Joint Dictionary Learning-based Non-Negative Matrix Factorization for Voice Conversion to Improve Speech Intelligibility After Oral Surgery

*Szu-Wei Fu, Pei-Chun Li, Ying-Hui Lai, Cheng-Chien Yang, Li-Chun Hsieh, and Yu Tsao**

***Abstract*—*Objective*: This paper focuses on machine learning based voice conversion (VC) techniques for improving the speech intelligibility of surgical patients who have had parts of their articulators removed. Because of the removal of parts of the articulator, a patient's speech may be distorted and difficult to understand. To overcome this problem, VC methods can be applied to convert the distorted speech such that it is clear and more intelligible. To design an effective VC method, two key points must be considered: (1) the amount of training data may be limited (because speaking for a long time is usually difficult for post-operative patients); (2) rapid conversion is desirable (for better communication). *Methods*: We propose a novel joint dictionary learning-based non-negative matrix factorization (JD-NMF) algorithm. Compared to conventional VC techniques, JD-NMF can perform VC efficiently and effectively with only a small amount of training data. *Results*: The experimental results demonstrate that the proposed JD-NMF method not only achieves notably higher short-time objective intelligibility (STOI) scores (a standardized objective intelligibility evaluation metric) than those obtained using the original unconverted speech but is also significantly more efficient and effective than a conventional exemplar-based NMF VC method. *Conclusion*: The proposed JD-NMF method may outperform the state-of-the-art exemplar-based NMF VC method in terms of STOI scores under the desired scenario. *Significance*: We confirmed the advantages of the proposed joint training criterion for the NMF-based VC. Moreover, we verified that the proposed JD-NMF can effectively improve the speech intelligibility scores of oral surgery patients.**

***Index Terms*—Joint dictionary learning, nonnegative matrix factorization, sparse representation, voice conversion.**

I. INTRODUCTION

The speech of a person after oral surgery is often difficult to understand for untrained listeners [1-3]. Therefore, such patients may desire a voice conversion (VC) system that can convert their voice into clear speech. In this study, we investigated the use of a VC approach to improve the intelligibility of the speech of patients who have had parts of their articulators removed during surgery.

Normal VC tasks are designed to transform the speech of the source speaker to make it sound like that of another (target) speaker. More recently, VC methods have been adopted for various medical applications. Aihara *et al.* [4, 5] proposed a VC system for articulation disorders that attempts to preserve the speaker's individuality based on a combined dictionary containing the source speaker's vowels and the target speaker's consonants. Toda *et al.* [6, 7] tried to apply VC to transform nonaudible murmurs into normal speech. Liu *et al.* [8] proposed a method for applying VC-based frequency-lowering technology for Mandarin-speaking hearing aid users.

Numerous VC methods have been proposed in the past. One notable class of methods uses a parametric model to map the acoustic features of the source speaker to those of the target speaker. The joint density Gaussian mixture model (JD-GMM) is known as an effective mapping model for VC [9-11]. JD-GMM implements a linear transformation function based on a Gaussian mixture model (GMM). Conversion parameters are estimated using the maximum likelihood [9], minimum mean-square error [10], or maximum mutual information [11] criteria. Numerous extensions of JD-GMM have been proposed to solve its intrinsic over-smoothing problem caused by statistical averaging [9, 12]. An artificial neural network (ANN) is another notable model that has been confirmed effective for VC [13-16]. Owing to its complex structure, an ANN model is capable of characterizing the nonlinear relationship between the utterances by different speakers. Since the emergence of deep learning, VCs based on deep neural networks have gained considerable attention [17-20].

Although model-based VC methods have been confirmed as being effective for various tasks, they usually require a certain amount of training data. When there is insufficient training data, the models may incur an overfitting problem, such that the sound quality of the converted speech is poor. To overcome the

Szu-Wei Fu is with Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan and Research Center for Information Technology Innovation (CITI) at Academia Sinica, Taipei, Taiwan.
Pei-Chun Li is with Department of Audiology and Speech Language Pathology, Mackay Medical College, Taipei, Taiwan.
Ying-Hui Lai is with Department of Electrical Engineering, Yuan Ze University.
Cheng-Chien Yang and Li-Chun Hsieh are with Mackay Memorial Hospital, Taipei, Taiwan.
Yu Tsao is with the Research Center for Information Technology Innovation (CITI) at Academia Sinica, Taipei, Taiwan (e-mails: yu.tsao @citi.sinica.edu.tw).

possible overfitting problem, several nonparametric exemplar-based VC methods [21-24] have been proposed as alternatives to model-based frameworks. This class of methods assumes that a target spectrogram can be generated from a set of basis target spectra (a dictionary), namely exemplars, through weighted linear combinations. Based on the nonnegative nature of the spectrogram, the nonnegative matrix factorization (NMF) technique [25] is employed to estimate the nonnegative weight. At runtime, the activations for each source spectrogram are estimated through the source dictionary and are then applied to the target dictionary to generate the corresponding target spectrogram. Therefore, the converted utterances are directly produced from real target exemplars rather than from model parameters. To include the temporal contextual constraint, multiple-frame exemplars are used in the source dictionary [21]. Wu *et al.*[22, 26] proposed a joint NMF framework to efficiently estimate the activations by simultaneously taking two distinct acoustic features (one being low-resolution, and the other high-resolution) into consideration. Although only limited training data are needed by exemplar-based NMF models, most of the data are crudely used as exemplars, implying that a large dictionary will be constructed. The main limitation of using a large dictionary is the long conversion time, which violates our rapid conversion requirement.

In this study, we focused our attention on NMF-based VC techniques for patients of oral surgery, for which two key points should be addressed: (1) the amount of training data may be limited (because speaking for a long time is usually difficult for post-surgical patients); (2) a rapid conversion is desirable (to facilitate those users with better communication). To address these two points, we propose a novel joint dictionary learning-based NMF (JD-NMF) VC algorithm. The JD-NMF algorithm simultaneously learns source and target dictionaries (joint dictionary). By specifying a small number of bases using the NMF technique, JD-NMF can learn a set of bases that are representative of the entire set of exemplars (estimated from the training data). Accordingly, the size of the dictionary in JD-NMF can be significantly reduced relative to exemplar-based NMF, thus improving the online conversion efficiency [20, 22].

The remainder of this paper is organized as follows: Section II reviews the conventional NMF-based VC. Section III details the proposed method. The experimental results are evaluated in Section IV. Finally, Section V presents our conclusions.

II. Related Work

A. NMF-based Speech Representation

The basic concept of NMF-based VC is to represent a magnitude spectrum as a linear combination of a set of bases; this collection of bases is called a dictionary. In the conventional exemplar-based NMF model, each basis in the matrix is a speech frame (exemplar) in the training data. More specifically, the bases are directly copied from the training data, and no learning process is involved to build the dictionary. Assume that I exemplars were collected, we then have a dictionary $\mathbf{A}=[\mathbf{a}_1,\mathbf{a}_2,...\mathbf{a}_I]\in R^{F\times I}$, where $\mathbf{a}_i$ is the i^{th} exemplar, and F is the feature dimension. Then, the speech sample at the l-th frame, $\mathbf{x}_l \in R^{F\times 1}$, can be represented by:

$$\mathbf{x}_l \approx \mathbf{A}\mathbf{h}_l = \sum_{i=1}^{I}\mathbf{a}_i h_{i,l} \quad \text{subject to } \mathbf{A}\geq 0, \mathbf{h}_l \geq 0 \tag{1}$$

where $\mathbf{h}_l = [h_{1,l}, h_{2,l}, ...h_{I,l}]\in R^{I\times 1}$ is the activation vector, and $h_{i,l}$ is the nonnegative weight (or activation) of the i^{th} exemplar.

As each speech sample is modeled independently, the spectrogram of each utterance can be represented in matrix form as

$$\mathbf{X}\approx\mathbf{AH} \tag{2}$$

where $\mathbf{X}\in R^{F\times M}$ is the spectrogram, M is the number of frames in the utterance, and $\mathbf{H}\in R^{I\times M}$ is the corresponding activation matrix for which the column vector is an activation vector $\mathbf{h}_l$. To minimize the distance between $\mathbf{X}$ and $\mathbf{AH}$, Lee [27] proposed multiplicative updating rules to alternately optimize $\mathbf{A}$ and $\mathbf{H}$ by gradient descent with a specific learning rate.

B. Exemplar-based NMF for Voice Conversion

1.) Offline Stage

For VC, paired source–target dictionaries $\mathbf{A}_E$ and $\mathbf{B}_E$ are required with acoustically aligned exemplars. In exemplar-based NMFs, both the source and target dictionaries are directly obtained from the data itself. To construct the paired dictionaries, a parallel dataset (between the source and target speakers) is collected. However, because of their different speech rates, the two dictionaries may not align with each other. Therefore, dynamic programing techniques such as dynamic time warping (DTW) [28] must be applied to obtain frame-wise source–target alignment [23]. Fig. 1 shows an example of the source–target dictionaries. For visual presentation, only 40 frames (bases) were randomly selected from the training data. The *x*-axis shows the basis index, and the *y*-axis denotes the frequency bins. Further, the intensity is represented by colors. In this example, we used 512 discrete-Fourier transform points to characterize 16-KHz speech sounds.

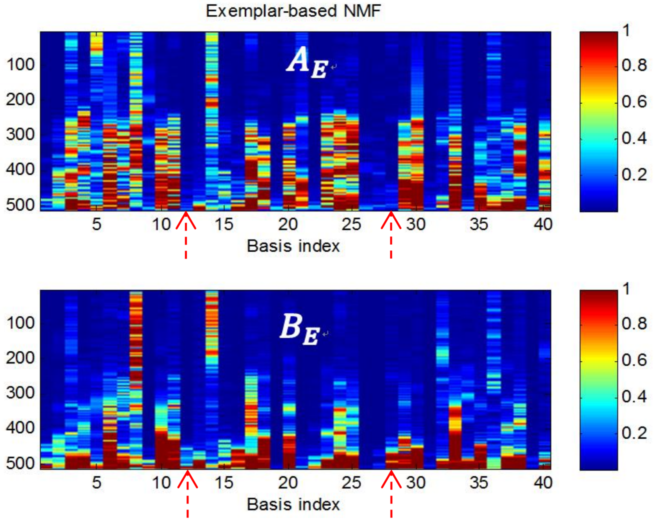


Fig. 1. Source and target dictionaries used in exemplar-based NMF.

2.) Online Stage

To generate the converted speech spectrogram, we assume that the aligned source and target dictionaries can share the same activation matrix **H**. Therefore, the converted spectrogram can be represented as

$$\mathbf{Y}_s = \mathbf{B}_E\mathbf{H} \qquad (3)$$

where $\mathbf{Y}_s \in R^{F\times T}$ is the converted spectrogram, $\mathbf{B}_E \in R^{F\times I}$ is the fixed target dictionary of the exemplars from the target training data, and **H** is determined by the source spectrogram $\mathbf{X}_s \in R^{F\times T}$ and source dictionary $\mathbf{A}_E$, as shown in (4).

Owing to the nonnegative nature of the magnitude spectrogram, the NMF technique is employed to estimate the activation matrix **H** by minimizing the following objective function:

$$\mathbf{H} = \arg\min_{\mathbf{H}\geq 0} \; d(\mathbf{X}_s, \mathbf{A}_E\mathbf{H}) + \lambda \|\mathbf{H}\|_1 \qquad (4)$$

where λ is the sparsity penalty factor, $\|.\|_1$ represents the L1-norm and $d(\cdot)$ is a distance measure. As the number of exemplars I is usually large in exemplar-based NMFs, the sparsity constraint [29-31] is adopted such that only a few exemplars are activated at any one time. A multiplicative updating rule for two criteria (the Euclidean distance and Kullback–Liebler (KL) divergence) was proposed in [27]. Other divergence measures and updating rules can be found in [32]. However, in the application of VC, KL divergence is observed to be more suitable [33]. Therefore, (4) can be minimized by iteratively applying the following multiplicative updating rule:

$$\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\mathbf{A}_E{}^T \dfrac{\mathbf{X}_s}{\mathbf{A}_E\mathbf{H}}}{\mathbf{A}_E{}^T\mathbf{1}+\lambda} \qquad (5)$$

where $\otimes$ represents element-wise multiplication (divisions are also performed element-wise), and $\mathbf{1} \in \mathrm{R}^{F\times M}$ is an all-ones matrix.

Fig. 2 illustrates the overall framework for exemplar-based VC.

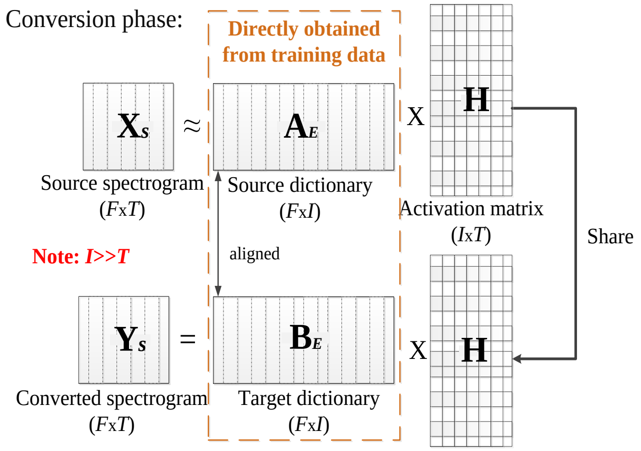


Fig. 2. Online stage of exemplar-based NMF for VC. Note that the number of exemplars I is usually much larger than the number of frames T in a source utterance under this framework.

III. Proposed Joint Dictionary Learning NMF for Voice Conversion

In other applications of NMF (e.g., speech enhancement [34, 35] and source separation [36, 37]), the dictionary is learned from the training data. However, in conventional exemplar-based NMF, the dictionary is directly copied from the training data. In other words, there is no training stage involved in the exemplar-based NMF framework, which save the offline training procedure but gives rise to a drawback: when the number of bases is large, the computational cost can be high in the conversion phase. This implies the exemplar-based NMF may not be suitable for our application scenario (a rapid conversion is desirable for better communication). For example, in [23], although the obtained level of performance was better than that of other well-known methods (namely, JD-GMM), the exemplar-based method required 19.02 s to generate 1 s of target speech. To solve the problem, we propose the JD-NMF framework, which spends more time in the offline (training) stage extracting a set of more meaningful (i.e., compact) basis representations. In the online (runtime) stage, based on the estimated bases, JD-NMF estimates the activation matrix and performs VC.

A. *Offline Stage*

In addition to applying DTW to align the training data in the same way as in the exemplar-based NMF, the proposed JD-NMF includes a training phase in the offline stage. In previous studies [8, 21], it has been confirmed that when performing NMF-based voice conversion, preparing a pair of coupled dictionaries is important since the activation matrix is shared by the source and target basis matrices. This suggests that the two dictionaries should be trained simultaneously instead of independently. Therefore, we propose the JD-NMF framework and modify the objective function to simultaneously learn the two dictionaries, as follows:

$$\mathbf{A}_J, \mathbf{B}_J = \arg\min_{\mathbf{A}_J,\mathbf{B}_J\geq 0} \; d(\mathbf{X}, \mathbf{A}_J\mathbf{H}) + d(\mathbf{Y}, \mathbf{B}_J\mathbf{H}) + \lambda \|\mathbf{H}\|_1$$

(6)

where $\mathbf{X} \in R^{F\times I}$ and $\mathbf{Y} \in R^{F\times I}$ are the paired source and target training data, respectively; $\mathbf{A}_J \in R^{F\times r}$ and $\mathbf{B}_J \in R^{F\times r}$ are learned dictionaries and r is the number of bases, which can be set by the users. Note that (6) is used to approximate the source and target spectrograms, provided the same activation matrix **H** is used. More specifically, to reconstruct the coupled training data ($\mathbf{X}$ and $\mathbf{Y}$) with shared $\mathbf{H}$, the learned dictionaries ($\mathbf{A}_J$ and $\mathbf{B}_J$) are forced to couple with each other to minimize the distance (KL divergence). Therefore, if the source and target training data are aligned, the learned i^{th} source basis $\mathbf{a}_i$ will represent the same basic speech unit as the i^{th} target basis $\mathbf{b}_i$.

To solve (6) by using the KL divergence as the criterion [27], the first two terms can be formulated as follows:

$$d(\mathbf{X},\mathbf{A}_J\mathbf{H})+d(\mathbf{Y},\mathbf{B}_J\mathbf{H})$$

$$=\sum_{\substack{1\le f\le F\\1\le i\le I}}(\mathbf{X}_{f\,i}\log\frac{\mathbf{X}_{f\,i}}{(\mathbf{A}_J\mathbf{H})_{f\,i}}-\mathbf{X}_{f\,i}+(\mathbf{A}_J\mathbf{H})_{f\,i})$$

$$+\sum_{\substack{1\le f\le F\\1\le i\le I}}(\mathbf{Y}_{f\,i}\log\frac{\mathbf{Y}_{f\,i}}{(\mathbf{B}_J\mathbf{H})_{f\,i}}-\mathbf{Y}_{f\,i}+(\mathbf{B}_J\mathbf{H})_{f\,i})$$

$$=\sum_{\substack{1\le f\le F\\1\le i\le I}}(\mathbf{X}_{f\,i}\log\frac{\mathbf{X}_{f\,i}}{(\mathbf{A}_J\mathbf{H})_{f\,i}}-\mathbf{X}_{f\,i}+(\mathbf{A}_J\mathbf{H})_{f\,i}+\mathbf{Y}_{f\,i}\log\frac{\mathbf{Y}_{f\,i}}{(\mathbf{B}_J\mathbf{H})_{f\,i}}-\mathbf{Y}_{f\,i}+(\mathbf{B}_J\mathbf{H})_{f\,i})\quad(7)$$

As the operations in (7) are all element-wise, we can cascade $\mathbf{X}$ with $\mathbf{Y}$, and $\mathbf{A}_J$ with $\mathbf{B}_J$, to further simplify the objective function (6), as follows:

$$\sum_{\substack{1\le f\le 2F\\1\le i\le I}}(\mathbf{S}_{f\,i}\log\frac{\mathbf{S}_{f\,i}}{(\mathbf{WH})_{f\,i}}-\mathbf{S}_{f\,i}+(\mathbf{WH})_{f\,i})+\lambda\,\|\,\mathbf{H}\,\|_1\qquad(8)$$

$$=d(\mathbf{S},\mathbf{WH})+\lambda\,\|\,\mathbf{H}\,\|_1\qquad(9)$$

where

$$\mathbf{S}=\begin{bmatrix}\mathbf{X}\\\mathbf{Y}\end{bmatrix}\in R^{2F\times I},\ \mathbf{W}=\begin{bmatrix}\mathbf{A}_J\\\mathbf{B}_J\end{bmatrix}\in R^{2F\times r}\qquad(10)$$

Therefore, the objective function in (6) is equivalent to the simplified objective function in (9). We can simply apply the conventional alternately multiplicative updating rules proposed in [27] to determine the joint dictionary $\mathbf{W}$.

$$\mathbf{W}\leftarrow\mathbf{W}\otimes\frac{\frac{\mathbf{S}}{\mathbf{WH}}\mathbf{H}^T}{\mathbf{1H}^T}\qquad(11)$$

$$\mathbf{H}\leftarrow\mathbf{H}\otimes\frac{\mathbf{W}^T\frac{\mathbf{S}}{\mathbf{WH}}}{\mathbf{W}^T\mathbf{1}+\lambda}\qquad(12)$$

Here, $\mathbf{1}\in\mathrm{R}^{2F\times I}$ is an all-ones matrix. Note that although we also update $\mathbf{H}$ by using (12), our goal is merely to obtain the joint dictionary $\mathbf{W}$. Fig. 3 shows an example of learned dictionary $\mathbf{W}$ with the source dictionary ($\mathbf{A}_J$) and the target dictionary ($\mathbf{B}_J$) occupy the upper and lower halves of $\mathbf{W}$, respectively. In this example, we used 512 fast-Fourier transform points to characterize 16-KHz speech sounds, and thus F = 513 in (10), and $\mathbf{W}$ is a 1026 × 40 matrix. From Fig. 3, we can see that: 1) the bases of $\mathbf{A}_J$ and $\mathbf{B}_J$ are aligned and learned jointly. During implementation, the "DTW alignment process" is actually very important for both exemplar-based NMF (Fig. 1) and JD-NMF. When the source and target speech signals are not aligned precisely, the dictionaries may not be coupled very well. 2) the middle frequency components of $\mathbf{A}_J$ are relatively noisy compared to those of $\mathbf{B}_J$, and the bases of $\mathbf{B}_J$ are more discriminative toward each other than those of $\mathbf{A}_J$. Note that $\mathbf{A}_J$ and $\mathbf{B}_J$ are learned from distorted speech (after surgery) and clear speech (before surgery), respectively. The second observation can explain why the distorted speech sounds blurred, leading to the speech having poor intelligibility. When comparing Fig.1 and Fig. 3, we can note that the bases in Fig.1 are not very representative (e.g. four bases, namely the 26, 27, 31and 39 bases are silence). In addition, some bases of the two dictionaries in Fig.1 do not couple well (e.g. the bases in index 12, 28, as indicated by. the red arrows) due to imperfect alignments of DTW. On the other hand, since our joint dictionaries are learned from the whole training data, the issue of imperfect alignments can be mitigated. In the next section, we introduce the conversion of distorted speech to clear speech by using $\mathbf{A}_J$ and $\mathbf{B}_J$ with a shared activation matrix. To reduce the computational cost during the online stage, the number of bases r should be minimized. In Section IV, we show that only a few representative bases, learned using the joint training criterion, are sufficient to obtain a satisfactory result.

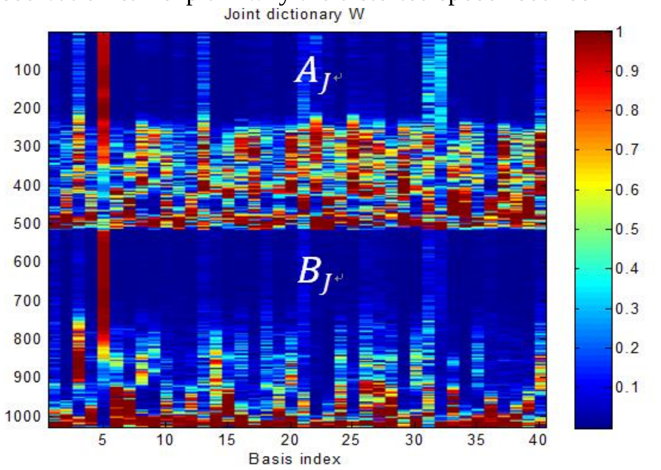


Fig. 3. Source and target dictionaries; the two dictionaries can be obtained by separating the upper and lower halves of the joint dictionary, **W**, respectively.

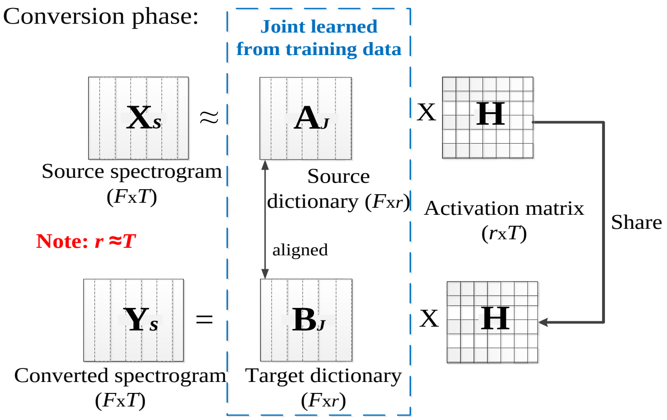


Fig. 4. Online stage of the proposed JD-NMF for VC. Compared to those in Fig. 2, the dictionaries and activation matrix are much smaller.

B. *Online Stage*

As the proposed JD-NMF and conventional exemplar-based NMF methods differ mainly in the training phase, the conversion process can be similarly presented, as shown in Fig. 4. Note that the sizes of the dictionaries and activation matrix are much smaller than those shown in Fig. 2. In the same way as in

(3), we can obtain the converted speech by

$$\mathbf{Y}_s = \mathbf{B}_J\mathbf{H} \tag{13}$$

where $\mathbf{Y}_s \in R^{F\times T}$ is the converted spectrogram. Note that the number of bases *r* in our framework is much smaller than that in the conventional method. The multiplicative updating rule in (5) can be modified as follows:

$$\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\mathbf{A}_J{}^T \dfrac{\mathbf{X}_s}{\mathbf{A}_J\mathbf{H}}}{\mathbf{A}_J{}^T\mathbf{1}+\lambda} \tag{14}$$

where $\mathbf{X}_s \in R^{F\times T}$ is the source spectrogram that is to be converted, and $\mathbf{1} \in \mathrm{R}^{F\times T}$ is an all-ones matrix.

From (13) and (14), we can see that as the size (number of columns) of $\mathbf{A}_J$ decreases, we can save large amounts of computational time when computing the activation matrix **H**, thus making rapid conversion feasible. To further analyze the computation cost, the number of multiplications or divisions needed for each iteration in (14) can be estimated as follows:

$$2FrT + 2rT + FT \tag{15}$$

which is a linear function of *r* when both *F* and *T* are fixed (given $\mathbf{X}_s$).

In the next section, we incorporate the context information to further improve the JD-NMF performance.

C. Contextual Information

To consider context information in many applications of speech processing, the features are cascaded such that they span multiple consecutive frames. However, in (3) to (5), no temporal information is considered, that is, each frame is modeled independently. Therefore, in [21, 22], multiple-frame exemplars were used in the source dictionary to more accurately estimate the activation matrix.

In our JD-NMF framework, we also proposed to cascade the spectrograms across multiple consecutive frames to train an extended joint dictionary. Accordingly, in the offline phase, **X** and **Y** become $R^{(2q+1)F\times I}$, which in turn causes $\mathbf{A}_J$ and $\mathbf{B}_J$ to expand into $R^{(2q+1)F\times r}$, where $2q+1$ is the window size of each frame. Note that, when calculating the computational cost, the dimensionality of spectra *F* in (15) also must expand to $(2q+1)F$. During the conversion phase, to utilize the extended dictionary, the source spectrogram $\mathbf{X}_s$ is also cascaded to estimate the activation matrix. Meanwhile, the cascaded target dictionary can also consider contextual information to yield other benefits. Fig. 5 illustrates a sequence of speech frames. In the figure, when the fifth frame is to be converted, it will consider five multiple-frame vectors (within the blue dotted lines) obtained from the first to the ninth frames, that is, up to $\pm 2q$ frames. Therefore, to produce the final converted fifth frame, we can calculate the average to integrate the information provided in the five multiple-framed vectors indicated with the red arrows. In addition, the average operation can reduce noise and smooth the transition between speech sounds. Thus, cascading of the training spectrogram can greatly improve the quality of the converted speech in the proposed JD-NMF framework.

IV. Experiments

The goal of the present study is to propose a rapid VC system for patients after oral surgery. Two objective evaluations are considered: the intelligibility of the converted speech and the computational cost of the conversions. A standardized evaluation method, short-time objective intelligibility (STOI) [38, 39], was employed as our objective intelligibility measure. The calculation of STOI is based on the correlation between the temporal envelopes of the target and the converted speech for short segments. The output STOI score ranges from 0 to 1, and is expected to be monotonically related to the average intelligibility of the converted speech. Hence, a higher STOI value indicates better speech intelligibility. To evaluate the computational cost, the number of multiplications or divisions for each iteration is employed, as presented in (15). In addition, we measured the real execution time for the online phase for comparison.

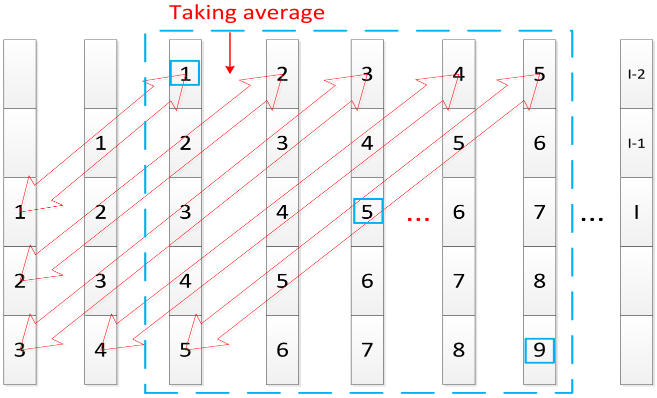


Fig. 5. Calculating average in multiple-frame vectors to reduce noise during conversion phase [here in this example, window size (2*q* + 1) = 5].

In the experiments of this study, we prepared 150 short sentences as our corpus; from among these, 70 utterances were randomly selected as a training set, 40 utterances were randomly selected as a development set, and the remaining 40 utterances were used as an evaluation set. A physically unimpaired male was chosen as the target speaker. We recorded 150 sentences uttered by four patients after oral surgery and also by the target speaker. The procedures were reviewed and approved by the local institutional review board committees. The speech signals were sampled at 16 kHz and windowed with a 20-ms Hamming window every 10 ms. The parameters in the dictionaries and the activation matrix are initialized with random numbers from normal distribution (mean=0 and standard deviation=1, with absolute value). With the initialized dictionaries and the activation matrix, we then update them by (11), (12) and (14) [29]. To reduce the effect of the random initializations of the matrix in NMF, each set of experiments was repeated 10 times and average values were obtained [40, 41]. Because the proposed JD-NMF uses a much smaller number of bases than the exemplar-based NMF, the sparsity constraint is not applied ($\lambda = 0$) in (6) and (14).

In the following discussion, experiments *A* to *C* were conducted using the training data and development set for the offline and online stages, respectively. The best parameters were then used to test the performance by using the data in the evaluation set for the online stage. The results were presented in experiment *D*.

A. Cascaded Dictionaries

We first examined the effect of using multiple-frame bases to check whether the contextual information is useful. Fig. 6 presents the STOI results (*y*-axis) of the development sets as a function of the window size *L* (*x*-axis). Here, we compare three different sets of results: 1) cascading of only the source data (only $\mathbf{A}_J$ in Fig. 4 was extended); 2) cascading of only the target training data (only $\mathbf{B}_J$ in Fig. 4 was extended); 3) cascading of both the source and target training data (both $\mathbf{A}_J$ and $\mathbf{B}_J$ in Fig. 4 were extended). These are represented by the blue, green, and red lines, respectively. Fig. 6 shows three observations. (1) Cascading of the source training spectrogram can facilitate the estimation of the activation matrix more accurately. Therefore, as the window size increases, the STOI values consistently increase. (2) When only the target training spectrogram is cascaded, the performance can be significantly improved because of the averaging operation and the large amount of contextual information. However, if too many frames are considered at one time, the performance may degrade. (3) To obtain the greatest improvement, both the source and target training data should be cascaded with a carefully selected window size. As indicated by the red line, a good trade-off between intelligibility and computation cost can be achieved when the window size is 5. Therefore, we used this window size in the following experiments.

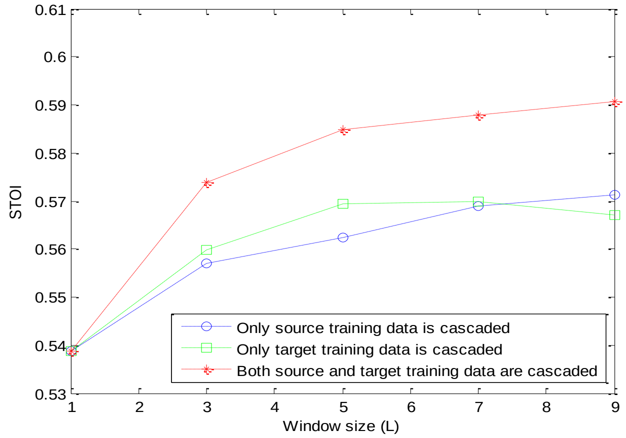


Fig. 6. STOI results for development set as a function of window size *L*.

B. Effect of Number of Bases and Iterations

There are two other parameters that can affect the performance (in terms of speech intelligibility and computational cost) of the JD-NMF framework: the number of bases in the dictionary and the number of iterations during conversion. To determine their degrees of influence, we calculated the STOI with different settings for the development set. Fig. 7 presents the STOI scores (*y*-axis) as a function of the number of bases (*x*-axis) under different iteration numbers. The results show that the number of bases and iterations influence each other. Therefore, we examined them separately in detail, as follows.

1.) Effect of Number of Bases

First, we examined the changes in STOI for different numbers of bases *r*. As the dictionary is learned from the training data in our proposed method, we can set different sizes for the dictionary. Fig. 7 shows that the STOI increases with the number of bases when the iteration number is small. However, if the algorithm iterates too many times, it leads to overfitting, suggesting that any more bases will degrade the intelligibility of the converted speech.

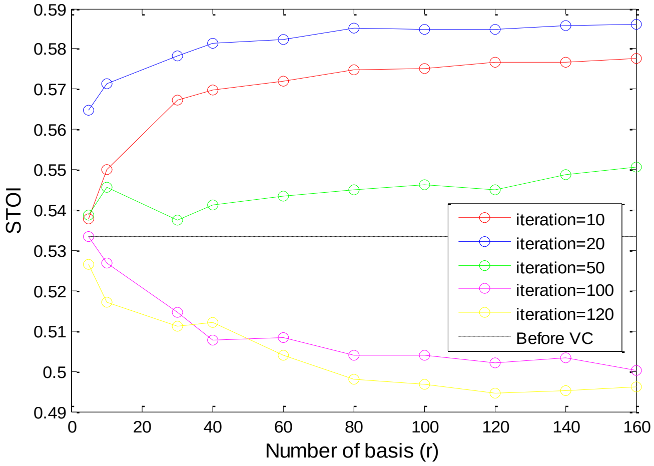


Fig. 7. STOI as a function of number of bases *r* in dictionary under different iteration numbers.

Hence the effect of adding more bases is different and depends on the iteration numbers.

Note that, with 80 bases, the STOI improvement starts to saturate when the iteration number is small (10 and 20). In this case, adding more bases brings very limited improvement. As indicated in (15), because a smaller dictionary can expedite conversion, we use *r* = 80 bases in the evaluation set.

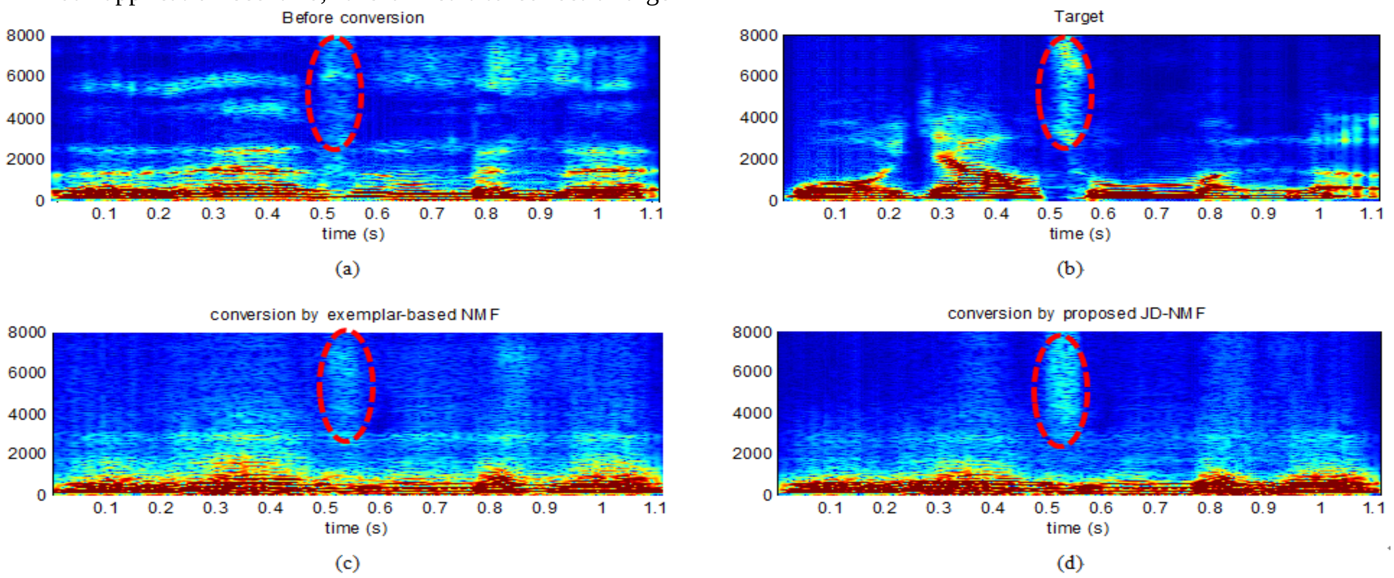
2.) Effect of Number of Iterations

The iteration number is usually experimentally determined from a development set [21, 22]. If the iteration number is too few/many, the learned model will be underfitting/overfitting to the training data. Fig. 7 shows that, when the algorithm iterates too many times, the STOI values begin to decrease because of overfitting. Although the difference between the source spectrogram $\mathbf{X}_s$ and modeled spectrogram $\mathbf{A}_J\mathbf{H}$ is always guaranteed to be reduced after each iteration (from the derivation of NMF [27]) by updating $\mathbf{H}$, there is no theoretical guarantee that the converted speech can be accordingly improved by running more iterations. Hence, if $\mathbf{H}$ overfits $\mathbf{X}_s$, it may produce distorted and unsmooth converted speech $\mathbf{B}_J\mathbf{H}$. To overcome this problem, we can simply stop the iteration earlier; this regularization method is also called early stopping [42]. From Fig. 7, with 20 iterations, we can achieve the highest STOI value without spending too much computational time; therefore, the iteration number is set to 20 for the evaluation set.

In brief, in the proposed method, the sizes of both the source and target dictionaries are set to (513 × 5) × 80 with 20 iterations during the conversion.

C. *Amount of Training Data*

In our application scenario, it is difficult to collect a large amount of training data because speaking for too long is difficult for post-surgical patients. Therefore, we examined the robustness of the proposed method for different amounts of training data; the results are presented in Fig. 8. The *x*- and *y*-axes denote the number of sentences used for training and STOI scores, respectively.



(a) (b) (c) (d)

Fig. 10. Spectrograms of source, target, and converted speech after DTW in evaluation set. (a) Source speech (before conversion), (b) target speech, and (c) speech converted by exemplar-based NMF and (d) by proposed JD-NMF. (More examples are available at http://jackylai.comli.com/TBME_web/Reviewer2_C5_v2.htm)

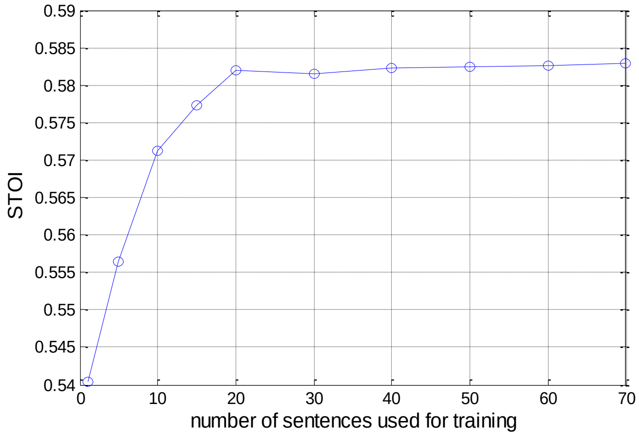


Fig. 8. STOI as a function of the number of sentences used for training.

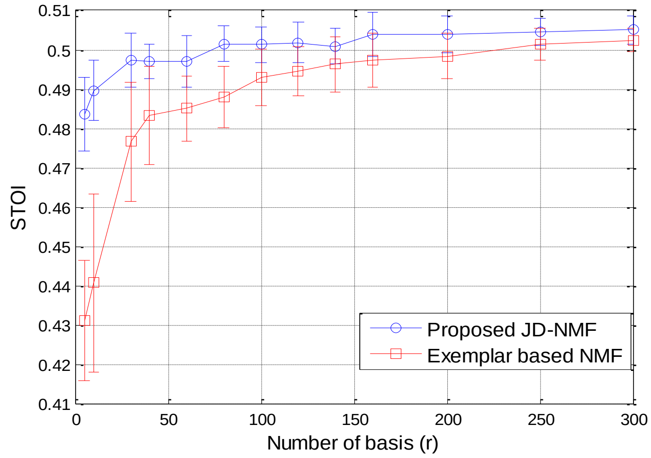


Fig. 9. STOI as function of number of bases for proposed and baseline methods.

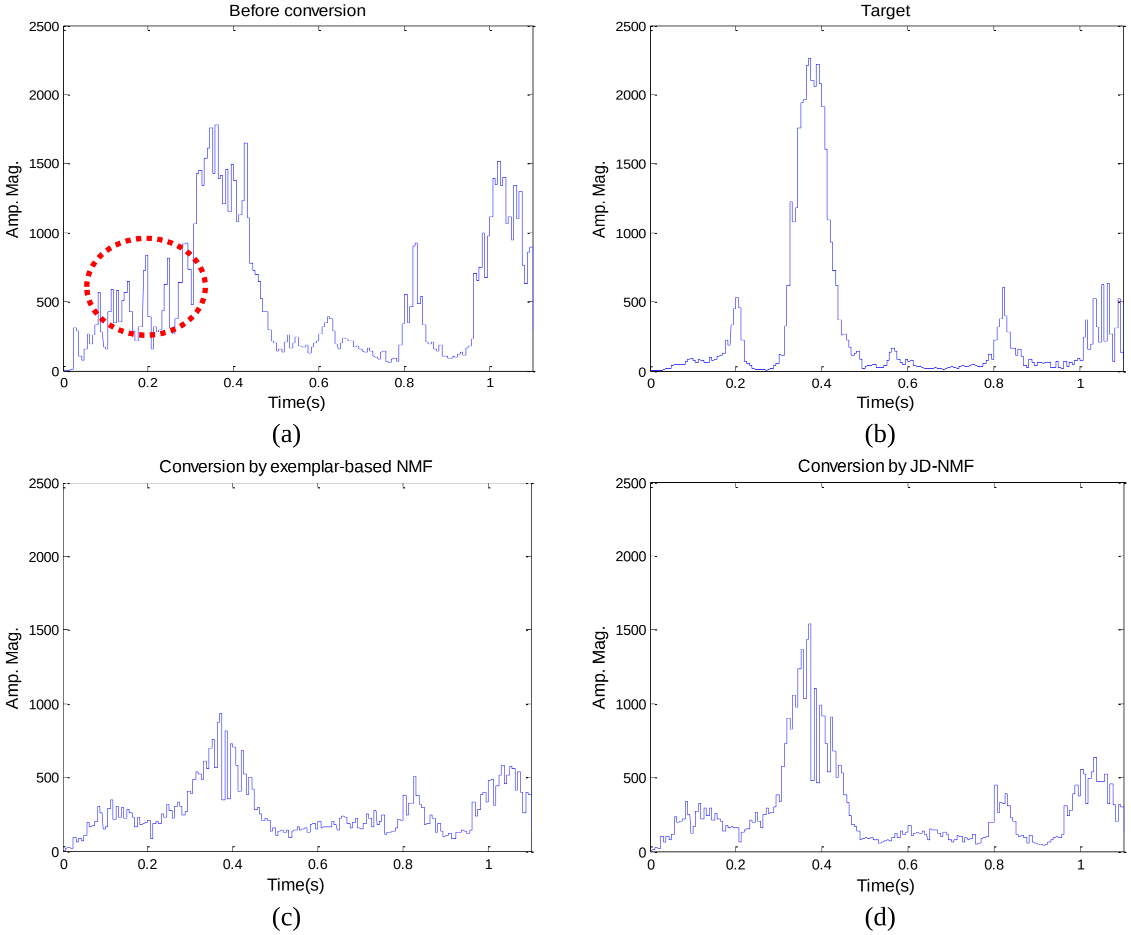
The number of sentences varied from 1 to 70; the sentences were randomly selected from the original training set. Fig. 8 shows that the STOI improvement begins saturating when roughly 20 sentences are used for training. In other words, our system can be trained with only 20 sentences.

D. *Overall Performance Comparison*

Finally, we compared the proposed JD-NMF and baseline (exemplar-based NMF) methods using the evaluation set. Figure 9 presents the STOI scores (mean and standard deviation) as a function of the number of bases for comparison. The same number of bases was used for JD-NMF and exemplar-based NMF, and the bases used for the exemplar-based NMF were randomly selected from the prepared exemplars of the training data. For the results of exemplar-based NMF, the window size and iteration number are optimized based on the development set. Notably, each mean result in Fig. 9 was an average of 1600 scores (40×4×10): 40 testing utterances recorded by 4 patients along with 10 random initials to avoid the randomness issue [29, 40]. Meanwhile, each standard deviation in Fig. 9 is estimated from 10 results (obtained from 10 different random initials), and each of these 10 results is the average of 160 STOI scores (40 testing utterances recorded by 4 patients).

From the figure, it is noted that when the size of the dictionary is small, JD-NMF significantly outperforms exemplar-based NMF. For example, the STOI of JD-NMF with 80 bases is roughly the same as that of exemplar-based NMF with 300 bases. This implies that the jointly learned bases provide more meaningful information than the directly obtained exemplars. We would like to emphasize that the present study focuses on two major requirements: limited training data and rapid online conversion. Thus, we only present the results of exemplar-based NMF and JD-NMF with bases less than 300.

To estimate the computational cost savings, the number of multiplications and divisions per frame in (15) can be applied with a feature dimension *F* set of 513 × 5. To practically compare the computation load, we also compared the execution time required to generate the activation matrix during conversion of one target utterance (1.2 s), on a 3.6-GHz Intel i7 core PC implemented in MATLAB. Both results are listed in Table I, in which we can observe that the proposed JD-NMF and the exemplar-based NMF require 413,125 and 1,542,165 numbers of multiplications and divisions, respectively. In other words, the ratio of the computation of the two methods (denoted as J/E) is 0.268. The execution times of JD-NMF and exemplar-based NMF were 0.1177 and 0.3332 s, respectively. The ratio of J/E in terms of execution time is thus 0.3532. The aforementioned results confirm that our proposed method can reduce the online computation by around a factor of three, relative to the conventional method.



(a) (b)

(c) (d)

Fig. 11. Amplitude envelopes from fifth channel of source, target, and converted speech: (a) Source speech (before conversion); (b) target speech; and (c) speech converted through exemplar-based NMF and (d) by the proposed JD-NMF. (More examples are available at http://jackylai.comli.com/TBME_web/Reviewer2_C5_v2.htm).

TABLE I
COMPARISON OF COMPUTATIONAL LOAD

Methods	# of multiplications and divisions (Eq. (15))	Execution time (s)
Proposed JD-NMF (80 bases) (J)	413,125	0.1177
Exemplar-based NMF (300 bases) (E)	1,542,165	0.3332
Ratio (J/E)	0.2679	0.3532

Next, we visually investigated the effect of JD-NMF VC on the distorted speech by using spectrogram plots. A spectrogram plot displays the variations of frequencies present in a speech signal [43, 44]. The *x*-axis denotes the time index while the *y*-axis denotes frequency bins, with the intensity represented by color (red corresponds to high intensities, while blue corresponds to low intensities). Fig. 10 shows spectrograms of a source and target speech pair after alignment through DTW (Fig. 10 (a) and (b), respectively), with speech conversion using the exemplar-based NMF and proposed JD-NMF (Fig. 10 (c) and (d), respectively). The figures show that the consonant sound (the region in the red circle) before conversion is unclear because of the articulators being removed. Moreover, the middle frequency components for the speech before conversion are relatively noisy. This observation is also observed in Fig. 2. Next, we note that, although the exemplar-based NMF can slightly enhance a consonant, it also produces a wide range of noise, especially at high frequencies. In contrast, our proposed JD-NMF can significantly improve the consonant part while maintaining a clean high-frequency part, which is more akin to the property of the target speech.

Finally, we provide another qualitative comparison of exemplar-based and JD-NMF VCs through the processed envelopes. Previous studies suggested that the modulation depth is also an important factor affecting speech perception [45, 46]. A higher modulation depth accounts for better speech intelligibility. In this study, we applied an eight-channel tone vocoder used in [46] as a tool to extract the envelopes under different frequency bands. In [47], it was pointed out that the middle frequency band is extremely important for speech intelligibility; therefore, only envelopes in the fifth channel (1158–1790 Hz) were adopted for comparison. Fig. 11 shows the amplitude envelopes from the fifth channel of one source and target speech pair after alignment through DTW (Fig. 11. (a) and (b), respectively), with speech conversion by the exemplar-based and proposed JD-NMFs (Fig. 11. (c) and (d), respectively). The *x*- and *y*-axes denote the time index and amplitude magnitude, respectively. Fig. 11 shows that the envelope before conversion has distortion at around 0.2 s (in the red circle) and less modulation depths than that of the target speech. Moreover, while both the exemplar-based and JD-NMF VCs can reduce the distortion, the modulation depth of the latter is much higher. Finally, a comparison of Fig. 11 (b) and (d) shows that the envelope of JD-NMF VC closely resembles that of the target speech, implying better speech intelligibility.

V. Conclusion

We are proposing JD-NMF-based VC for oral surgery patients. The overall JD-NMF process can be divided into two phases: offline and online. In the offline phase, the JD-NMF learns a paired source and target dictionary matrix. To ensure the alignment of the bases of the source and target dictionary matrices, the two matrices are jointly learned. In the online phase, the activation matrix is shared by the source and target speakers when performing VC. We evaluated the proposed JD-NMF by using real-world speech data obtained from patients after their oral surgeries. Our experimental results first showed that JD-NMF greatly improved the original speech with a high STOI score. In addition, JD-NMF is significantly more efficient and effective than a conventional exemplar-based NMF VC method. Finally, through quantitative analyses using a spectrogram and speech envelope plots, it was found that the proposed JD-NMF produces clearer spectrograms with a more obvious modulation depth than those of the original speech, converted by conventional exemplar-based NMF. In summary, the contribution of this paper is two-fold. First, we verified the effectiveness of the proposed joint-training criterion for NMF-based VC. Second, we confirmed that JD-NMF can greatly enhance the speech intelligibility of patients who have undergone oral surgery.

In the present study, we confirmed the effectiveness of the proposed JD-NMF method in terms of objective STOI scores and online computational cost. In the future, we plan to undertake the following: (1) Conduct objective recognition tests to further confirm the clinical applicability of the proposed JD-NMF, even though STOI has been verified as being able to accurately predict the speech intelligibility. (2) This study has confirmed the effectiveness of the proposed JD-NMF running on a PC; thus, we plan to implement it either as a standalone electronic device or as an app for a smartphone.

References

[1] K. Mády, *et al.*, "Speech evaluation and swallowing ability after intra-oral cancer," *Clinical Linguistics and Phonetics,* vol. 17, pp. 411-420, 2003.

[2] G. Rentschler and M. Mann, "The effects of glossectomy on intelligibility of speech and oral perceptual discrimination," *Journal of Oral Surgery,* vol. 38, pp. 348-354, 1980.

[3] B. R. Pauloski, *et al.*, "Speech and swallowing function after anterior tongue and floor of mouth resection with distal flap reconstruction," *Journal of Speech, Language, and Hearing Research,* vol. 36, pp. 267-276, 1993.

[4] R. Aihara, *et al.*, "Consonant enhancement for articulation disorders based on non-negative matrix factorization," in *Proc. APSIPA* 2012, pp. 1-4.

[5] R. Aihara, *et al.*, "Individuality-preserving voice conversion for articulation disorders based on non-negative matrix factorization," in *Proc. ICASSP*, 2013, pp. 8037-8040.

[6] T. Toda, *et al.*, "Voice conversion for various types of body transmitted speech," in *Proc. ICASSP*, 2009, pp. 3601-3604.

[7] K. Nakamura, *et al.*, "Speaking aid system for total laryngectomees using voice conversion of body transmitted artificial speech," in *Proc. INTERSPEECH*, 2006, pp. pp. 1395–1398.

[8] Y.-T. Liu, *et al.*, "Nonnegative matrix factorization-based frequency lowering technology for mandarin-speaking hearing aid users," in *Proc. ICASSP*, 2016.

[9] T. Toda, *et al.*, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 15, pp. 2222-2235, 2007.

[10] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. ICASSP*, 1998, pp. 285-288.

[11] H.-T. Hwang, *et al.*, "A study of mutual information for GMM-based spectral conversion," in *Proc. Interspeech*, 2012, pp. 78-81.

[12] H.-T. Hwang, *et al.*, "Incorporating global variance in the training phase of GMM-based voice conversion," in *Proc. APSIPA*, 2013, pp. 1-6.

[13] M. Narendranath, *et al.*, "Transformation of formants for voice conversion using artificial neural networks," *Speech Communication,* vol. 16, pp. 207-216, 1995.

[14] S. Desai, *et al.*, "Spectral mapping using artificial neural networks for voice conversion," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 18, pp. 954-964, 2010.

[15] F.-L. Xie, *et al.*, "Sequence error (SE) minimization training of neural network for voice conversion," in *Proc. Interspeech*, 2014, pp. 2283-2287.

[16] H.-T. Hwang, *et al.*, "A probabilistic interpretation for artificial neural network-based voice conversion," in *Proc. APSIPA*, 2015.

[17] S. H. Mohammadi and A. Kain, "Voice conversion using deep neural networks with speaker-independent pre-training," in *IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 19-23.

[18] L. Sun, *et al.*, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Proc. ICASSP*, 2015, pp. 4869-4873.

[19] S. H. Mohammadi and A. Kain, "Semi-supervised training of a voice conversion mapping function using a joint-autoencoder," in *Proc. Interspeech*, 2015.

[20] M. Dong, *et al.*, "Mapping frames with DNN-HMM recognizer for non-parallel voice conversion," in *Proc. APSIPA*, 2015, pp. 488-494.

[21] Z. Wu, *et al.*, "Exemplar-based voice conversion using non-negative spectrogram deconvolution," in *8th ISCA Speech Synthesis Workshop*, 2013, pp. 201-206.

[22] Z. Wu, *et al.*, "Joint nonnegative matrix factorization for exemplar-based voice conversion," in *Proc. Interspeech*, 2014.

[23] Z. Wu, *et al.*, "Exemplar-based sparse representation with residual compensation for voice conversion," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 22, pp. 1506-1521, 2014.

[24] K. Masaka, *et al.*, "Multimodal voice conversion using non-negative matrix factorization in noisy environments," in *Proc. ICASSP*, 2014, pp. 1542-1546.

[25] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature,* vol. 401, pp. 788-791, 1999.

[26] Z. Wu, *et al.*, "Exemplar-based voice conversion using joint nonnegative matrix factorization," *Multimedia Tools and Applications,* vol. 74, pp. 9943-9958, 2015.

[27] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. NIPS*, 2001, pp. 556-562.

[28] M. Müller, "Dynamic time warping," *Information retrieval for music and motion,* pp. 69-84, 2007.

[29] P. O. Hoyer, "Non-negative sparse coding," in *IEEE Workshop on Neural Networks for Signal Processing*, 2002, pp. 557-565.

[30] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *The Journal of Machine Learning Research,* vol. 5, pp. 1457-1469, 2004.

[31] R. Peharz and F. Pernkopf, "Sparse nonnegative matrix factorization with ℓ 0-constraints," *Neurocomputing,* vol. 80, pp. 38-46, 2012.

[32] A. Cichocki, *et al.*, "Csiszar's divergences for non-negative matrix factorization: Family of new algorithms," in *Independent Component Analysis and Blind Signal Separation*, ed: Springer, 2006, pp. 32-39.

[33] J. F. Gemmeke, *et al.*, "Exemplar-based sparse representations for noise robust automatic speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 19, pp. 2067-2080, 2011.

[34] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 18, pp. 550-563, 2010.

[35] K. W. Wilson, *et al.*, "Speech denoising using nonnegative matrix factorization with priors," in *Proc. ICASSP*, 2008, pp. 4029-4032.

[36] N. Mohammadiha, *et al.*, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 21, pp. 2140-2151, 2013.

[37] H.-T. Fan, *et al.*, "Speech enhancement using segmental nonnegative matrix factorization," in *Proc. ICASSP*, 2014, pp. 4483-4487.

[38] C. H. Taal, *et al.*, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. ICASSP*, 2010, pp. 4214-4217.

[39] C. H. Taal, *et al.*, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 19, pp. 2125-2136, 2011.

[40] C. Févotte, *et al.*, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural computation,* vol. 21, pp. 793-830, 2009.

[41] P. Sajda, *et al.*, "Recovery of constituent spectra using non-negative matrix factorization," in *Optical Science and Technology, SPIE's 48th Annual Meeting*, 2003, pp. 321-331.

[42] L. Prechelt, "Early stopping-but when?," in *Neural Networks: Tricks of the trade*, ed: Springer, 1998, pp. 55-69.

[43] J. L. Flanagan, *Speech analysis synthesis and perception* vol. 3: Springer Science & Business Media, 2013.

[44] S. Haykin, *Advances in spectrum analysis and array processing* vol. 3: Prentice-Hall, Inc., 1995.

[45] R. van Hoesel, *et al.*, "Amplitude-mapping effects on speech intelligibility with unilateral and bilateral cochlear implants," *Ear and hearing,* vol. 26, pp. 381-388, 2005.

[46] Y.-H. Lai, *et al.*, "Effects of Adaptation Rate and Noise Suppression on the Intelligibility of Compressed-Envelope Based Speech," *PloS one,* vol. 10, p. e0133519, 2015.

[47] A. N. S. Institute, *American National Standard: Methods for Calculation of the Speech Intelligibility Index*: Acoustical Society of America, 1997.