\boldsymbol{xx})^{-1}} \boldsymbol{\Sigma}_m^{(\boldsymbol{xd})}. \quad (7)$$

Using the conventional method described by Stylianou *et al.* [5] and Kain and Macon [6], the conversion is performed based on the minimum mean-square error (MMSE). Finally, we can obtain the estimated spectral differential features $\hat{\boldsymbol{d}}_t$ as following:

$$\hat{\boldsymbol{d}}_t = \sum_{m=1}^{M} P\left(m \mid \boldsymbol{x}_t, \boldsymbol{\lambda}^{(\boldsymbol{z})}\right) \boldsymbol{E}_{m,t}^{(\boldsymbol{d})}. \quad (8)$$

3.1.2. DNN-based modeling

For the case of training the M-layer NN model ($m = 1, \cdots, M$), let $h^{(m)}$ be the value at the m^{th} middle layer of network. The forward propagation processing is performed as follows for the input feature vector $\boldsymbol{X}_t$ in every layer as described by the following equations:

$$\boldsymbol{h}_t^{(1)} = f\left(\boldsymbol{W}^{(1)} \boldsymbol{X}_t + \boldsymbol{b}^{(1)}\right) \quad (9)$$

$$\boldsymbol{h}_t^{(m+1)} = f\left(\boldsymbol{W}^{(m+1)} \boldsymbol{h}_t^{(m)} + \boldsymbol{b}^{(m+1)}\right) \quad (10)$$

$$\hat{\boldsymbol{d}}_t = f\left(\boldsymbol{W}^{(M)} \boldsymbol{h}_t^{(M-1)} + \boldsymbol{b}^{(M)}\right) \quad (11)$$

where $\boldsymbol{W}^{(m)}$ and $\boldsymbol{b}^{(m)}$ are the weight matrix and bias vector at the m^{th} layer, respectively. Eq. (9), Eq. (10) and Eq. (11) show, respectively, the input layer, the middle layer ($m < M$), and the output layer. The function $f(\cdot)$ is the activation function. The network is trained by a backpropagation algorithm to minimize mean squared error calculated between $\hat{\boldsymbol{d}}_t$ in Eq. (11) and $\boldsymbol{d}_t$ obtained from the training data set.

3.2. Conversion process

In the conversion process $\hat{\boldsymbol{d}}_t$ is estimated from the source speaker's features $\boldsymbol{X}_t$ by the trained GMM or DNN. The speech of the source speaker is converted to the target speaker's by directly filtering the speech waveform with $\hat{\boldsymbol{d}}_t$.

Here, we explain how the differential spectral method works in the conversion process. According to speech production theory, we can obtain the residual signal $R(z)$ as follows:

$$R(z) = \frac{1}{\boldsymbol{X}(z)} S(z) \tag{12}$$

where $S(z)$ is the input speech and $\boldsymbol{X}(z)$ is a transfer function that characterizes a vocal tract of the source speaker in the Z domain. Let $\hat{\boldsymbol{Y}}(z)$ be the converted transfer function that characterizes a vocal tract of the target speaker. Then we obtain the converted speech $T(z)$ by filtering the residual signal as follows:

$$\begin{aligned} T(z) &= \hat{\boldsymbol{Y}}(z) R(z) \\ &= \frac{\hat{\boldsymbol{Y}}(z)}{\boldsymbol{X}(z)} S(z) \\ &= \frac{\exp\left(\sum_{m=0}^{M} \hat{c}_m^{(y)} \tilde{z}^{-m}\right)}{\exp\left(\sum_{m=0}^{M} c_m^{(x)} \tilde{z}^{-m}\right)} S(z) \\ &= \exp\left(\sum_{m=0}^{M} \left(\hat{c}_m^{(y)} - c_m^{(x)}\right) \tilde{z}^{-m}\right) S(z) \end{aligned} \tag{13}$$

where $c_m^{(x)}$ and $\hat{c}_m^{(y)}$ are m^{th} mel-cepstrum coefficients of the input speech and target speech, respectively. Here $\tilde{z}^{-1}$ represents a first-order all-pass filter. Therefore, the filtering based on the estimated mel-cepstrum for the residual signal is equal to the direct filtering of the input speech using differential mel-cepstrum between mel-cepstrum of input speech and estimated mel-cepstrum of target speech. Here, mel-cepstrum coefficients are used to characterize the transfer function. If the input and target voices are from the same speaker, the assumption that the residual signal of the input speech is equal to that of the target speech is holds.

4. Experiments

4.1. Experimental conditions

As mentioned in Section 2, for training and evaluations we used, respectively, 100 sentences uttered phrase-by-phrase and 53 sentences uttered by sentence-by-sentence were used. In the DNN training ten sentences were used for validation. The sampling frequency was 20 kHz. There were two male speakers (M1 and M2) and a female (F1). To fix notation, in the remainder of the paper speech uttered by M1 with the appliance is denoted SPM1 (Simulated Patient Male1), and similarly for the others.

Spectral envelopes were extracted by WORLD [11] and parameterized to the 0-25th mel-cepstral coefficients and their dynamic features. The frame shift was 5 ms. In the spectral differential method, the mel log spectrum approximation (MLSA) filter [12] was used as the synthesis filter. In the vocoder-based method, speech was synthesized by the WORLD vocoder.

We denote the combinations of the conversion models and methods as follows:

- **GMM-D**: GMM-based VC using the spectral differential method;
- **GMM-V**: GMM-based VC using the vocoder-based method;
- **DNN-D**: DNN-based VC using the spectral differential method;
- **DNN-V**: DNN-based VC using the vocoder-based method.

Thirty-two mixture components were used in the GMM method, and the full covariance matrices were employed. The GMM is trained based on trajectory-concerned training, and modification of global variance (GV) [13] is applied only for the subjective evaluation but not for objective evaluation. [1]

In the case of DNN we adopted multilayer perceptron (MLP) as the conversion model. In each layer, the number of units is set as [52, 1024, 1024, 1024, 1024, and 52]. The structure of the DNN showed the best performance in our preliminary experiments, which has the greatest quantity of both layer-levels and units-numbers. The rectified linear units (ReLU) 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. For training the DNN, a mini-batch adaptive moment estimation (Adam)-based backpropagation algorithm was used.

4.2. Objective evaluation

Mel-cepstral distortion (MelCD) is used as an objective measure of the spectral distance between converted speech and target speech. For fair comparison in both spectral differential method and vocoder-based method, the mel-cepstrum parameters were calculated from post-converted speech; this is because we cannot obtain mel-cepstrum parameters without synthesizing speech in the case of spectral differential method.

Figure 1 shows the results of objective evaluation for the three speakers. Comparing pre-conversion speech with post-conversion speech, MelCD score is reduced by 18%, 26%, and 19% for SPM1 to M1, SPM2 to M2, SPF1 to F1 in DNN-D, respectively. The spectral differential method consistently showed slightly better performance than the vocoder-based method. But in the case of DNN for M2, the spectral differential method demonstrated significantly better performance than the vocoder-based method. This indicates direct waveform modification using spectrum differential can successfully convert spectral characteristics and perform well even in the case of speech of glossectomy patients. Comparing DNN with GMM, there is no significant difference in the spectral differential method, but GMM consistently performed better than DNN for the vocoder-based method.

4.3. Subjective evaluation

We carried out a Mean Opinion Score (MOS) test to evaluate naturalness of speech. In the MOS test, the opinion score was set to a 5-point scale (5:excellent; 4:good; 3:fair; 2:poor; and 1:bad). There are six types of speech: the original simulated patient speech ORIG, and the five forms of converted speech: DNN-D, DNN-V, DNN-VGV, GMM-D, and GMM-VGV, where VGV indicates the vocoder-based method with GV. The subjects were 13 Japanese speakers who evaluated 10

[1] In a preliminary examination, the GV modification had improved the subjective score, but worsen the Mel-cepstral distortion.

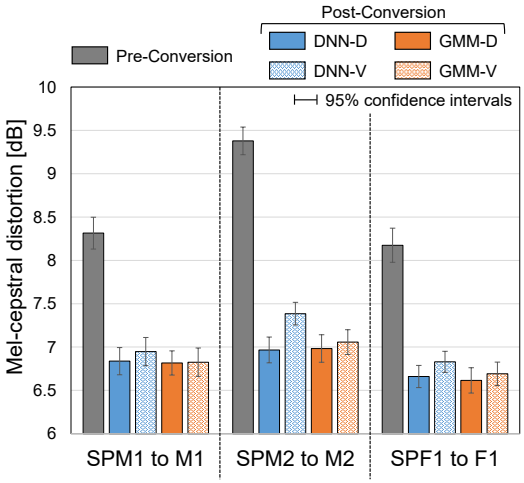

Figure 1: *Mel-cepstral distortion before and after VC for simulated patients (speaker-dependent).*

Table 2: *Mean opinion score (MOS) test result for speech naturalness.*

Conversion	ORIG	DNN			GMM	
		D	V	VGV	D	VGV
SPM1 to M1	4.94	**3.95**	2.09	2.58	**4.13**	2.85
SPM2 to M2	4.91	**3.14**	1.56	2.26	**3.97**	2.85
SPF1 to F1	4.88	**2.95**	1.80	2.39	**3.45**	2.50

sentences for each of the six types of speech. Note that the subjects evaluated only the naturalness of the speech, namely whether the sound was close to that of a real voice. Moreover, the subjects were asked to neglect the intelligibility of the speech.

Table 2 shows the result of the MOS test, in particular, that the spectral differential methods (DNN-D and GMM-D) achieved higher MOS scores than the vocoder-based methods with and without GV (DNN-V, DNN-VGV and GMM-VGV). We attribute the success of the fact that the spectral differential method has no F0 extraction error and has smooth energy changes in consecutive phonemes, which results in the synthesis of more natural speech than vocoder-based method.

4.4. Comparison of the spectrograms

The causes of improving the phoneme intelligibility can be observed in spectrograms. Figure 2 shows spectrograms that compare the VC from SPM1 to M1 by DNN-D and DNN-V. (a) 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), spectral differential method (c) reconstructed the fricative more clearly than vocoder-based method (d). Because the spectral differential method uses the input speech directly in conversion, it can pick up a slight excitation signal of the input speech that is missed by the vocoder-based method. In essence, the spectral differential method works as a filter to emphasize high-frequency components of fricatives.

Our algorithm also shows advantages in the stop, however, there was a negative side-effect on the preceding closure. The high-frequency components of stop /t/, that are shown within the green dotted lines in the Figure 2, which illustrates that these components are reconstructed by the spectral differential method (c) with larger power than the vocoder-based method

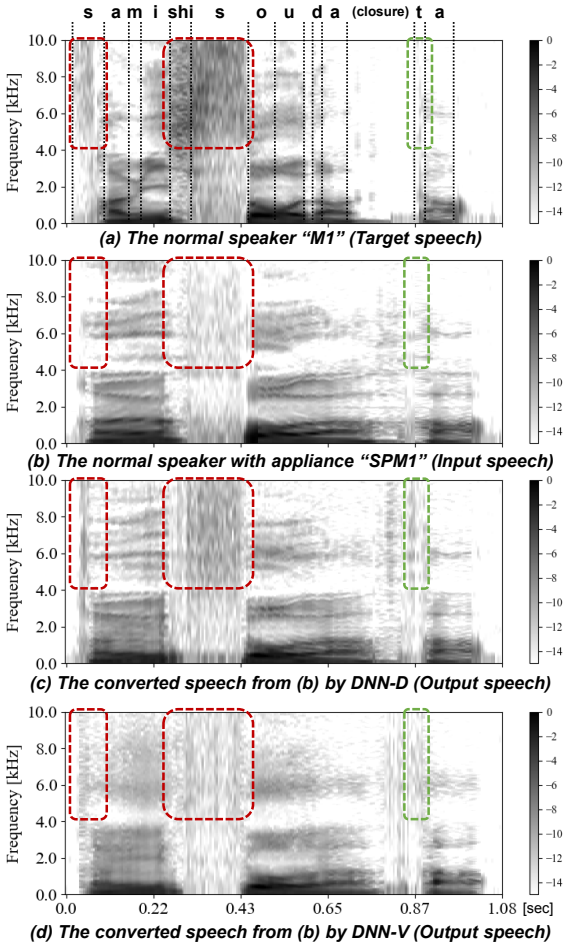

Figure 2: *Comparisons of spectrograms in the case of SPM1 to M1.*

(d). However, we can also observe power in the high-frequency band in surrounding closure. This implies that similar functions are generated for both stops and closure. An advanced conversion algorithm is therefore needed for generating appropriate mapping functions. Promising future works are, for example, to extend the feature vectors so that they contain additional information and to investigate the DNN model structures that can treat contexts of phonemes.

5. Conclusions

We proposed an algorithm to improve naturalness of reconstructed glossectomy patients' speech. The basic idea of our method is to directly modify waveforms using a spectrum differential to avoid the artificiality of speech caused by the vocoder-based method. To evaluate the effectiveness of our method, we carried out subjective experiments that compared the proposed algorithm with the vocoder-based method. The experimental results showed the effectiveness of our method in MOS score. Moreover, from observations of the spectrogram, power in high-frequency band fricatives and stops are reconstructed as similar to those of natural speech.

As part of our future work, we have plans to apply other DNN structures to improve the performance of conversions. For example, RNN-based modeling and CTC-based modeling are promising. Furthermore, we will consider applying a post-filtering approach to enhance the power of the high-frequency band of the converted speech.

6. References

[1] R. Cantor, T. Curtis, T. Shipp, J. Beume, and B. Vogel, "Maxillary speech prostheses for mandibular surgical defects," *J. Prosthet Dent*, 22:253–60. (1969)

[2] R. Leonard, and R. Gillis, "Differential effects of speech prostheses in glossectomized patients," *J. Prosthet Dent*, 64:701–8. (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] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *Proc. ICASSP*, S14.1, pp.655–658. (1988)

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

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

[7] 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)

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

[9] 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)

[10] 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)

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

[12] 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)

[13] 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. (2007)