A Revisit to Feature Handling for High-quality Voice Conversion Based on Gaussian Mixture Model

Hitoshi Suda*, Gaku Kotani*, Shinnosuke Takamichi†, and Daisuke Saito*
* Graduate School of Engineering, The University of Tokyo, Japan
E-mail: {hitoshi,kotani,dsk_saito}@gavo.t.u-tokyo.ac.jp
† Graduate School of Information Science and Technology, The University of Tokyo, Japan
E-mail: shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp

***Abstract*—This paper discusses influences of handling acoustic features on the quality of generated sounds in voice conversion (VC) systems based on Gaussian mixture models (GMMs). In the context of improving the quality of VC, mapping models, which are used to convert acoustic features, have been widely discussed. Nevertheless, the components other than the mapping models have rarely been studied. The experimental results show that the quality of VC depends on not only the models but also the methods of analysis and synthesis of utterances. This paper also investigates filtering methods for synthesis. In order to avoid buzzy sounds generated from vocoders, differential-spectrum compensation is applied as an alternative method of synthesizing waveforms. Although mel log spectral approximation (MLSA) filtering is traditionally used for differential-spectrum compensation, the experimental results indicate the approximation in MLSA filtering degrades the quality of the synthesized speech. In order to avoid this approximation, this paper introduces an alternative filtering method, which is named SP-WORLD, inspired by the WORLD vocoder framework. The subjective experiments demonstrate that SP-WORLD is comparable to MLSA filtering, and outperforms it in some cases.**

I. Introduction

Voice conversion (VC), or speaker conversion, is a technique to alter an input utterance to make it sound like another speaker's utterance without changing its linguistic contents [1]. VC systems generally consist of three steps: feature extraction, feature conversion, and waveform synthesis. Firstly, acoustic features are extracted from input utterances by analysis. Secondly, the mapping models convert the extracted features. Finally, output utterances are synthesized from the mapped features. There are several mapping methods which provide nonlinear and continuous conversion of features. Methods based on Gaussian mixture models (GMMs) express mapping functions as conditional probability distributions [2], [3]. Exemplar-based methods separate speaker individuality and linguistic information by non-negative matrix factorization (NMF) and conversion can be achieved by the exchange of speaker information [4], [5]. Methods based on neural networks directly estimate nonlinear mapping functions, and the complexity of the frameworks makes the models precise [6], [7]. To train these models, as a rule, source and target speakers' utterances that have the same linguistic contents, or parallel data, are required. Once models are trained, any input features can be converted by the mapping functions derived from the models.

In GMM-based techniques, mixtures of Gaussian components model probability density of joint vectors of source and target feature vectors [3], [8]. The mapping functions are described as conditional probability derived from the GMM, and equivalent to the weighted sum of locally linear transformations. Since GMM-based frameworks are flexible and easy to handle compared with other statistical approaches, several techniques can be additionally applied: speaker adaptation techniques based on maximum likelihood linear regression (MLLR) [9] or maximum a posterior (MAP) adaptation [10], a technique using a target speaker model as prior knowledge in parameter generation [11], et cetra.

The quality of converted utterances depends on not only mapping models but also synthesizers. Traditional VC techniques use vocoders such as STRAIGHT [12] and WORLD [13] for waveform generation. Although these frameworks model voices as precisely as possible, vocoders inevitably degrade the naturalness of the generated utterances because of various errors in excitation modeling, voiced/unvoiced analysis, and so on. In order to avoid these vocoding errors, differential-spectrum compensation is also used to generate waveforms [14]. In contrast to vocoders, the differential-spectrum compensation methods use source waveforms of input utterances directly. Since these methods cannot modify excitations unlike vocoders, another technique must be applied for fundamental frequency (F_0) conversion. Several methods for modifying F_0 are proposed such as the waveform-similarity-based synchronized overlap-add algorithm (WSOLA) [15].

For differential-spectrum compensation, the mel log spectral approximation (MLSA) filtering method is traditionally used. Since this method approximately derives filter coefficients from mel-cepstral coefficients, the approximation possibly degrades the quality of converted utterances. In order to avoid this approximation, inspired by the WORLD vocoder, this paper introduces a new filtering method named SP-WORLD, which uses the minimum phase reconstruction technique. The

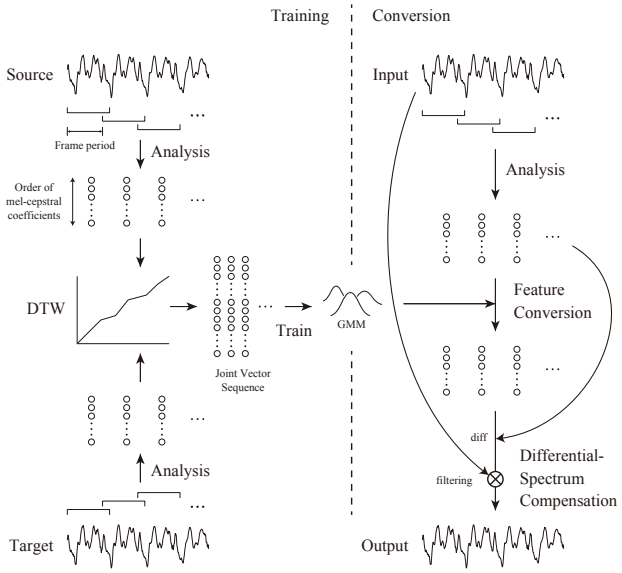


Fig. 1. Overview of the GMM-based VC framework discussed in this paper.

subjective experiments show SP-WORLD is comparable to MLSA filtering, and superior in some conditions.

The quality of synthesized utterances is also determined by parameters of analysis such as frame periods and the order of mel-cepstral coefficients. Although these values are often selected implicitly and have rarely been discussed, the subjective experiments reveal the effectiveness of optimization of these hyperparameters. Since the analysis of utterances is essential for training mapping models, the contribution of the paper spreads to approaches based on not only GMM but also neural networks.

The remainder of this paper is organized as follows. Section II describes the components of VC systems: feature analysis, dynamic time warping algorithm (DTW), GMM-based VC and differential-spectrum compensation. Section III details the several experiments for investigating the methods of analysis and synthesis in GMM-based VC. Finally, Section IV concludes this paper.

II. Main Components of VC Systems

This section describes four main components in VC systems: feature analysis, DTW, GMM-based feature conversion [3], [8], and differential-spectrum compensation [14]. Fig. 1 shows the overview of the system including these components. The discussed VC systems in this paper are based on these techniques.

A. Feature Analysis

For voice conversion, acoustic features are obtained by analysis of utterances. For feature analysis, several hyperparameters must be specified. This paper focuses on two parameters: frame periods and the order of mel-cepstral coefficients.

Frame periods, or frame shift sizes, determine how precisely the sounds are analyzed in the time domain. With shorter frame periods, higher quality re-synthesis is expected to be performed.

In analysis and synthesis systems such as STRAIGHT and WORLD, waveforms are analyzed into three acoustic features: pitch (F_0) information, spectral envelopes, and aperiodic information. Among these features, spectral envelopes play an important role for speaker information, and therefore the target of conversion is generally spectral envelopes in VC. Although spectral envelopes can be obtained as power spectra, the mel-cepstral coefficients are often used as features. The mel-cepstral coefficients are compressive expression of spectral envelopes, and is easier to model than power spectra. The order of mel-cepstral coefficients determines how precisely the spectral envelopes are represented.

B. Dynamic Time Warping Algorithm (DTW)

In order to obtain mapping models, the time alignment of features must be performed. As an alignment method, DTW is often used, which minimizes the mean square error between two feature vector sequences. That is, the difference in the mel-cepstral coefficients is the criterion for alignment. Since mel-cepstral coefficients include not only linguistic information but also speaker individuality, the difference in speakers degrades alignment performance. To overcome this defect, this paper introduces an improved algorithm named affine-DTW. Affine-DTW iterates following three steps: performing general DTW, estimation of transformation matrix, and affine transformation of source features. Since affine transformation is equivalent to GMM-based conversion with one Gaussian component, affine-DTW can be regarded as iteration of rough conversion and DTW. Thus affine-DTW performs alignment with less speaker identities. Fig. 2 shows an example of a result of affine-DTW. The figure shows alignment path converges via iterations.

C. GMM-based Feature Conversion

Let $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n_x}]$ and $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{n_y}]$ be D-dimensional vector sequences. These sequences represent the acoustic features of the source and target speakers' utterances respectively, which have the same linguistic contents. After $\boldsymbol{x}$ and $\boldsymbol{y}$ are aligned by DTW, $2D$-dimensional joint vectors $\boldsymbol{z}_t = \left[\boldsymbol{x}_t^\top \boldsymbol{y}_t^\top\right]^\top$ and these sequence $\boldsymbol{z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_T]$ are created. The notation $^\top$ denotes transposition of the vectors. In GMM-based VC, a GMM models probability density of the joint vectors $\boldsymbol{z}_t$ as follows:

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

In (1), $\mathcal{N}\left(\boldsymbol{z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}\right)$ denotes the multivariate Gaussian distribution with the mean vector $\boldsymbol{\mu}_m^{(z)}$ and the covariance matrix $\boldsymbol{\Sigma}_m^{(z)}$, m is the mixture component index, M is the total number of the components, and w_m denotes the positive weight of the m-th component where $\sum_{m=1}^{M} w_m = 1$. $\boldsymbol{\lambda}^{(z)}$

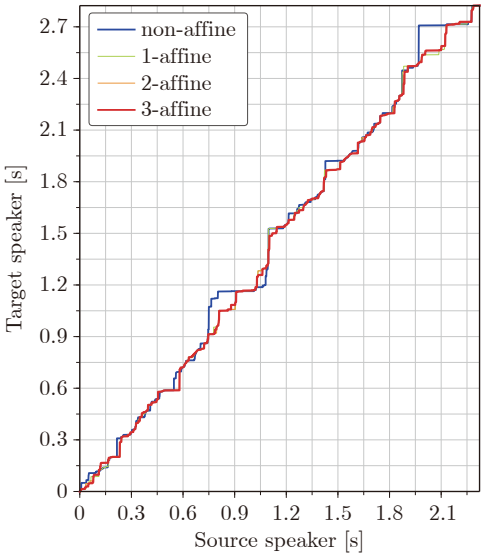


Fig. 2. Example of results of affine-DTW. Non-affine denotes the result of traditional DTW, and n-affine denotes the result of affine-DTW after n-th affine transformation.

denotes the parameter set of the GMM consisting of the number of mixtures and the weights, mean vectors and covariance matrices of all components. Since the stochastic variable $\boldsymbol{z}_t$ is a joint vector, $\boldsymbol{\mu}_m^{(z)}$ and $\boldsymbol{\Sigma}_m^{(z)}$ can be written as

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix}. \tag{2}$$

The mean vectors, covariance matrices and weights of the GMM can be iteratively estimated by using the EM algorithm.

The conditional probability density of $\boldsymbol{y}_t$ given $\boldsymbol{x}_t$ is approximately represented by the parameters of the joint density model as follows:

$$P\left(\boldsymbol{y}_t \middle| \boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}\right) = \sum_{m=1}^{M} P\left(m \middle| \boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}\right) P\left(\boldsymbol{y}_t \middle| \boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(z)}\right), \tag{3}$$

where

$$P\left(m \middle| \boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}\right) = \frac{w_m \mathcal{N}\left(\boldsymbol{x}_t; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)}\right)}{\sum_{m'=1}^{M} w_{m'} \mathcal{N}\left(\boldsymbol{x}_t; \boldsymbol{\mu}_{m'}^{(x)}, \boldsymbol{\Sigma}_{m'}^{(xx)}\right)}, \tag{4}$$

$$P\left(\boldsymbol{y}_t \middle| \boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(z)}\right) = \mathcal{N}\left(\boldsymbol{y}_t; \boldsymbol{E}_{m,t}^{(y)}, \boldsymbol{D}_m^{(y)}\right), \tag{5}$$

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

$$\boldsymbol{D}_m^{(y)} = \boldsymbol{\Sigma}_m^{(yy)} - \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)^{-1}} \boldsymbol{\Sigma}_m^{(xy)}. \tag{7}$$

The aim of VC is to obtain a mapping function $\mathcal{F}(\cdot)$ that converts source feature vectors into target vectors. Based on the minimum mean square error criterion, the function is derived as follows:

$$\mathcal{F}(\boldsymbol{x}_t) = \sum_{m=1}^{M} P\left(m \middle| \boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}\right) \boldsymbol{E}_{m,t}^{(y)}. \tag{8}$$

A parameter generation method on the basis of the maximum likelihood criterion is also proposed [8]. The estimated parameter $\hat{\boldsymbol{y}}_t$ is obtained from iteration of following equations:

$$\hat{\boldsymbol{y}}_t = \left(\sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{(y)^{-1}}\right)^{-1} \sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{(y)^{-1}} \boldsymbol{E}_{m,t}^{(y)}, \tag{9}$$

$$\gamma_{m,t} = P\left(m \middle| \boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{\lambda}^{(z)}\right). \tag{10}$$

In this parameter generation method, each precision matrix $\boldsymbol{D}_m^{(y)^{-1}}$ measures the confidence of the corresponding conditional mean vector $\boldsymbol{E}_{m,t}^{(y)}$.

The simple GMM-based VC system converts input features frame by frame, and therefore inappropriate transitions of features are easily observed. This problem can be effectively avoided by using static and dynamic features [8]. The source and target feature vector sequences are expanded to $\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top\right]^\top$ and $\boldsymbol{Y}_t = \left[\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top\right]^\top$ respectively, and the trained joint vector sequence is $\boldsymbol{Z}_t = \left[\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top\right]^\top$.

In addition, the generated features are often excessively smoothed compared with the ones of natural utterances. Compensation of the variances of the features in each utterance, or global variances, makes this oversmoothing suppressed [8].

D. Differential-spectrum Compensation

In traditional VC systems, vocoders are used to synthesize converted utterances. Briefly speaking, vocoders generate waveforms by convoluting source waveforms with vocal tract filters. Let $s(n)$ be a source waveform and $f_n(n)$ be impulse responses of vocal tract filters. Here, n denotes an index of the waveform sample. Note that vocal tract filters are time-variant. The vocoder system can be represented as follows:

$$v(n) = s(n) \otimes f_n(n), \tag{11}$$

where $v(n)$ is the generated waveform and $\otimes$ denotes time-varying convolution. It is difficult to model source waveforms, on which the quality of synthesized speech directly depends.

In VC systems, a source waveform $s^{(S)}(n)$ can be estimated from an input utterance $v^{(S)}(n)$ by convoluting their inverted vocal tract filters $f^{(S)}(n)$. This method can be expressed as follows:

$$s^{(S)}(n) = v^{(S)}(n) \oslash f_n^{(S)}(n), \tag{12}$$

where $\oslash$ denotes time-varying convolution with inverse filters. The target waveform $v^{(T)}(n)$ is generated by convolution of the estimated source waveform $s^{(S)}(n)$ with the target filters $f_n^{(T)}(n)$, or

$$v^{(T)}(n) = s^{(S)}(n) \otimes f_n^{(T)}(n). \tag{13}$$

This operation is equivalent to convoluting the filters of differential spectrum to the source utterances, or

$$v^{(T)}(n) = v^{(S)}(n) \otimes \left(f_n^{(T)}(n) \oslash f_n^{(S)}(n)\right). \tag{14}$$

This method is called differential-spectrum compensation or differential-spectrum filtering.

Traditionally, MLSA filtering is used for differential-spectrum compensation with mel-cepstral coefficients. This method uses the approximated filter coefficients derived by Padé approximation. To reduce the effects of the approximation, this paper introduces a new filtering method named SP-WORLD. This method reconstructs minimum-phase impulse responses from real cepstra. WORLD vocoder also uses this reconstruction method, and SP-WORLD is designed with the help of diversion of the reconstruction method to differential-spectrum compensation.

The reconstruction method is based on a simple signal processing approach [16]. Cepstra of minimum-phase impulse responses have the property of being causal. In other words, $\hat{h}_{\min}(n) = 0$ for $n < 0$ where $\hat{h}_{\min}(n)$ denotes the real cepstrum of the minimum-phase response $h_{\min}(n)$. Hence, the causal impulse response with the same frequency response as given can be derived as follows:

$$\hat{h}_{\min}(n) = \begin{cases} 0 & n < 0 \\ \hat{h}(n) & n = 0 \\ 2\hat{h}(n) & n > 0 \end{cases}, \tag{15}$$

where $\hat{h}(n)$ denotes the real cepstrum of the given impulse response.

As described above, differential-spectrum compensation can be interpreted as applying two vocal tract filters: the inverse filters of the source speaker and the forward filters of the target speaker. Since MLSA filtering generates the filter coefficients in the mel-cepstral domain, the orders of both filter coefficients are implicitly same. However, because the filters used in SP-WORLD are derived from division of the spectra, the denominators, or the vocal tract filters of the source speaker, do not depend on the mapping models. In short, the spectra obtained by analysis also can be used as the inverse filters instead of the reconstructed spectra from the mel-cepstral coefficients. Since the effectiveness of the inverse filters from the power spectra have not been studied, this paper experimentally investigates it in Section III-D.

III. Experiments

This section details several experiments conducted to evaluate the quality of generated utterances under three conditions: analysis and re-synthesis, ideal feature conversion, and statistical feature conversion. Fig. 3 shows the overview of these experiments.

A. Experimental Setups

The prepared data were the speech data of four Japanese speakers that uttered the ATR Japanese phonetically balanced sentence sets [17]. In this paper, the subset A (50 sentences) of the dataset was used. The speakers were composed of two male and two female speakers. The sampling frequency was 22 050 Hz. The conducted tests were preference AB tests for quality evaluations and ABX tests for speaker similarity evaluations described in Section III-D. In each test, 23 listeners answered the test via our crowdsourcing system. Each listener answered 10 questions and earned approximately $0.46 for his/her participation.

B. Analysis and Re-synthesis

This section examines the quality of re-synthesized sounds without feature modification to investigate parameters of analysis. As parameters, frame periods and the order of mel-cepstral coefficients were considered. Here, WORLD vocoder was used as a synthesis method.

First, the frame period was selected from $5\,\mathrm{ms}$, $1\,\mathrm{ms}$, $500\,\mu\mathrm{s}$, and $50\,\mu\mathrm{s}$. Here $50\,\mu\mathrm{s}$ means 1 sample, or $1/22050$ seconds. The spectral envelopes were not compressed to mel-cepstral coefficients in these experiments. Fig. 4 shows the results. On the whole, the shorter frame period made the quality higher, probably because of the fewer interpolations between frames. However, no significant difference appeared when the frame period was shorter than $1\,\mathrm{ms}$. Consequently, $1\,\mathrm{ms}$ is the appropriate frame period for re-synthesis.

Second, the order of the mel-cepstral coefficients was selected from a range of 24–99. The frame period was $1\,\mathrm{ms}$ or $50\,\mu\mathrm{s}$. Fig. 5 shows the results. The higher the order of the features, the higher the quality of the synthesized utterances. The results also show 79 and 39 were the significant upper limits of the quality improvement in $1\,\mathrm{ms}$ and $50\,\mu\mathrm{s}$ analysis respectively.

For objective evaluation, the log-spectral distances (LSD) between original and re-synthesized waveforms were calculated. LSD is defined as follows:

$$\mathrm{LSD[dB]} = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{2}{N}\sum_{n=1}^{N/2}\left(20\log_{10}\frac{A_{t,n}}{\hat{A}_{t,n}}\right)^2}, \tag{16}$$

where $A_{t,n}$ and $\hat{A}_{t,n}$ are amplitude spectra, N denotes the length of Fourier transformation, and T is the number of frames [18]. Fig. 6 shows the results. These results follow the results of the subjective experiments on the whole. The results show significant difference between $5\,\mathrm{ms}$ and $1\,\mathrm{ms}$ analysis. However, the shorter analysis than $1\,\mathrm{ms}$ made no difference probably because the time resolution of F_0 analysis of WORLD is always $1\,\mathrm{ms}$. The results also show that the higher order of mel-cepstral coefficients made quality better.

C. Ideal Conversion

In this section, the evaluated utterances were generated from converted features. Here, the feature mapping method was based on aligned parallel data and statistical models were not used. Hence, the conversion can be regarded as ideal conversion. As parameters, filtering methods, the order of mel-cepstral coefficients, and frame periods were considered. The alignment was derived by affine-DTW with the 24-order mel-cepstral coefficients. The differential-spectrum compensation was applied for speech synthesis, and the fundamental frequencies were not converted. The order of Padé approximation was 4 in MLSA filtering. The source and target speakers were

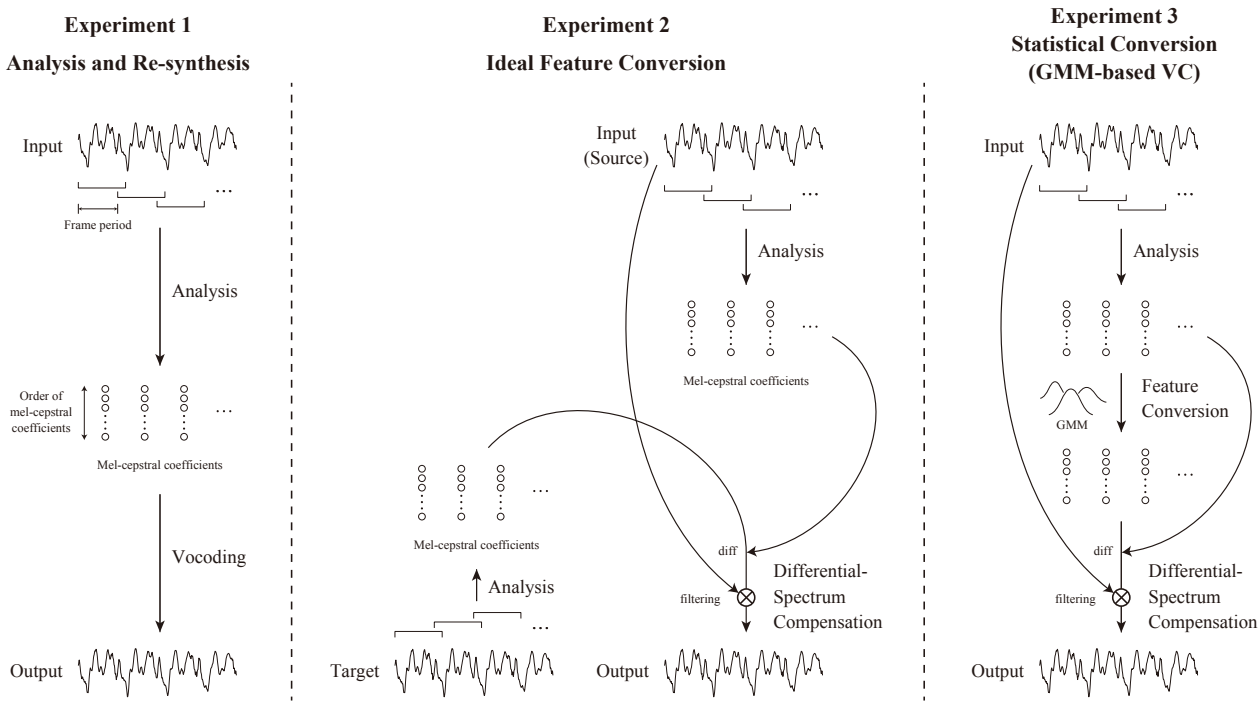


Fig. 3. Overview of the GMM-based VC framework discussed in this paper. Figure presenting experiment 3 is same as the right part of Fig. 1.

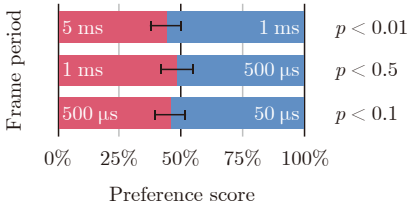
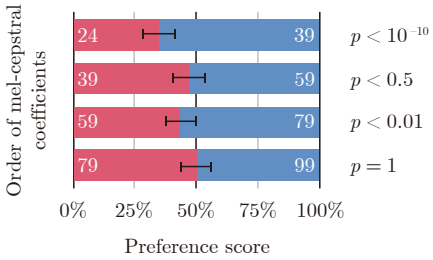
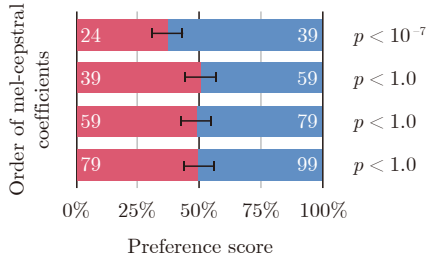


Fig. 4. Results of subjective evaluations of synthesis quality with different frame periods. Error bars denote 95% confidential intervals.



(a) Results of the 1 ms analysis.



(b) Results of the 50 µs analysis.

Fig. 5. Results of subjective evaluations of synthesis quality with the different order of mel-cepstral coefficients. Error bars denote 95% confidential intervals.

both male, and the fifty sentences were divided into two sets: 40 for training the transformation matrix in affine-DTW and 10 for evaluation.

First, the synthesis methods were compared. Fig. 7 shows the results. The effectiveness of SP-WORLD was observed especially with the high order of mel-cepstral coefficients.

Next, the order of mel-cepstral coefficients was examined. Fig. 8 shows the results. No significant difference was observed when the filtering method was MLSA. However, with SP-WORLD, the lower dimensional features were effective. Considering that complete conversion cannot be performed even with affine-DTW, this is probably because more precise conversion exposed conversion errors in higher dimensions. In other words, the ambiguous features were superior to the precise features in an auditory sense.

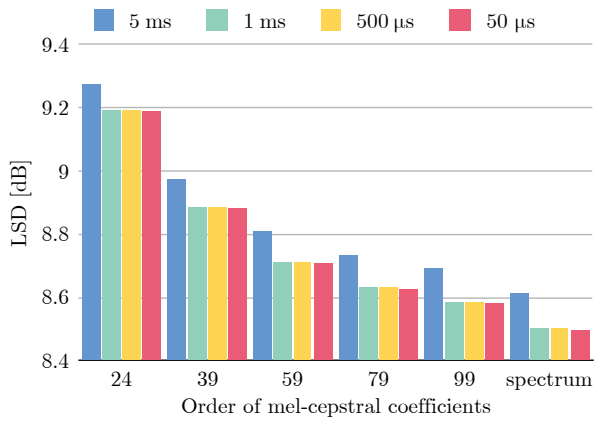


Fig. 6. Results of objective evaluations of synthesis quality.

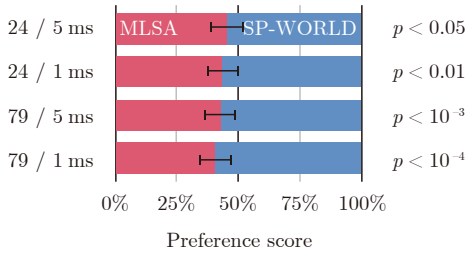


Fig. 7. Results of the subjective evaluations with different synthesis methods. Labels on the left side indicate the order of mel-cepstral coefficients and frame periods. Error bars denote 95% confidential intervals.

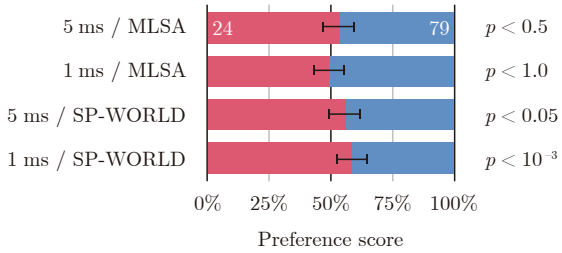


Fig. 8. Results of subjective evaluations with the different order of mel-cepstral coefficients. Labels on the left side indicate frame periods and synthesis methods. Error bars denote 95% confidential intervals.

Finally, the frame periods were investigated. When the order of mel-cepstral coefficients was 24 and SP-WORLD was used, 1 ms analysis was slightly effective.

These results indicate the best combination in ideal conversion is as follows: the filtering method is SP-WORLD, the order of mel-cepstral coefficients is 24, and the frame period is 1 ms.

D. Statistical Conversion

In this section, synthesis methods and two feature generation methods are evaluated in complete GMM-based VC frameworks. The order of mel-cepstral coefficients was fixed to 24, and the frame period was 1 ms. The used features were selected from three candidates: static features only, static and dynamic features, both features and the global variances. The synthesis method was also selected from three methods: MLSA, SP-WORLD with the reconstructed spectra from the mel-cepstral coefficients for the inverse filters (WM), and SP-WORLD with the analyzed spectra for the inverse filters (WS). The fifty sentences were divided into two sets: 40 for development and 10 for evaluation. The number of mixture components of GMM was optimized on the basis of the development set. There were 64 and 128 Gaussian components for only static features and for static and dynamic features, respectively. The autocovariance matrices and the cross-covariance matrices were assumed to be diagonal.

First, the synthesis methods were investigated. In these

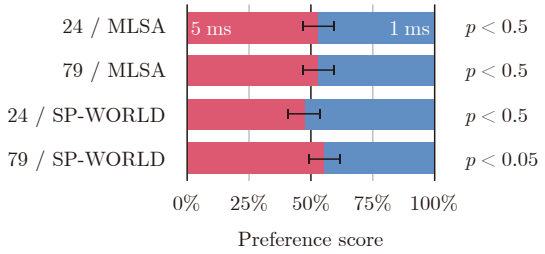
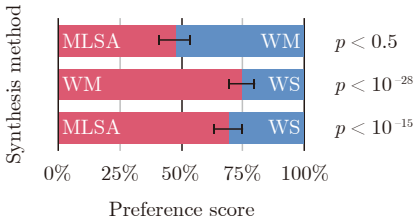
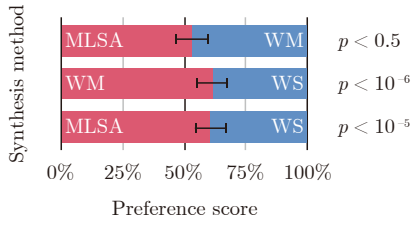


Fig. 9. Results of subjective evaluations with different frame periods. Labels on the left side indicate the order of mel-cepstral coefficients and synthesis methods. Error bars denote 95% confidential intervals.

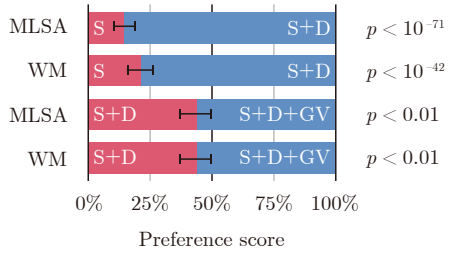


(a) Results of subjective evaluations of naturalness.

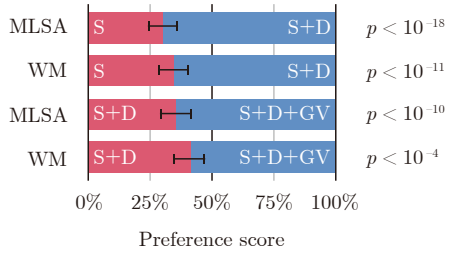


(b) Results of subjective evaluations of speaker similarity.

Fig. 10. Results of subjective evaluations with different synthesis methods in GMM-based VC. WM denotes SP-WORLD with the spectra derived from the mel-cepstral coefficients, and WS is SP-WORLD with the spectra obtained by analysis. Error bars denote 95% confidential intervals.



(a) Results of subjective evaluations of naturalness.



(b) Results of subjective evaluations of speaker similarity.

Fig. 11. Results of subjective evaluations with different sequential features in GMM-based VC. S denotes static features only, S+D means static and dynamic features, and S+D+GV denotes both features and global variances. Error bars denote 95% confidential intervals.

experiments, the static and dynamic features were used and the global variances were considered. Fig. 10 shows the results. No significant difference between MLSA and WM was observed. However, WS was inferior to the other two methods probably because the precise filters obtained by analysis deprived the converted features of the high order elements. On the other hand, the reconstructed filters from mel-cepstral coefficients did not contain high order features, and therefore these features remained in converted utterances.

Second, the effectiveness of the dynamic features and the global variances was examined. Fig. 11 shows that these sequential features were effective with both filtering methods.

IV. Conclusions

In this paper, in order to build a higher quality GMM-based VC system, the effects of hyperparameters were investigated by means of subjective experiments. The results showed the effectiveness of high time-resolution analysis and sequential features such as dynamic features and global variances. This paper also introduced a new filtering method named SP-WORLD for differential-spectrum compensation, and was found to be comparable to MLSA filtering. For future works, the effectiveness of the order of features and conversion methods of fundamental frequencies in total GMM-based VC systems also should be investigated. Additionally, the 1 ms barrier in F_0 analysis should be broken for higher time-resolution analysis. Moreover, these insights can be applied to the VC systems based on neural networks, and therefore the effectiveness of the application to these systems needs to be investigated experimentally.

Acknowledgment

This work was partly supported by SECOM Science and Technology Foundation.

References

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1988, pp. 655–658.

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

[3] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1998, pp. 285–288.

[4] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *IEEE Spoken Language Technology Workshop*, 2012, pp. 313–317.

[5] R. Aihara, T. Takiguchi, and Y. Ariki, "Activity-mapping non-negative matrix factorization for exemplar-based voice conversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 4899–4903.

[6] M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks," *Speech Communication*, vol. 16, no. 2, pp. 207–216, 1995.

[7] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 3893–3896.

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

[9] A. Mouchtaris, J. V. der Spiegel, and P. Mueller, "Non-parallel training for voice conversion by maximum likelihood constrained adaptation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2004, pp. I–1–4.

[10] C.-H. Lee and C.-H. Wu, "MAP-based adaptation for speech conversion using adaptation data selection and non-parallel training," in *INTERSPEECH*, 2006, pp. 2254–2257.

[11] D. Saito, S. Watanabe, A. Nakamura, and N. Minematsu, "Statistical voice conversion based on noisy channel model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1784–1794, 2012.

[12] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Reconstructing 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*, vol. 27, no. 3–4, pp. 187–207, 1999.

[13] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

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

[15] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1993, pp. II–554–557.

[16] S.-C. Pei and H.-S. Lin, "Minimum-phase FIR filter design using real cepstrum," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 53, no. 10, pp. 1113–1117, 2006.

[17] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990.

[18] J. Benesty, M. M. Sondhi, and Y. Huang, Eds., *Springer Handbook of Speech Processing*. Springer, 2008.