IMPROVING SINGING AID SYSTEM FOR LARYNGECTOMEES WITH STATISTICAL VOICE CONVERSION AND VAE-SPACE

Li Li[1], Tomoki Toda[2], Kazuho Morikawa[2], Kazuhiro Kobayashi[2], Shoji Makino[1]
[1] University of Tsukuba, Japan [2] Nagoya University, Japan
lili@mmlab.cs.tsukuba.ac.jp, tomoki@icts.nagoya-u.ac.jp

ABSTRACT

This paper proposes an improved singing aid system for laryngectomees that converts electrolaryngeal (EL) speech produced using an electrolarynx to a more naturally sounding singing voice. Although the previously proposed system employing a noise suppression process and a rule-based pitch control approach has achieved preliminary success in converting EL speech into a singing voice, there are still two major limitations. First, the converted singing voice still sounds mechanical and unnatural owing to the adverse impacts of spectrograms extracted from EL speeches, also making the effect of pitch control limited. Second, the capability and flexibility of the rule-based pitch control in modeling various singing styles are insufficient, causing the converted singing voices to lack variety. To address these limitations, this paper proposes an improved system that uses 1) a statistical voice conversion approach to convert spectrograms extracted from EL speeches into those of natural speeches and 2) a deep generative model-based approach called VAE-SPACE for pitch modification, which generates pitch patterns in a data-driven manner instead of following manually designed rules. The experimental results revealed that 1) the conversion of spectrograms was effective in improving the naturalness of singing voices, and 2) the statistical pitch control approach was able to achieve comparable results with the rule-based approach, which was very carefully designed to be specialized in singing.

1. INTRODUCTION

The voice is an essential tool used by most of people to communicate with others or express themselves. However, it is difficult for laryngectomees whose larynxes have been removed in surgery to speak or sing in a common manner since they are unable to generate glottal excitation sounds owing to the loss of their vocal folds. In consequence, this vocal disorder may significantly degrade the quality of life of laryngectomees. One popular approach that enables laryngectomees to speak again is to use an external medical device called *electrolarynx* to produce intense mechanical vibrations as an alternative to glottal excitation sounds. Electrolaryngeal (EL) speech produced using an electrolarynx is noteworthy for its intelligibility, and furthermore, it is easy for laryngectomees to learn how to use an electrolarynx, even for people with low physical fitness. However, the perceived naturalness of EL speech is unsatisfactory owing to the use of mechanically generated source excitation sounds, the fundamental frequency (F_0) contours of which are usually flat or given as predetermined patterns. This further limits the capability of the electrolarynx to assist laryngectomees in singing, where F_0 contours play an important role in providing both melodic information and details related to the naturalness and singing style [1].

To develop singing aid systems for laryngectomees, it is essential to suitably control the pitch of EL speech, i.e., F_0 contours. One existing approach is to set F_0 contours corresponding to melodies of predetermined songs and embed them in advance into the electrolarynx as a function to allowing these songs to be sung. However, the flexibility in singing with this approach is unsatisfactory because the number of embedded songs is limited and laryngectomees are solely allowed to sing in predetermined styles.

To achieve a more flexible singing aid, a system based on pitch control has recently been proposed [2]. With this system, laryngectomees are allowed to freely control melodic information such as musical scores and tempo by playing a musical instrument themselves while singing with an electrolarynx. Singing voices are then generated by applying a voice conversion approach that converts EL speeches into singing voices containing well-sung F_0 contours that are modified from the inputted musical scores according to a set of manually predefined rules [3, 4]. Furthermore, noise suppression [5] is employed to reduce the source excitation signals emitted from the electrolarynx. Although this system has achieved preliminary success as a singing aid, there are still two limitations. First, the effect of pitch control in improving the naturalness of singing voices was limited because of the fluctuations originating from the spectral features extracted from EL speeches [2], which resulted in the converted singing voices still sounding mechanical and unnatural. Second, both the capability and flexibility of the rule-based pitch control approach in modeling various singing styles are insufficient. Once the rules are determined, the system outputs certain F_0 pat-

terns containing similar characteristics without considering the personalities and emotions of singers, which is an undesirable property of singing aid systems.

To develop a system that is capable of aiding laryngectomees to sing naturally and distinctly, this paper proposes an improved system that uses 1) a statistical voice conversion (VC) approach based on the Gaussian mixture model (GMM) [6, 7] to convert spectral features extracted from EL speeches into those of natural speeches to alleviate the fluctuation problem and 2) a deep generative model-based approach called VAE-SPACE [8] to generate F_0 contours of singing voices from input musical scores. As a data-driven approach, it is expected that VAE-SPACE can learn the rules for generating natural F_0 contours from data automatically, making it possible to model different singing styles and expressions with a unified model.

2. OVERALL FRAMEWORK OF SINGING AID SYSTEM FOR LARYNGECTOMEES

Fig. 1 shows an overview of the conventional singing aid system proposed in [2] that takes a sequence of musical scores $\boldsymbol{N} = [n_1, \ldots, n_t, \ldots, n_T]$ provided by playing an instrument as melodic information in addition to an EL speech $\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_t, \ldots, \boldsymbol{s}_T]$. Here, $\boldsymbol{s}_t = [s_t(1), \ldots, s_t(f), \ldots, s_t(F)]^\mathsf{T}$ denotes the short-time Fourier transform (STFT) coefficients of the EL speech at frame t, and f and $(\cdot)^\mathsf{T}$ denote the frequency index and transpose operator, respectively. The system mainly consists of three modules, namely, for voice quality enhancement, pitch control, and synthesis. With this system, singing voices are generated by a vocoder-based synthesis approach [9] that takes phonetic information and pitch information as inputs, where the former is extracted from the EL speech $\boldsymbol{S}$ enhanced by the voice quality enhancement module and the latter is obtained by modifying the input musical scores $\boldsymbol{N}$ via the pitch control module.

Note that this system can also serve laryngectomees who do not play instruments by allowing them to sing with played accompaniments, where a sequence of predetermined musical scores is given in synchronization with the accompaniments. Different from the method of embedding preset F_0 patterns into an electrolarynx and controlling the pitch by pushing a button, this system can obtain more natural singing voices since singing voices obtained with the former method are usually interrupted when the musical score changes owing to the limitation of mechanical excitation generation, and those obtained in the latter way are converted from more fluent EL speeches.

3. CONVENTIONAL SYSTEM WITH NOISE SUPPRESSION AND RULE-BASED PITCH CONTROL

3.1 Noise suppression

It is important to enhance the quality of both phonetic information and pitch information to achieve a better transformation. To obtain correct phonetic information from

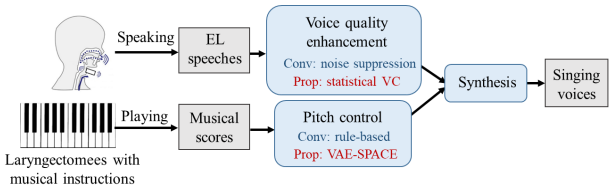

Figure 1. Flowchart of singing aid system.

an EL speech, which is usually mixed with a noisy source signal that radiates from the position of the EL attachment, the aforementioned system employs a spectral subtraction (SS) method [5] to enhance the voice quality of the recorded EL speech. A prototype noise amplitude spectrum $|l(f)|$ is calculated by averaging the amplitude spectra of the EL noise recorded with a close-talking microphone in advance. The enhanced EL speech is obtained with the enhanced amplitude spectrum $|\hat{s}_t(f)|$ and the noisy phase spectrum, where

$$|\hat{s}_t(f)| = \begin{cases} |s_t(f)| - 2|l(f)|, & (|s_t(f)| > 2|l(f)|), \\ 0, & (\text{otherwise}). \end{cases} \quad (1)$$

The phonetic information used for synthesis, i.e., spectral features and aperiodic components, is obtained by analyzing the enhanced EL speech with fixing F_0 at a constant value and using the on/off information of the electrolarynx as unvoiced/voiced information.

3.2 Rule-based pitch control

For pitch control, a rule-based F_0 modification technique [3,4] is applied to add overshoot, vibrato, preparation, and fine fluctuation, which are four characteristics typically observed in F_0 contours of natural singing voices, into the musical scores n_t. Specifically, overshoot, vibrato, and preparation are added by applying the following infinite impulse response filter to the musical scores:

$$H(s) = \frac{k}{s^2 + 2\zeta\omega s + \omega^2}, \quad (2)$$

where ω, ζ, and k denote the natural frequency, damping coefficient, and proportional gain, respectively. Overshoot and preparation are expressed as the second-order damping model ($0 < |\zeta| < 1$), while vibrato is expressed as the second-order oscillation model ($|\zeta| = 0$). The fine fluctuation is generated from white noise that is high-pass-filtered with the cutoff frequency set at 10 Hz followed by a normalization. The modified F_0 contour can be expressed as $o_t = n_t + e_t$, where o_t and e_t denote the generated F_0 and the component including all four characteristics that is finally added to the musical score at frame t, respectively.

3.3 Limitations

The effectiveness of this system in converting EL speech into a singing voice was experimentally confirmed in [2]. However, it was also reported that undesirable fluctuations reside in F_0 contours of synthesized singing voices that

may have originated from the spectral features extracted from the enhanced EL speeches, which cause the singing voices to still sound mechanical and unnatural. The upper figure in Fig. 6 shows an example of the reanalyzed F_0 contour of a singing voice obtained with the system. Another limitation originates from the rule-based pitch control. Although the system allows laryngectomees to sing an arbitrary song with the desired melody, it is difficult for this framework to further improve the capability to express various singing styles or to generate expressive singing voices.

4. PROPOSED SYSTEM WITH STATISTICAL VC AND VAE-SPACE

To remove these indefinite spectral components affecting F_0 contours, one of the promising approaches is to transform the spectral features of EL speeches of songs into those of natural singing voices not containing these components. Statistical VC techniques [6, 7] have the potential to be used for developing such a transformation based on training data consisting of utterance pairs of the source and target voices, namely, singing voices sung using an electrolarynx (EL speeches) and in a natural way. Furthermore, it is expected that EL noise can be reduced together by training a model with the source voice being noisy EL speech.

To address the second limitation, motivated by the high flexibility of a statistical approach in modeling different voice characteristics and singing styles [10–12], we propose using a statistical parametric model for pitch control instead of the rule-based approach. Specifically, we employ VAE-SPACE [8], which uses a variational autoencoder (VAE) as an analysis-synthesis model to discover the structure of an F_0 generating process for the singing voice in a data-driven manner as well as an inverse process for estimating the underlying musical scores.

4.1 Statistical VC for converting EL speech into singing voice

Let $\boldsymbol{x}_t = [x_t(1), \ldots, x_t(d), \ldots, x_t(D)]^\mathsf{T}$ denotes the D-dimensional spectral feature extracted from EL speech $\boldsymbol{s}_t$ at frame t, where d denotes the index of the feature dimension. The aim of VC is to estimate the spectral features, F_0 contours including unvoiced/voiced (U/V) information, and aperiodic components of the corresponding natural singing voice, which are denoted by the same variable $\boldsymbol{y}_t = [y_t(1), \ldots, y_t(d), \ldots, y_t(D)]^\mathsf{T}$ for simplicity, from the noisy spectral sequence $\boldsymbol{x}_t$.

In the training step, a joint probability density function (p.d.f.) of the joint acoustic feature vectors $[\boldsymbol{X}_t^\mathsf{T}, \boldsymbol{Y}_t^\mathsf{T}]^\mathsf{T}$ is modeled with a GMM as follows:

$$P\big(\boldsymbol{X}_t, \boldsymbol{Y}_t \big| \boldsymbol{\Theta}^{(\mathrm{X,Y})}\big) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\big([\boldsymbol{X}_t^\mathsf{T}, \boldsymbol{Y}_t^\mathsf{T}]^\mathsf{T}; \boldsymbol{\mu}_m^{(\mathrm{X,Y})}, \boldsymbol{\Sigma}_m^{(\mathrm{X,Y})}\big). \quad (3)$$

Here $\boldsymbol{X}_t$ denotes spectral segment feature vectors that are obtained by performing principal component analysis (PCA) for the joint vectors concatenating the spectral feature vectors of the current frame, preceding L frames, and succedding L frames extracted from source voices. $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\mathsf{T}, \Delta\boldsymbol{y}_t^\mathsf{T}]^\mathsf{T}$ denotes vectors combining static and dynamic features extracted from target voices. $\boldsymbol{\Theta}^{(\mathrm{X,Y})}$ denotes a parameter set of the GMM, which consists of the weights α_m, mean vectors $\boldsymbol{\mu}_m^{(\mathrm{X,Y})}$, and convariance matrices $\boldsymbol{\Sigma}_m^{(\mathrm{X,Y})}$ of all the mixture components. Moreover, the p.d.f. of the global variance (GV) [13] of the target static feacture vectors over an utterance $\boldsymbol{v}(\boldsymbol{y}) = [v(1), \ldots, v(D)]^\mathsf{T}$ is also modeled with a Gaussian distribution, which is expressed with a set of parameters $\boldsymbol{\Theta}^{(v)} = \{\boldsymbol{\mu}^{(v)}, \boldsymbol{\Sigma}^{(v)}\}$ as

$$P\big(\boldsymbol{v}(\boldsymbol{y}) \big| \boldsymbol{\Theta}^{(\boldsymbol{v})}\big) = \mathcal{N}\big(\boldsymbol{v}(\boldsymbol{y}); \boldsymbol{\mu}^{(v)}, \boldsymbol{\Sigma}^{(v)}\big). \quad (4)$$

Here, the GV $\boldsymbol{v}(\boldsymbol{y})$ of a time sequence of the target static feature $\boldsymbol{y} = [\boldsymbol{y}_1^\mathsf{T}, \ldots, \boldsymbol{y}_T^\mathsf{T}]^\mathsf{T}$ is calculated utterance by utterance as

$$v(d) = \frac{1}{T}\sum_{t=1}^{T}\big(y_t(d) - \frac{1}{T}\sum_{\tau=1}^{T} y_\tau(d)\big)^2. \quad (5)$$

In the conversion process, a time sequence vector of the converted static feature vectors $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\mathsf{T}, \ldots, \hat{\boldsymbol{y}}_T^\mathsf{T}]^\mathsf{T}$ is determined by maximizing the product of the conditional p.d.f. of $\boldsymbol{Y}$ given $\boldsymbol{X}$ and the p.d.f. of the GV as

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\ P\big(\boldsymbol{Y}\big|\boldsymbol{X}, \boldsymbol{\Theta}^{(\mathrm{X,Y})}\big) P\big(\boldsymbol{v}(\boldsymbol{y})\big|\boldsymbol{\Theta}^{(v)}\big)^\lambda, \quad (6)$$

$$\text{subject to } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}, \quad (7)$$

where $\boldsymbol{X} = [\boldsymbol{X}_1^\mathsf{T}, \ldots, \boldsymbol{X}_T^\mathsf{T}]^\mathsf{T}$ and $\boldsymbol{Y} = [\boldsymbol{Y}_1^\mathsf{T}, \ldots, \boldsymbol{Y}_T^\mathsf{T}]^\mathsf{T}$ are time sequence vectors of the source and target feature vectors, respectively. $\boldsymbol{W}$ denotes a $2DT$-by-DT matrix that extends a time sequence vector of the static feature vectors into that of the joint static and dynamic feature vectors [14], and λ is a weight parameter, which is commonly set to $2T$. By adopting an approximation with a suboptimum mixture component sequence $\boldsymbol{m} = \{m_1, \ldots, m_T\}$, the converted static feature vector sequence is determined as follow [7]:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P\big(\boldsymbol{Y}\big|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\Theta}^{(\mathrm{X,Y})}\big) P\big(\boldsymbol{v}(\boldsymbol{y})\big|\boldsymbol{\Theta}^{(v)}\big)^\lambda. \quad (8)$$

The enhanced EL speech is generated by filtering the mixed excitation signal, which is designed according to the F_0 values, U/V information, and aperiodic components estimated from the spectral segment feature vectors, with the converted spectral features.

4.2 VAE-SPACE for pitch control

VAE-SPACE [8] has been proposed as a generative model that can represent and generate F_0 contours of both speeches and singing voices. Let $\boldsymbol{z}$ denotes a sequence of parameters governing the generating process of F_0 contours $\boldsymbol{o} = [o_1, \ldots, o_T]^\mathsf{T}$. VAE-SPACE uses an encoder to estimate the parameters of a conditional distribution $q_\phi(\boldsymbol{z}|\boldsymbol{o})$ of the latent variable $\boldsymbol{z}$ given an F_0 contour $\boldsymbol{o}$, and a decoder to estimate the parameters of a conditional distribution $p_\theta(\boldsymbol{o}|\boldsymbol{z})$ of the F_0 contour $\boldsymbol{o}$ given the latent variable $\boldsymbol{z}$. The encoder and decoder are trained simultaneously so that $q_\phi(\boldsymbol{z}|\boldsymbol{o})$ becomes consistent with the true

posterior distribution $p_\theta(\boldsymbol{z}|\boldsymbol{o}) \propto p_\theta(\boldsymbol{o}|\boldsymbol{z})p(\boldsymbol{z})$. The parameters of the networks ϕ and θ can be trained by maximizing the following variational lower bound [15]:

$$\mathcal{L}(\theta, \phi; \boldsymbol{o}) = \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{o})}[\log p_\theta(\boldsymbol{o}|\boldsymbol{z})] - D_{KL}[q_\phi(\boldsymbol{z}|\boldsymbol{o})||p(\boldsymbol{z})], \quad (9)$$

where $D_{KL}[\cdot||\cdot]$ denotes Kullback-Leibler (KL) divergence. In VAE-SPACE, the latent variable $\boldsymbol{z}$ is associated with a set of interpretable parameters so that the decoder can be seen as a generative model for synthesizing F_0 contours and the encoder can be seen as an inverse problem solver that analyzes the underlying parameters of an observed F_0 contour. In the case of speech, $\boldsymbol{z}$ is associated with a phrase/accent command sequence defined in the Fujisaki model [16], while in the case of a singing voice, it is associated with a sequence of musical scores. The name "VAE-SPACE" comes from the former case, where the VAE-based method is designed to perform statistical phrase/accent component estimation (SPACE).

A typical way of modeling $q_\phi(\boldsymbol{z}|\boldsymbol{o})$ and $p_\theta(\boldsymbol{o}|\boldsymbol{z})$ is to assume a Gaussian distribution

$$q_\phi(\boldsymbol{z}|\boldsymbol{o}) = \mathcal{N}\big(\boldsymbol{z}|\boldsymbol{\mu}_\phi(\boldsymbol{o}), \operatorname{diag}\boldsymbol{\sigma}^2_\phi(\boldsymbol{o})\big), \quad (10)$$

$$p_\theta(\boldsymbol{o}|\boldsymbol{z}) = \mathcal{N}\big(\boldsymbol{o}|\boldsymbol{\mu}_\theta(\boldsymbol{z}), \operatorname{diag}\boldsymbol{\sigma}^2_\theta(\boldsymbol{z})\big), \quad (11)$$

where $\boldsymbol{\mu}_\phi(\boldsymbol{o})$, $\boldsymbol{\sigma}^2_\phi(\boldsymbol{o})$ are the encoder outputs and $\boldsymbol{\mu}_\theta(\boldsymbol{z})$, $\boldsymbol{\sigma}^2_\theta(\boldsymbol{z})$ are the decoder outputs. While the prior distribution $p(\boldsymbol{z})$ is typically modeled as a standard Gaussian distribution with zero mean and unit variance, VAE-SPACE designs it as a specific form based on the assumption that $\boldsymbol{z}$ indicates the underlying musical scores of an F_0 contour in the case of a singing voice. Specifically, in a supervised setting where pairs of F_0 contours and musical scores are availiable, we can train the VAE by maximizing the following loss function, whici tends to maximize the likelihood of $\boldsymbol{z}$, instead of minimizing the KL divergence since the prior distribution of $\boldsymbol{z}$ is known:

$$\mathcal{L}(\theta, \phi; \boldsymbol{o}) = \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{o})}[\log p_\theta(\boldsymbol{o}|\boldsymbol{z})] + \mathbb{E}_{(\boldsymbol{o},\boldsymbol{z})\sim p_D(\boldsymbol{o},\boldsymbol{z})}[\log q_\phi(\boldsymbol{z}|\boldsymbol{o})], \quad (12)$$

where $\mathbb{E}_{(\boldsymbol{o},\boldsymbol{z})\sim p_D(\boldsymbol{o},\boldsymbol{z})}[\cdot]$ denotes the sample mean over the training data. In the generation process, a $\boldsymbol{z}$ sampled from a Gaussian distribution with the musical scores $\boldsymbol{N}$ as the mean and a variance matrix with a small constant value in the diagonal is used as the input of the trained decoder. The generated F_0 contour is then used to replace that obtained by VC to generate the excitation signal with the U/V information estimated by VC.

For network architectures, a gated convolutional neural network (CNN) [17] is used to construct the encoder and decoder to capture long- and short-term dependencies in F_0 contours. The gated CNN uses a data-driven gate called gated linear unit (GLU) $\sigma(\mathbf{H}_{l-1} * \mathbf{W}^{\mathrm{g}}_l + \mathbf{b}^{\mathrm{g}}_l)$ as a nonlinear activation function to control the information passed on in the hierarchy, where $\mathbf{H}_{l-1}$ denotes the output of the $(l-1)$-th layer, $\mathbf{b}^{\mathrm{f}}_l$ and $\mathbf{b}^{\mathrm{g}}_l$ are the weight and bias parameters of the l-th layer and σ is the sigmoid function.

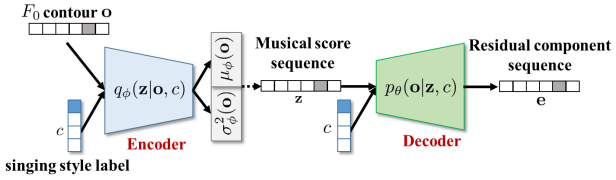

Figure 2. An overview of conditional VAE-SPACE.

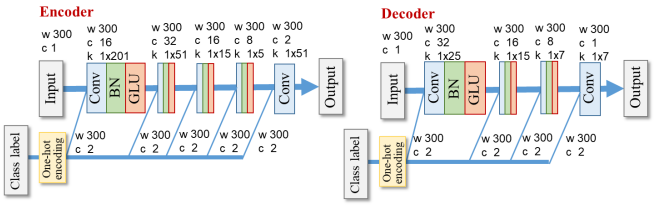

Figure 3. Network architectures of encoder and decoder used for cVAE. VAE used the same architecture excluding the class label inputs. "w", "c" and "k" denote the width, channel number and kernel size, respectively. "Conv", "BN" and "GLU" denote 1D convolution, batch normalization and gated linear unit, respectively.

To improve the performance of VAE-SPACE in modeling F_0 contours and minimize the cost of preparing pair data, we investigate two specific implementation-level problems in this paper: 1) whether a score-level alignment is necessary to train a supervised VAE-SPACE and 2) whether the performance can be improved by fine-tuning the trained decoder with the real musical score sequences generated by a sampling process during VAE pretraining. Furthermore, aiming to model different singing styles in a controllable manner, we also attempt to adopt a conditional CNN with GLUs to construct networks, which takes labels of singing styles c represented as one-hot vectors as additional inputs. The criterion of training a conditional VAE (cVAE) can be obtained merely by extending (12) [15].

5. EXPERIMENTAL EVALUATIONS

5.1 Experimental conditions

We prepared two datasets containing different pairs of data. *Dataset1* consists of EL speech samples of 21 Japanese children's songs recorded by a laryngeal speaker using an electrolarynx and the corresponding natural singing voices. 17 songs were manually segmented into 157 short phrases and used to train GMMs, and the other 4 songs were segmented into 19 phrases, and used as a test set. Each phrase was about 3∼8s long. *Dataset2* includes two versions of one Japanese song (about 4min 30s long), which were sung by a female person in normal and expressive singing styles. We segmented each song into 53 phrases and manually took alignments at the phrase and score levels, which were referred to as "unaligned" and "aligned" data, respectively.

We used the amplitude spectra as spectral feature vectors of EL speeches, and STRAIGHT analysis [9] to extract acoustic features of normal singing voices. The shift

Table 1. RMSE and standard deviation of VAE with different implementation conditions. "ft" denotes fine-tuning.

Models	RMSE
VAE-unaligned	26.1031 ± 5.5028
VAE-aligned	22.7670 ± 5.2874
VAE-aligned-ft	**21.3409 ± 5.3566**
cVAE-aligned-ft	22.2933 ± 6.3934

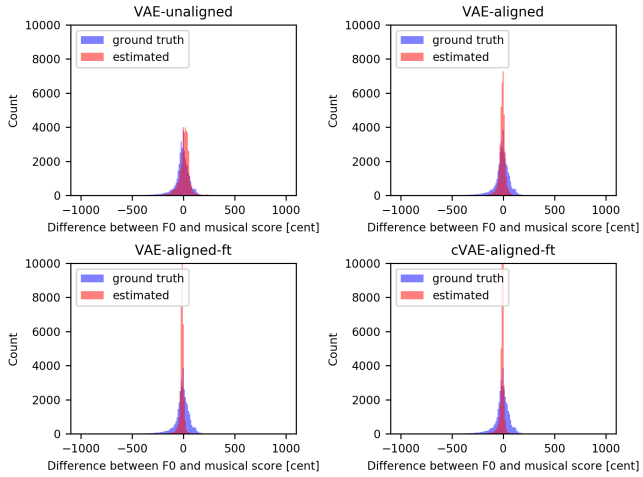

Figure 4. Histograms of the residual components e_t.

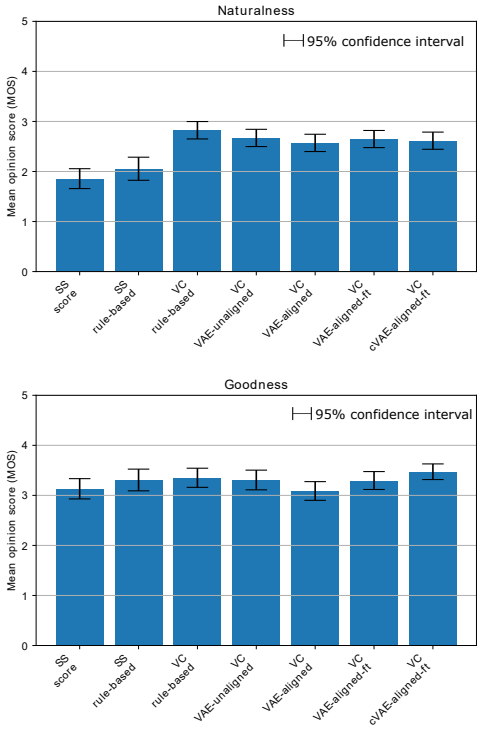

Figure 5. MOS in terms of naturalness and goodness.

length was 5 ms. The 0th through 24th mel-cepstral coefficients were used as spectral features of normal singing voices. As excitation features, a log-scaled F_0 value and aperiodic components on five frequency bands (i.e., 0-1, 1-2, 2-4, 4-6, and 6-8 kHz) were used. To obtain segmental feature vectors at each frame, we concatenated spectral feature vectors with the adjacent frames by setting $L = 4$ and performed PCA to reduce the feature dimension to 50. The numbers of mixture components of the GMMs used to estimate spectral features, aperiodic components, and F_0 including U/V information were all set at 16. Following the original VAE-SPACE paper, the output of the decoder was designed to be a sequence of residual components $\boldsymbol{e} = [e_1, \ldots, e_T]$, as shown in Fig. 2. Fig. 3 shows the architectures of the encoder and decoder. To train VAE, we used the songs sung in a normal style in *Dataset2*, and those sung in an expressive style were used as additional data for training the cVAE.

5.2 Objective evaluation

We first conducted an objective evaluation to demonstrate the performances of VAE-SPACE with different implementation conditions. We divided the songs into 7 folds and performed cross-validation. Root mean square error (RMSE) between the estimated and target F_0 contours was used as a metric to evaluate the estimation accuracy of F_0 generation. Table 1 shows the results. The objective results show that applying a score-level alignment and fine-tuning the decoder were effective in improving the accuracy of generating the target F_0 contours. However, the histograms of the residual components e_t shown in Fig. 4 reveal that both the alignment and fine-tuning decreased the dynamics of the generated F_0 contours.

Table 2. p-values calculated for method c).

	naturalness	goodness
SS+score	2.25E-11	0.1257
SS+rule-based	5.64E-07	0.7673
VC+VAE-unaligned	0.2223	0.7551
VC+VAE-aligned	0.0460	0.0561
VC+VAE-aligned-ft	0.1677	0.3413
VC+cVAE-aligned-ft	0.0976	0.3413

5.3 Subjective evaluation

To further investigate the performances of VAE-SPACE with different implementations and demonstrate the effectiveness of VC, we conducted a subjective evaluation that compared 7 methods, namely, a) SS+musical scores (SS+score), b) SS+rule-based F_0 modification (SS+rule-based), c) VC+rule-based F_0 modification (VC+rule-based), d) VC+VAE-SPACE using unaligned data (VC+VAE-unaligned), e) VC+VAE-SPACE using aligned data (VC+VAE-aligned), f) VC+VAE-SPACE using aligned data and fine-tuning (VC+VAE-aligned-ft), g) VC+cVAE-SPACE using aligned data and fine-tuning (VC+cVAE-aligned-ft). 13 evaluators participating in the experiments scored the converted singing voices in terms of the naturalness of the song and the goodness of singing using a 5-point opinion scale. The mean and 95% confidence interval of the two criteria are shown in Fig. 5 and the p-values calculated for method c) are shown in Table 2.

The results show that VC significantly improved the

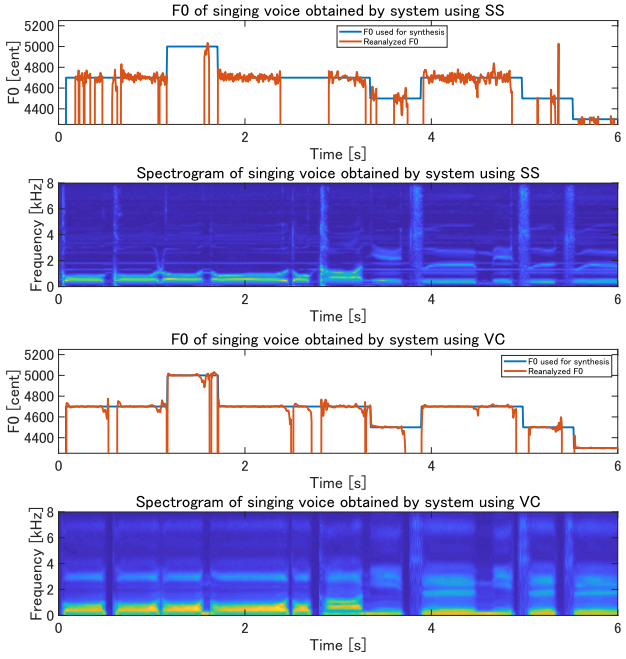

Figure 6. Examples of reanalyzed F_0 contours and spectrograms of synthesized singing voices obtained by employing SS (upper) and VC (bottom).

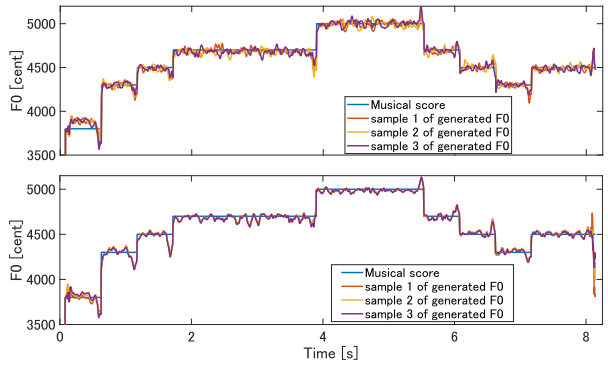

Figure 7. Generated F_0 contours with various $\boldsymbol{z}$ (upper) and class labels (bottom).

quality of converted singing voices in terms of the naturalness, which confirmed the effectiveness of the VC approach in removing the undesirable spectral components. An example of the reanalyzed F_0 contour and spectrogram is shown at the bottom of Fig. 6, which also confirmed this result. VAE-SPACE implemented with a conditional CNN achieved comparable results to the rule-based pitch control with carefully designed rules, and showed a high potential for contributing to the system. Compared with the method using unaligned data and the cVAE, the subjective results for the method using aligned data and fine-tuning were slightly lower, exhibiting the same tendency as the results shown in Fig. 4. This suggests that alignment and fine-tuning are not important in our system. We also investigated the variety of the F_0 contours generated with different sampled $\boldsymbol{z}$ and style class labels, the results of which are shown in Fig. 7. We observed some reasonable differences between the F_0 contours generated with various sampled $\boldsymbol{z}$. However, there was no notable difference observed between the F_0 contours generated with different class labels, which may have been due to the high ability of the networks to model the conditional distributions while ignoring the class labels [18, 19]. This issue will be addressed in future work.

6. DISCUSSION AND REUSABLE INSIGHTS

From the above results, it is concluded that spectrogram modification is useful and must be considered as well as F_0 control to improve the quality of singing voices. On the other hand, data alignment and fine-tuning are not essential, which means that we can increase the number of pair data at a relatively low cost to improve the estimation accuracy and the variety of styles. Furthermore, since the amount of pair data required for the VC approach is small and VAE-SPACE allows semi-supervised training with unlabeled data in addition to labeled data [8,15], it is expected that the entire system will be allowed to play its potential data efficiently, which is important for such a data-driven framework. Although cVAE-based implementation failed to represent different styles in the experiments, the framework for modeling various styles with a unified model provides us with a simple and straightforward way to control singing styles and apply style interpolation/morphing. Note that although we applied the method to model singing voices, it can also be used with other audio signals such as to generate suitable F_0 contours of musical instruments from musical instrument digital interface (MIDI) information. In addition to modeling various styles, we can extend this method to generate the F_0 contours of multiple instruments.

7. CONCLUSIONS

This paper proposed an improved singing aid system for laryngectomees based on a previously proposed system that converts EL speeches into singing voices according to the additionally inputted melodic information. The proposed system uses a statistical VC approach to transform the phonetic information extracted from EL speeches into those of natural speeches, and VAE-SPACE to perform pitch control. We investigated the importance of well-aligned pair data and the fine-tuning process for improving the performance and a conditional version of VAE-SPACE for modeling multiple singing styles with a unified model. The experimental results demonstrated that 1) the VC approach was effective in significantly improving the naturalness of singing voices, 2) the effectiveness of well-aligned pair data and fine-tuning was limited, and 3) VAE-SPACE was able to achieve comparable results to a carefully designed rule-based approach in generating F_0 contours.

8. ACKNOWLEDGEMENTS

This work was supported by JST, PRESTO Grant Number JPMJPR1657, and JSPS KAKENHI Grant 18J20059.

9. REFERENCES

[1] M. Umbert, J. Bonada, M. Goto, T. Nakano, and J. Sundberg. "Expression control in singing voice synthesis: Features, approaches, evaluation, and challenges," *IEEE Signal Processing Magazine*, Vol. 32, No. 6, pp. 55–73, 2015.

[2] K. Morikawa and T. Toda. "Electrolaryngeal speech modification towards singing aid system for laryngectomees," in Proc. *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 610-–613. 2017.

[3] T. Saitou, M. Unoki, and M. Akagi. "Development of an f0 control model based on f0 dynamic characteristics for singing-voice synthesis," *Speech communication*, Vol. 46, No. 3-4, pp. 405-–417, 2005.

[4] T. Saitou, M. Goto, M. Unoki, and M. Akagi. "Speech-to-singing synthesis system: Vocal conversion from speaking voices to singing voices by controlling acoustic features unique to singing voices," in Proc. *the 10th National Conference on Man-Machine Speech Communication (NCMMSC)*, 2009.

[5] S. Boll. "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 27, No. 2, pp. 113-–120, 1979.

[6] T. Toda, M. Nakagiri, and K. Shikano. "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE Trans. on Audio, Speech, and Language Processing*, Vol. 20, No. 9, pp. 2505-–2517, 2012.

[7] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano. Speaking-aid systems using gmm-based voice conversion for electrolaryngeal speech. Speech Communication, 54(1):134–146, 2012.

[8] K. Tanaka, H. Kameoka, and K. Morikawa. "Vaespace: Deep generative model of voice fundamental frequency contours," in Porc. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, pp. 5779-–5783, 2018.

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

[10] M. Nishimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda. "Singing voice synthesis based on deep neural networks," in Proc. Interspeech, pp. 2478-–2482, 2016.

[11] Y. Ohishi, H. Kameoka, D. Mochihashi, and K. Kashino. "A stochastic model of singing voice f0 contours for characterizing expressive dynamic components," in Proc. *13th Annual Conference of the International Speech Communication Association*, 2012.

[12] K. Saino, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda. "An hmm-based singing voice synthesis system," in Proc. *9th International Conference on Spoken Language Processing*, 2006.

[13] T. Toda, A. W. Black, and K. Tokuda. "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. on Audio, Speech, and Language Processing*, Vol. 15, No. 8, pp. 2222-–2235, 2007.

[14] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. "Speech parameter generation algorithms for HMM-based speech synthesis," in Proc. *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (Cat. No. 00CH37100)*, Vol. 3, pp. 1315-–1318, 2000.

[15] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. "Semi-supervised learning with deep generative models," in *Advances in neural information processing systems*, pp. 3581—3589, 2014.

[16] H. Fujisaki. "A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour," *Vocal physiology: Voice production, mechanisms and functions*, pp. 347–355, 1988.

[17] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier. "Language modeling with gated convolutional networks," in Proc. *the 34th International Conference on Machine Learning* Vol. 70, pp. 933-–941, 2017.

[18] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in neural information processing systems*, pp. 2172-–2180, 2016.

[19] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo. "ACVAE-VC: Non-parallel many-to-many voice conversion with auxiliary classifier variational autoencoder," arXiv preprint *arXiv:1808.05092*, 2018.