# Enhancing a Glossectomy Patient's Speech via GMM-based Voice Conversion

Kei Tanaka*, Sunao Hara*, Masanobu Abe*, and Shogo Minagi†
*Graduate School of Natural Science and Technology, Okayama University, Japan
E-mail: pjot7wfu@s.okayama-u.ac.jp, {abe, hara}@cs.okayama-u.ac.jp
†Graduate School of Medicine Dentistry and Pharmaceutical Sciences, Okayama University, Japan
E-mail: minagi@md.okayama-u.ac.jp

***Abstract*— In this paper, we describe the use of a voice conversion algorithm for improving the intelligibility of speech by patients with articulation disorders caused by a wide glossectomy and/or segmental mandibulectomy. As a first trial, to demonstrate the difficulty of the task at hand, we implemented a conventional Gaussian mixture model (GMM)-based algorithm using a frame-by-frame approach. We compared voice conversion performance among normal speakers and one with an articulation disorder by measuring the number of training sentences, the number of GMM mixtures, and the variety of speaking styles of training speech. According to our experiment results, the mel-cepstrum (MC) distance was decreased by 40% in all pairs of speakers as compared with that of pre-conversion measures; however, at post-conversion, the MC distance between a pair of a glossectomy speaker and a normal speaker was 28% larger than that between pairs of normal speakers. The analysis of resulting spectrograms showed that the voice conversion algorithm successfully reconstructed high-frequency spectra in phonemes /h/, /t/, /k/, /ts/, and /ch/; we also confirmed improvements of speech intelligibility via informal listening tests.**

## I. INTRODUCTION

Speech is the primary means of communication for human beings and plays a crucial role in maintaining one's quality of life (QoL) 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. The palatal augmentation prosthesis (PAP) is one such promising method, and its efficacy has been widely recognized [1,2]. However, use of a PAP is insufficient, especially for patients with severe articulation disorders associated with a wide glossectomy and/or segmental mandibulectomy. To help these individuals, we recently proposed a new method using a kinematic artificial tongue (KAT) together with a PAP and reported good overall performance of our proposed method [3]. Unfortunately, prosthesis approaches such as PAP and KAT (or combinations of such approaches) have the drawback of requiring patients to use a wearable device in their mouth. One complication here is that patients cannot wear them during meals. In this paper, we therefore propose another approach to improve speech quality by using digital signal processing, particularly a voice conversion algorithm. Our approach does not require that an individual wear any special equipment, which results in more natural opportunities for speech communication.

Voice conversion is a technique that changes a speaker's individuality, i.e., speech uttered by speaker A is changed to sound as if another speaker B had uttered it [4]. Hereinafter, speakers A and B are referred to as source and target speakers, respectively. Voice conversion has been applied to a variety of applications [5,6,7], including a clinical application of speech enhancement to esophageal speech [8]. In this latter work, the source speaker is an esophageal speaker, while the target speaker is a normal speaker. In this study, we propose another clinical application of voice conversion to enhance speech uttered by individuals with articulation disorders (whereas esophageal speech of individuals is referred to as voice source disorders). We start with the following two fundamental questions: (1) can we reconstruct speech from degraded speech caused by one or more articulation disorders; and (2) how can we handle large variations in terms of disorder levels among patients, which are based on varying types and number of surgeries. First, we demonstrate how difficult the specific voice conversion task is by comparing the performance of vocally impaired and normal speakers. For our comparisons, we applied a conventional voice conversion algorithm.

The rest of the paper is organized as follows. In Section 2, we describe a patient whose speech we focus on reconstructing via voice conversion. In Section 3, we explain our voice conversion algorithm, which is based on the Gaussian mixture model (GMM). In Section 4, we show our evaluation results and provide a discussion. Finally, in Section 5, we present our conclusions and suggest avenues for future work.

## II. PATIENT AND SPEECH MATERIAL

### A. Patient

In April 2014, a 50-year-old man was diagnosed with tongue cancer via a CT scan and biopsy and was subsequently treated using combination chemotherapy the following month. After this treatment, surgical intervention occurred in June 2014, the surgery involving subtotal glossectomy, right cervical dissection, right cricopharyngeus muscle amputation, and laryngeal elevation. Given a recurrence of the cancer in August 2014, oropharyngeal carcinoma removal surgery, segmental mandibulectomy, mesopharyngeal tumor resection, mandibular bone debridement, and reconstruction with anterolateral thigh flap were undertaken. Nonetheless, the

cancer recurred in October 2014, leading to a right mandibulectomy, left cervical dissection, reconstruction with free flap of the jawbone rolled letter paper evisceration, and left neck dissection with reconstruction by the right-front outside thigh free flap. Given these extensive surgeries, the patient's speech was unfortunately quite unintelligible.

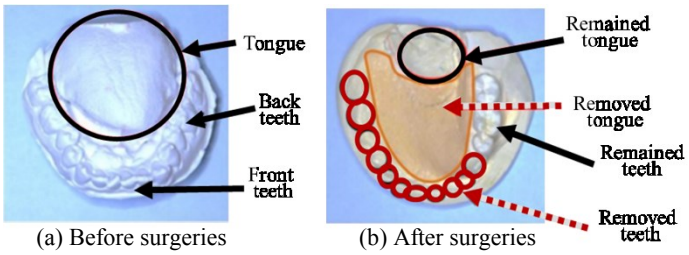
Figures 1 shows an intraoral mock-up following the three operations. The patient was referred to the Department of Oral Rehabilitation and Occlusion, Okayama University for treatment with a PAP. We applied a palatal plate (PP) to the patient's maxilla and a KAT to his mandibular to improve his articulation abilities.

### *B. Speech material*

After the surgeries noted above and the application of a PAP and KAT, the patient was asked to speak in three sessions, summarized in Table 1. In these sessions, we considered the aspects described below.

(1) Speaking styles

As described in the next section, the voice conversion algorithm first determines corresponding feature vectors between two speakers using a parallel corpus in which two speakers utter the same text. These correspondences are determined via dynamic time warping (DTW) to take varying speeds into account. Because differences between the speech of a glossectomy speaker and a normal speaker are typically much larger than between two normal speakers, DTW might not work properly in some cases. To reduce the incidence of this problem, we propose using shorter lengths of speech, i.e., phrase-by-phrase utterances instead of sentence-by-sentence utterances. Moreover, the shorter the text, the less likely patients will mispronounce portions of the text. Repeating sentences several times can place a large burden on patients. Nonetheless, one drawback of using trained mapping functions using phrase-by-phrase utterances instead of sentence-by-sentence utterances is that speech used in daily life is more likely to consist of sentence-by-sentence utterances. As a result, trained mapping functions are degraded not only by differences in individual speakers but also by differences in styles.




(a) Before surgeries (b) After surgeries

Fig. 1 Intraoral mock-up of a patient.

Table 1 Speech material

| Session | Number of sentence | Speaking style | Devices (PAP and KAT) |
|---|---|---|---|
| 1 | Full text (503 sentence) | Phrase-by-phrase utterances | without |
| 2 | Subset texts (103 sentence) | Sentence-by-sentence utterances | without |
| 3 | Subset texts (103 sentence) | Sentence-by-sentence utterances | with |

(2) Number of sentences

In general, the more data we have, the better the performance; however, from the perspective of the burden on the patient, the amount of data collected should be as small as possible. To ensure the best performance, as a reference, we recorded 503 sentences, i.e., we use this as the maximum number of sentences for patients to utter. For the recordings, it took approximately nine hours, i.e., three hours per day for three days. We currently reduced the target number of sentences to 100, but this is not small enough in terms of the burden on the patient and is relatively large compared with the target number of sentences required for the voice conversion between normal speakers. This is primarily because we assume that voice conversion between a glossectomy speaker and a normal speaker is more difficult than between normal speakers.

## III. GMM-BASED VOICE CONVERSION ALGORITHM

### *A. Probability density function*

Let $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ be D-dimensional source and target feature vectors at frame $t$, respectively. The joint probability density of the source and target feature vectors is modeled by a GMM as follows:

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

where $\boldsymbol{z}_t$ is joint vector $[\boldsymbol{x}_t^{\mathrm{T}}, \boldsymbol{y}_t^{\mathrm{T}}]^{\mathrm{T}}$, T denotes the transposition of a vector, $m$ is the mixture component index, $M$ is the total number of mixture components, and $w_m$ is the weight of the $m^{\text{th}}$ mixture component. Further, the normal distribution with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is denoted as $N\left(\cdot;\ \boldsymbol{\mu},\ \boldsymbol{\Sigma}\right)$. 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 $\boldsymbol{z}_t$ $(t = 1,2, \ldots N)$ 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.

Mean vector $\boldsymbol{\mu}_m^{(\boldsymbol{z})}$ and covariance matrix $\boldsymbol{\Sigma}_m^{(\boldsymbol{z})}$ of the mth mixture component are written as

$$\boldsymbol{\mu}_m^{(\boldsymbol{z})} = \begin{bmatrix} \boldsymbol{\mu}_m^{(\boldsymbol{x})} \\ \boldsymbol{\mu}_m^{(\boldsymbol{y})} \end{bmatrix}, \quad \boldsymbol{\Sigma}_m^{(\boldsymbol{z})} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(\boldsymbol{xx})} & \boldsymbol{\Sigma}_m^{(\boldsymbol{xy})} \\ \boldsymbol{\Sigma}_m^{(\boldsymbol{yx})} & \boldsymbol{\Sigma}_m^{(\boldsymbol{yy})} \end{bmatrix},$$

where $\boldsymbol{\mu}_m^{(\boldsymbol{x})}$ and $\boldsymbol{\mu}_m^{(\boldsymbol{y})}$ are the mean vectors of the $m^{\text{th}}$ mixture component for the source and target, respectively. Matrices $\boldsymbol{\Sigma}_m^{(\boldsymbol{xx})}$ and $\boldsymbol{\Sigma}_m^{(\boldsymbol{yy})}$ are the covariance matrices of the $m^{\text{th}}$ mixture component for the source and target, respectively. Matrices $\boldsymbol{\Sigma}_m^{(\boldsymbol{xy})}$ and $\boldsymbol{\Sigma}_m^{(\boldsymbol{yx})}$ are the cross-covariance matrices of the $m^{\text{th}}$ mixture component for the source and target, respectively. The GMM is trained with an expectation-maximization (EM) algorithm using the joint vectors, which are automatically aligned by DTW, in a training set.

### *B. Mapping function*

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

$$P(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}) = \sum_{m=1}^{M} P(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}) P(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(z)}),$$

where

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

and

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

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

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

and

$$\boldsymbol{D}_m^{(y)} = \boldsymbol{\Sigma}_m^{(yy)} - \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)^{-1}} \boldsymbol{\Sigma}_m^{(xy)}.$$

Using the conventional method described in [9] and [10], the conversion is performed based on the minimum mean-square error (MMSE) as follows:

$$\begin{aligned}\hat{\boldsymbol{y}}_t &= E[\boldsymbol{y}_t|\boldsymbol{x}_t] \\ &= \int P(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}) \boldsymbol{y}_t \, d\boldsymbol{y}_t \\ &= \int \sum_{m=1}^{M} P(m|\boldsymbol{x}_t,\ \boldsymbol{\lambda}^{(z)})\, P(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(z)}) \boldsymbol{y}_t \, d\boldsymbol{y}_t \\ &= \sum_{m=1}^{M} P(m|\boldsymbol{x}_t,\ \boldsymbol{\lambda}^{(z)}) \boldsymbol{E}_{m,t}^{(y)}\end{aligned}$$

Here, $E[\cdot]$ represents the expectation and $\hat{\boldsymbol{y}}_t$ is the converted target feature vector.

## IV. Experiments for Glossectomy Speech Enhancement

Table 2 presents common parameters for the experiments described in this section.

### A. Experiment 1

We designed this first experiment to ensure we identified the appropriate number of mixtures and the influence of training data size. Voice conversion was performed by a male glossectomy speaker and a normal male speaker, then between two normal male speakers. As described in Section II. B, the mapping function was trained using phrase-by-phrase speech and evaluated using sentence-by-sentence speech. To train via 100 sentences, we employed tenfold cross-validation.

Our experimental results are shown in Fig. 2. By comparing differences between pre-conversion values, MC distances decreased by 40%. However, at post-conversion, MC distances between a pair of a glossectomy and a normal speaker are 28% larger than those between normal speakers. In terms of training data size, 450 sentences achieved slightly better performance than 100 sentences. Judging from our results, we conclude that voice conversion is fruitful for the glossectomy speaker even when trained with only 100 sentences. Furthermore, the number of mixtures that yield the best performance is 16 for a pair of glossectomy and normal speakers and is 32 for a pair of normal speakers.

### B. Experiment 2

We designed this experiment to intuitively show the level of difficulty of voice conversion between a glossectomy speaker and a normal speaker and illustrate the influence caused by speaking style differences. For comparison, in addition to the pairs used in Experiment 1, voice conversions were performed between a male speaker and a female speaker. The number of mixtures was set to 16 and 32, according to the results of Experiment 1. In comparison of speaking styles (SS), we use the model learned by phrase-by-phrase utterances (Diff. SS) and by Sentence-by-sentence utterances (Same SS).

Figure 3 shows our experimental results. Judging from MC distances at pre-conversion, differences between a glossectomy speaker and a normal speaker are much larger than those between normal speakers. This phenomenon indicates that the vocal tract shape of the glossectomy speaker is substantially different from that of a normal speaker. In terms of the voice conversion performance, MC distances decreased by 40% in all pairs. However, at post-conversion, larger MC distances remained in a pair of a glossectomy and a normal speaker versus that of pairs of normal speakers. One reason here could be that there are one-to-many or many-to-one correspondences in frame-based feature vectors between glossectomy and normal speakers. Performance differences caused by speaking styles in the training data are relatively smaller as compared with those caused by speaker individualities. Judging from our results, we conclude that training with phrase-by-phrase

Table 2 Experimental conditions

| | |
|---|---|
| Sampling frequency | 20 kHz |
| Speech analysis | STRAIGHT[11] |
| Frame shift | 5 ms |
| Speech feature | $0^{\text{th}}$~$24^{\text{th}}$ mel-cepstral Coefficients and their Δ |

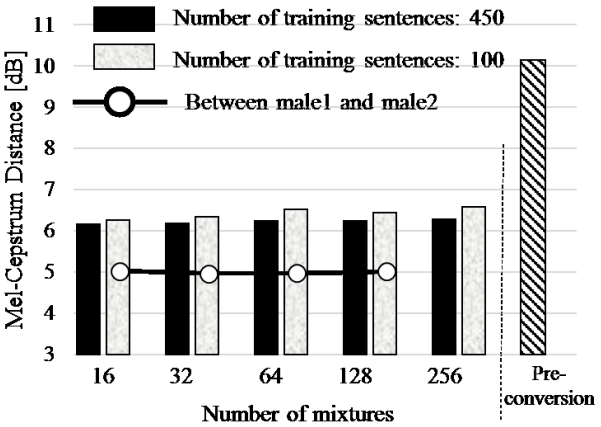



Fig. 2 MC distances per number of GMM mixtures (Experiment 1).

speech is fruitful for glossectomy speakers even if the trained function is applied to sentence-by-sentence speech.

Figure 4 shows spectrograms of glossectomy speech (i.e., input speech), converted speech (i.e., output speech), and normal speech (i.e., target speech). By comparing converted speech with glossectomy speech, in the frequency bands marked with an "x" in Fig. 4(b), the power spectrum is newly generated and located at similar regions to that of normal speech, indicating that the spectrum is properly reconstructed. Moreover, the marked bands are related to phonemes /h/, /t/, /k/, /ts/, and /ch/, which is reasonable here because these phonemes are articulated using the tongue. Another feature of converted speech is that the power spectrum is reconstructed around 1 kHz bands marked with a "y" in Fig. 4(a). Through informal listening tests, we confirmed improvements in speech intelligibility for these phonemes.

## V. CONCLUSIONS

To improve the intelligibility of speech by a patient who has one or more articulation disorders due to a wide glossectomy and/or segmental mandibulectomy, we applied a voice conversion algorithm based on GMM using a frame-by-frame approach. According to our experimental results, after conversion, MC distances decreased by 40% in all pairs of speakers; however, MC distances between a pair of glossectomy and normal speakers was 28% larger than those between normal speakers. Furthermore, for phonemes /h/, /t/, /k/, /ts/, and /ch/, the high-frequency spectrum was successfully reconstructed. However, we observed some problems of intelligibility in the converted speech. As part of our future work, we plan to perform formal listening tests to ensure the intelligibility of phonemes is properly reconstructed. In terms of problems of one-to-many or many-to-one mappings, we will try to implove by using a segmental approach instead of frame-by-frame approach.


## ACKNOWLEDGMENT

We express our genuine thanks to Dr. Ken-ichi Kozaki, a professor at the Graduate School of Medicine Dentistry and Pharmaceutical Sciences, Okayama University, Japan. Regardless of his difficult condition after surgical intervention, he willingly cooperated with us to record his speech. He unfortunately passed away on May 29, 2016. We sincerely pray for the repose of his soul.

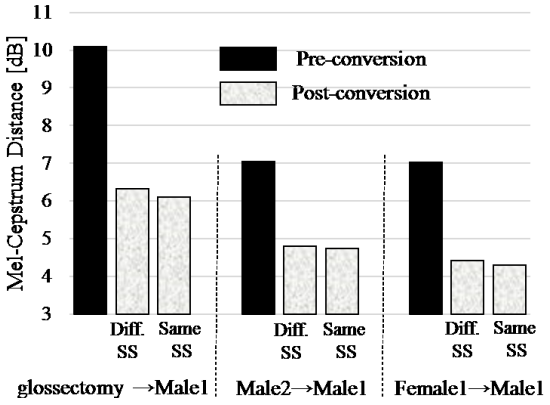




Fig.3 MC distances for different speaker pairs

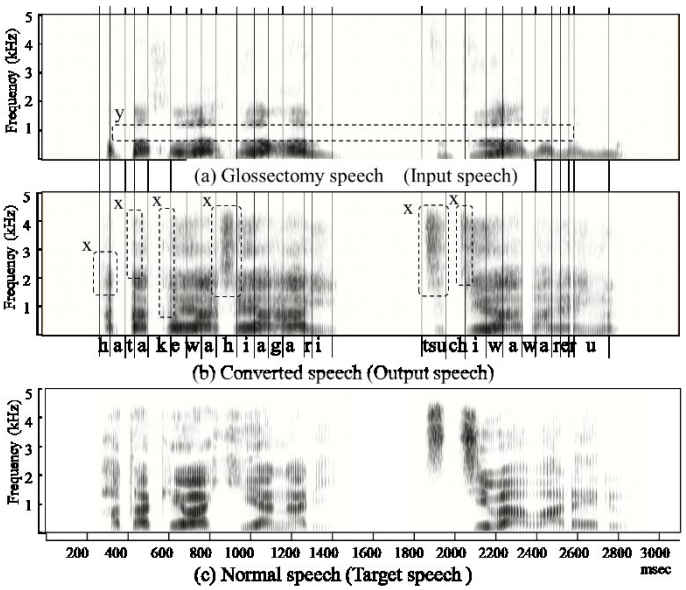



Fig. 4 Spectrograms of glossectomy speech, converted speech, and target speech.

## REFERENCES


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

[2] R. Leonard, 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, 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. (in printing)

[4] M. Abe, S. Nakamura, K. Shikano, H. Kuwabara, "Voice conversion through vector quantization," Proc. ICASSP'88, S14.1, pp.655-658. (1988)

[5] Y. Yoshida, M. Abe, "An algorithm to reconstruct wideband speech from narrowband speech based on codebook mapping," Proc. ICLSP'94, pp.1591-1594. (1994)

[6] K. Kobayashi, T. Toda et al., "Statistical singing voice conversion with direct waveform modification based on the spectrum differential," Proc. INTERSPEECH, pp. 2514-2518. (2014)

[7] Y. Tajiri, K. Tanaka, T. Toda et al., "Non-audible murmur enhancement based on statistical conversion using air- and body-conductive microphones in noisy environments," Proc. INTERSPEECH, pp. 2769-2773. (2015)

[8] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, "Statistical approach to enhancing esophageal speech based on Gaussian mixture models," Proc. ICASSP2010, pp. 4250–4253 (2010)

[9] 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)

[10] A. Kain, M. Macon , "Spectral voice conversion for text-to-speech synthesis," Proc. ICASSP'98, pp. 285–288. (1998)

[11] H. Kawahara, I. Katsue, and A. Cheveigne, "Restructure speech representations using a pitch adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Communication, pp. 187–207. (1999)