Articulatory Controllable Speech Modification based on Statistical Inversion and Production Mappings

Patrick Lumban Tobing, *Student member, IEEE,* Kazuhiro Kobayashi, *Student member, IEEE,*
Tomoki Toda, *Member, IEEE*

***Abstract*—In this paper, we present an innovative way of utilizing the natural relationship between speech sounds and articulatory movements by developing an articulatory controllable speech modification system. Specifically, we employ statistical acoustic-to-articulatory inversion mapping and articulatory-to-acoustic production mapping based on a Gaussian mixture model (GMM), allowing flexible modification of the model parameters and the independence of the text input features. Modification of an input speech signal through manipulation of the unobserved articulatory movements is achievable through a sequence of inversion and production mappings. To ensure the naturalness of articulatory movement trajectories, we introduce a method for manipulating articulatory parameters by considering their intercorrelation. Moreover, to generate high-quality modified speech sounds, we avoid the use of vocoder-based excitation generation by presenting several implementations of direct waveform modification capable of directly filtering an input speech signal using the differences in spectral parameters. The experimental results demonstrate that 1) higher accuracy in the estimation of spectral parameters is achieved by using sequential inversion and production mappings than for conventional production mapping using measured articulatory parameters, 2) the method for manipulating articulatory parameters by considering their intercorrelation makes it possible to generate more natural trajectories of modified articulatory movements, 3) the implementations of the direct waveform modification method significantly improve the quality of modified speech sounds, even under varying speaking conditions, and 4) the controllability of the system is ensured by its capability of producing modified vowel sounds through the manipulation of appropriate articulatory configurations.**

***Index Terms*—articulatory control, direct waveform modification, intercorrelation of articulators, Gaussian mixture model, speech modification, statistical inversion and production mappings**

I. Introduction

DURING speech production, both of the vocal folds and the articulators are used to achieve the so-called source-filter combination in the generation of speech signals [1], [2]. To accomplish this, the air pressure must be increased in the lungs. Then, the corresponding air flow is channeled in the trachea through the vocal folds. A particular characteristic of the excitation signal is then determined by the configuration of the vocal folds while the air is flowing, for example, a periodic signal is produced by constant vibrations of the vocal folds. Subsequently, this source-excitation signal is modulated within the vocal tract by the articulatory organs, including the tongue, teeth, and velum. Hence, a certain configuration of articulators appropriately determines the resonance/filter characteristics of the vocal tract, which, in turn naturally regulates the traits of the generated phonemic sounds.

This intimate relationship between speech and articulatory organs appears to be in contrast with their corresponding attributes. While producing speech sounds, the movements of the articulators in fact vary much more slowly than their counterparts in the speech signal [3]–[5], such as the trajectory of the vocal tract spectrum. Undeniably, a broad range of possibilities in the development of speech technologies would be viable through the utilization of slowly varying articulatory representations, such as articulatory parameters. Indeed, researchers have been extensively studying the use of articulatory parameters in speech technologies for several decades. Several notable comprehensive works have been reported on applications to low-bit-rate speech coding [3], [6], speech analysis and synthesis [3], [7], [8], speech recognition [9]–[11], and speech visualization [12], [13].

To establish a relationship between speech and vocal tract composition, it is widely known that the fundamental approach is based on mathematical functions [5], [6], [14], [15]. Unfortunately, the nature of the speech production mechanism itself does not provide a straightforward procedure for doing this. This is shown by the fact that there is no one-to-one mapping between speech signals and configurations of articulators. Such a peculiarity is observed in the so-called inverse mapping from acoustic to articulatory parameters [16], [17] as well as in the forward/production mapping from articulatory configurations to the vocal tract spectrum [18], [19]. A vast number of approximations must be considered when examining the affiliation between speech and articulators.

Recently, researchers have considered the use of statistical approaches. This has been made possible by the availability in parallel recording data of speech and articulatory movements [20]–[23]. Indeed, the elegance of capturing statistical traits within the accessible data has led to many notable works on the advancement of statistical data-driven methods for both acoustic-to-articulatory inversion and articulatory-to-acoustic production mappings. The statistical approach for the acoustic-to-articulatory inversion mapping was first introduced with the use of a codebook-based method [24]. Later, in [25], it was reported that by incorporating a constraint on the articulation dynamics, an improvement in the accuracy of inversion mapping can be achieved. The utilization of a neural-network method in inversion mapping was reported in [26],

P. L. Tobing, K. Kobayashi, and T. Toda are with the Graduate School of Information Science, Nagoya University, Nagoya, Aichi 464-8601, Japan.

Manuscript received September 14, 2017.

[27]; in [27], it was found that by employing multiple mixtures to model the density of articulatory features, significant improvement in the estimation accuracy can be obtained. Then, in [28], [29], phonetic information was used to improve the mapping effectiveness. Meanwhile, in [30], a Gaussian mixture model (GMM)-based mapping approach was shown to be capable of preserving the effectiveness while allowing independence from textual input features. For the articulatory-to-acoustic production mapping, the progress can be described in a similar manner to that for the inversion mapping, where it was first used with a codebook-based method in [31]. In [32], [33], production mapping based on a neural network was reported, and phonetic information was shown to enhance the production mapping performance in [34]. Similarly to above, a GMM-based production mapping was reported to perform effectively in [35].

In this paper, we focus on the GMM-based statistical method for both the acoustic-to-articulatory inversion and articulatory-to-acoustic production mappings [36], where the GMM concept itself is widely used in voice conversion systems [37]. The GMM-based statistical feature mapping technique essentially has three notable advantages. First, it provides non-black-box procedures in both modeling and estimation mechanisms with low resources. Second, it allows the possibility of developing language-independent systems owing to its independence of textual input features. Third, its low computational complexity opens a wide range of possibilities for implementation, particularly for real-time processing. Thus, in this work, to maximize the potential of GMM-based inversion and production mappings, we must look back and attempt to utilize the close relationship between speech sounds and articulatory movements. One effective means of achieving this is by developing a system capable of producing modified speech sounds through adjustments of unobserved articulatory features, which is closely related to a system using an HMM-based technique [38]. A system that is capable of performing the aforementioned scheme will offer immense opportunities in the development of various speech applications, such as acoustic and/or articulatory visualization feedback for speech therapy [39], [40], language learning/pronunciation training [41], [42].

To make it possible to take advantage of the use of articulatory parameters in the above speech applications, in this paper, we present an articulatory controllable speech modification system that employs the GMM-based statistical inversion and production mappings. This system allows one to modify an input speech signal through manipulation of the unobserved articulatory movements. In a more advanced development to adjust for various speech applications, one can conveniently adapt the manipulation of articulatory parameters for different procedures. In this system, however, we integrate the statistical inversion and production mappings into a single sequential mapping procedure, which allows one to adjust the unobserved articulatory parameters. These unobserved movements of the articulators can be conveniently modified with an advanced manipulation procedure, which considers the intercorrelation between articulatory parameters [43]. Additionally, high-quality modified speech sounds can be generated with the implementation of direct waveform modification method, capable of avoiding vocoder-based waveform generation by straightforwardly filtering an input speech signal with spectrum differential parameters [44], [45]. The experimental evaluation results suggest that the system makes it possible to produce more accurate spectral parameters, generate natural trajectories of modified articulatory movements, yield high-quality modified speech sounds, and control appropriate articulatory configurations for the modification of several vowel sounds. Note that, distinct than our previous work [43], [45], in this paper, comprehensive experimental results are described for both objective and subjective evaluations. [1]

This paper is organized as follows. In Section II, the parallel acoustic-articulatory speech data used in this paper are introduced. In Section III, the GMM-based statistical inversion and production mapping methods are described. In Section IV, the sequential procedure of inversion and production mappings and the methods for manipulating articulatory parameters are presented. In Section V, several methods of direct waveform modification with spectrum differentials are elaborated. In Section VI, the results of experimental evaluations are given. Finally, the paper is summarized in Section VIII.

II. Acoustic-Articulatory Speech Data

In this paper, we use the Multichannel Articulatory Database (MOCHA) [46] as the acoustic-articulatory data, which is provided by the Centre for Speech Technology Research (CSTR), University of Edinburgh. The MOCHA database consists of speech and articulatory movement data, which were simultaneously recorded at Queen Margaret University College. It contains two sets of speaker data from one male speaker (msak0) and one female speaker (fsew0), with both speakers having a Southern England accent. There are a total of 460 British TIMIT sentences uttered by each speaker.

The speech data were recorded with a sampling rate of 16 kHz. For the articulatory movement data, an electromagnetic articulograph (EMA) device was used to record the positions of articulators while speaking. The EMA data provide recorded measurements of seven articulators: the upper lip, lower lip, lower incisor, tongue tip, tongue body, tongue dorsum, and velum. The locations of the articulators are measured as x- and y-coordinates in the mid-sagittal plane, where the bridge of the nose and the upper incisor are chosen as points of reference. The articulatory movement data were recorded with a sampling rate of 500 kHz. Preprocessing procedures were performed [36] to reduce the effect of noise from measurement errors and normalize data values into Z-scores.

[1] In this paper, in the objective evaluation, more elaborate performance assessment of the sequential inversion and production mappings is given. In the subjective evaluation, results from female speaker data are given for experiments on both speech quality and speech controllability. More details on experimental data examinations are given, i.e., spectrogram details, formant frequency plot, and global variance comparison. We also derive the equations for obtaining the parameters of the differential GMM, described later in Section V-B3.

III. GMM-BASED STATISTICAL INVERSION AND PRODUCTION MAPPING METHODS

Let $\boldsymbol{c}_t$, $\boldsymbol{s}_t$, and $\boldsymbol{x}_t$ be the spectral envelope parameters, i.e., mel-cepstral coefficients; the source excitation parameters, i.e., log-scaled F_0 and log-scaled waveform power; and the articulatory parameters at frame t, respectively. The time sequence vectors of these parameters over an utterance are respectively defined as $\boldsymbol{c} = [\boldsymbol{c}_1^\top, \ldots, \boldsymbol{c}_T^\top]^\top$, $\boldsymbol{s} = [\boldsymbol{s}_1^\top, \ldots, \boldsymbol{s}_T^\top]^\top$, and $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_T^\top]^\top$, where T denotes the number of frames and $\top$ denotes the transposition of the vector. Note that, we deliberately describe the procedure of each of the inversion and production mappings, because of their differences in the employment of respective source and target feature vectors. Mainly, wider contextual frames are needed in the inversion mapping, while the use of source excitation features is helpful in the production mapping [36].

A. GMM-based Acoustic-to-Articulatory Inversion Mapping

In the inversion mapping, spectral envelope parameters of an input speech signal are converted into their corresponding articulatory parameters.

1) Feature extraction: As the source feature, a mel-cepstral segment feature vector $\boldsymbol{O}_t$ is used at frame t, which is extracted from the mel-cepstral parameters $\boldsymbol{c}_t$ at multiple frames around the current frame t, as given by

$$\boldsymbol{O}_t = \boldsymbol{A}[\boldsymbol{c}_{t-L}^\top, \ldots, \boldsymbol{c}_t^\top, \ldots, \boldsymbol{c}_{t+L}^\top]^\top + \boldsymbol{b}, \tag{1}$$

where $\boldsymbol{A}$ and $\boldsymbol{b}$ denote the parameters for the linear transformation, which are calculated beforehand by principal component analysis using the training data. As the target feature, a joint static and dynamic feature vector of articulatory parameters, given by $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$, is used at frame t, where $\Delta\boldsymbol{x}_t$ is the dynamic feature vector of the articulatory parameters.

2) Training and conversion: In the training procedure of inversion mapping, a joint source and target feature vector $[\boldsymbol{O}_t^\top, \boldsymbol{X}_t^\top]^\top$ is developed at each frame t from all utterances in the training data. The joint probability density function of the source and target features is then modeled with a GMM of the inversion mapping as follows:

$$P(\boldsymbol{O}_t, \boldsymbol{X}_t|\boldsymbol{\lambda}^{(O,X)}) = \sum_{m=1}^{M} \alpha_m^{(O,X)} \mathcal{N}([\boldsymbol{O}_t^\top, \boldsymbol{X}_t^\top]^\top; \boldsymbol{\mu}_m^{(O,X)}, \boldsymbol{\Sigma}_m^{(O,X)}), \tag{2}$$

where $\mathcal{N}(;\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. The parameter set for the GMM of the inversion mapping is denoted as $\boldsymbol{\lambda}^{(O,X)}$, which consists of weights $\alpha_m^{(O,X)}$, the mean vector $\boldsymbol{\mu}_m^{(O,X)}$, and the covariance matrix $\boldsymbol{\Sigma}_m^{(O,X)}$ of individual mixture components. The mixture component index is m and the total number of mixture components is M [2]. These model parameters are trained with the expectation-maximization (EM) algorithm. The training scheme for the GMM of the inversion mapping is shown in the upper diagram of Fig. 1.

[2]In this paper, the effective total number of mixture component M, for both inversion and production mappings, is between 64 and 128.

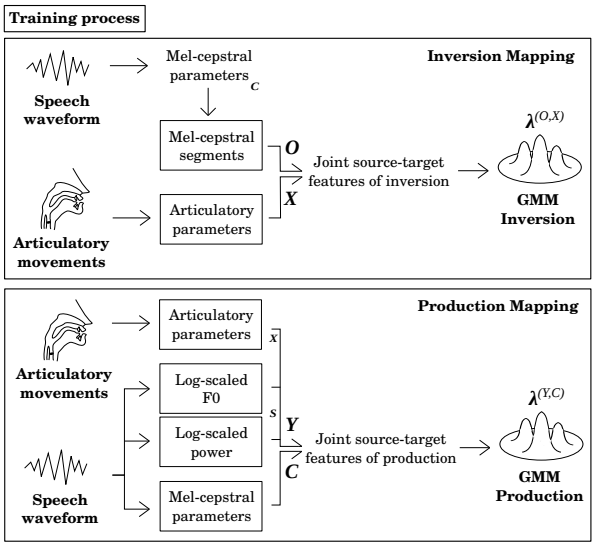


Fig. 1. Training scheme for GMM of the inversion mapping (top) and the production mapping (bottom).

In the conversion procedure, given a time sequence of mel-cepstral segment feature vectors $\boldsymbol{O} = [\boldsymbol{O}_1^\top, \ldots, \boldsymbol{O}_T^\top]^\top$, a time sequence of articulatory feature vectors $\boldsymbol{x}$ is estimated by employing a conditional probability density function, which is analytically derived from the GMM of the inversion mapping given in Eq. (2). In this paper, an approximation of the conditional probability density function is employed with the use of a single mixture component sequence $\boldsymbol{m} = \{m_1, \ldots, m_T\}$ [47], where m_t denotes the mixture component index at frame t. First, a suboptimum mixture component sequence $\hat{\boldsymbol{m}}^{(O)}$ is determined as

$$\hat{\boldsymbol{m}}^{(O)} = \arg\max_{\boldsymbol{m}} P(\boldsymbol{m}|\boldsymbol{O}, \boldsymbol{\lambda}^{(O,X)}). \tag{3}$$

Then, a time sequence of converted articulatory feature vectors $\hat{\boldsymbol{x}}$ is determined as follows:

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} P(\boldsymbol{X}|\boldsymbol{O}, \hat{\boldsymbol{m}}^{(O)}, \boldsymbol{\lambda}^{(O,X)}), \quad \text{subject to } \boldsymbol{X} = \boldsymbol{W}^{(x)}\boldsymbol{x}, \tag{4}$$

where $\boldsymbol{W}^{(\boldsymbol{x})}$ is the linear transformation matrix used to calculate the sequence of joint static and dynamic articulatory feature vectors $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_T^\top]^\top$ from a sequence of articulatory feature vectors $\boldsymbol{x}$. The conversion scheme using the GMM of the inversion mapping is shown in the upper diagram of Fig. 2.

B. GMM-based Articulatory-to-Acoustic Production Mapping

In the production mapping, articulatory parameters together with source excitation parameters are converted into their corresponding spectral envelope parameters.

1) Feature extraction: As the source feature, a joint static and dynamic feature vector of articulatory and source excitation parameters, given by $\boldsymbol{Y}_t = [\boldsymbol{x}_t^\top, \boldsymbol{s}_t^\top, \Delta\boldsymbol{x}_t^\top, \Delta\boldsymbol{s}_t^\top]^\top$, is used at frame t, where $\Delta\boldsymbol{s}_t$ is the dynamic feature vector of the source excitation parameters. As the target feature, a joint static and dynamic feature vector of mel-cepstral parameters,

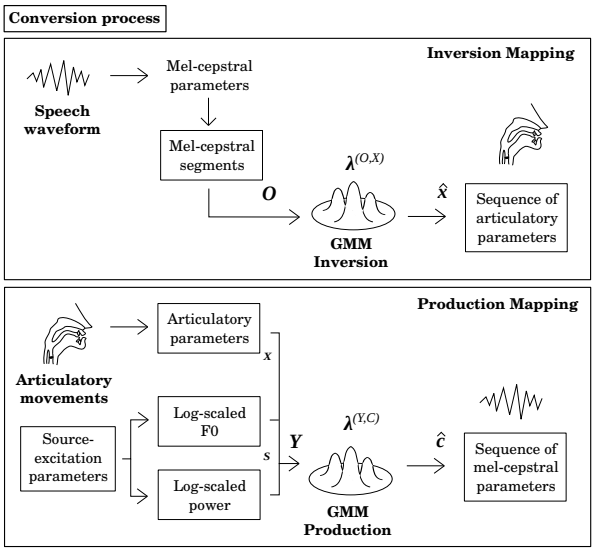


Fig. 2. Conversion scheme using GMM of the inversion mapping (top) and the production mapping (bottom).

given by $\boldsymbol{C}_t = [\boldsymbol{c}_t^\top, \Delta\boldsymbol{c}_t^\top]^\top$, is used at frame t, where $\Delta\boldsymbol{c}_t$ is the dynamic feature vector of the mel-cepstral parameters.

2) Training and conversion: In this training procedure, a joint source and target feature vector $[\boldsymbol{Y}_t^\top, \boldsymbol{C}_t^\top]^\top$ is developed at each frame t. A joint probability density function is then modeled with a GMM of the production mapping as follows:

$$P(\boldsymbol{Y}_t, \boldsymbol{C}_t|\boldsymbol{\lambda}^{(Y,C)}) = \sum_{m=1}^{M} \alpha_m^{(Y,C)} \mathcal{N}([\boldsymbol{Y}_t^\top, \boldsymbol{C}_t^\top]^\top; \boldsymbol{\mu}_m^{(Y,C)}, \boldsymbol{\Sigma}_m^{(Y,C)}), \quad (5)$$

where the parameter set for the GMM of the production mapping is denoted as $\boldsymbol{\lambda}^{(Y,C)}$, which consists of weights $\alpha_m^{(Y,C)}$, the mean vector $\boldsymbol{\mu}_m^{(Y,C)}$, and the covariance matrix $\boldsymbol{\Sigma}_m^{(Y,C)}$ of individual mixture components. These model parameters are also trained with the EM algorithm. The training scheme for the GMM of the production mapping is shown in the lower diagram of Fig. 1.

The conversion procedure for the production mapping is also performed in a similar way to in the inversion mapping. Given a time sequence of joint static and dynamic articulatory and source excitation feature vectors $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_T^\top]^\top$, first, the suboptimum mixture component sequence $\hat{\boldsymbol{m}}^{(Y)}$ is determined as

$$\hat{\boldsymbol{m}}^{(Y)} = \arg\max_{\boldsymbol{m}} P(\boldsymbol{m}|\boldsymbol{Y}, \boldsymbol{\lambda}^{(Y,C)}). \quad (6)$$

Then, a time sequence of converted mel-cepstral feature vectors $\hat{\boldsymbol{c}}$ is determined as follows:

$$\hat{\boldsymbol{c}} = \arg\max_{\boldsymbol{c}} P(\boldsymbol{C}|\boldsymbol{Y}, \hat{\boldsymbol{m}}^{(Y)}, \boldsymbol{\lambda}^{(Y,C)}), \quad \text{subject to } \boldsymbol{C} = \boldsymbol{W}^{(c)}\boldsymbol{c}, \quad (7)$$

where $\boldsymbol{W}^{(c)}$ is the linear transformation matrix used to calculate the sequence of joint static and dynamic mel-cepstral feature vectors $\boldsymbol{C} = [\boldsymbol{C}_1^\top, \ldots, \boldsymbol{C}_T^\top]^\top$ from a sequence of mel-cepstral feature vectors $\boldsymbol{c}$. The conversion scheme using the GMM of the production mapping is shown in the lower diagram of Fig. 2.

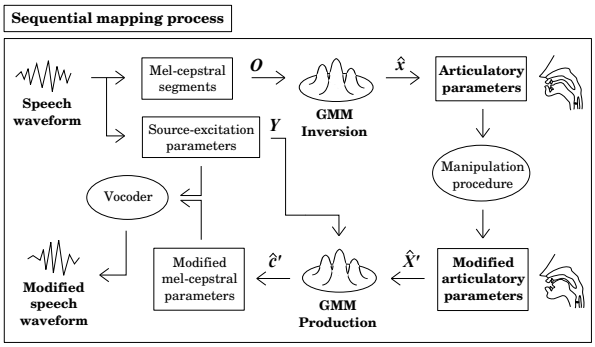


Fig. 3. Flow of the articulatory controllable speech modification system using the sequential inversion and production mapping procedure.

IV. Articulatory Controllable Speech Modification Using GMM-based Statistical Mapping Methods

A. Sequential Procedure of Inversion and Production Mappings

In this paper, to harness the use of articulatory parameters, we develop an articulatory controllable speech modification system based on a sequential mapping process using both the GMM of the inversion mapping and that of the production mapping. These two GMMs are trained by the procedures described in Sections III-A and III-B, respectively. By performing a sequence of inversion and production mappings, an input speech signal can be conveniently modified through manipulation of the unobserved articulatory movements. The methods for manipulating the articulatory parameters are elaborated in Section IV-B.

The flow of the sequential mapping is shown in Fig. 3. First, given an input speech signal, its spectral envelope parameters $\boldsymbol{c}$, i.e., mel-cepstral coefficients, and its source excitation parameters $\boldsymbol{s}$, i.e., log-scaled F_0 and log-scaled waveform power, are extracted. Then, the corresponding articulatory parameters $\hat{\boldsymbol{x}}$ are estimated from the mel-cepstral segments $\boldsymbol{O}$ by using the GMM of the inversion mapping as described in Section III-A2. To modify the spectral characteristics of the input speech signal, these estimated articulatory parameters $\hat{\boldsymbol{x}}$ are manually manipulated to produce a set of modified articulatory parameters $\hat{\boldsymbol{x}}'$. Then, the corresponding spectral envelope parameters $\hat{\boldsymbol{c}}'$ are estimated from the joint features $\boldsymbol{Y}$ of the modified articulatory parameters $\hat{\boldsymbol{x}}'$ and the source excitation parameters $\boldsymbol{s}$ by using the GMM of the production mapping as described in Section III-B2. Finally, the modified speech signal is generated from the modified spectral envelope parameters $\hat{\boldsymbol{c}}'$ and the source excitation parameters $\boldsymbol{s}$ by using a vocoder-based waveform generation procedure.

B. Methods for Manipulating Articulatory Parameters

To modify the unobserved articulatory movements, we present two methods for manipulating the articulatory parameters: a simple manipulation method and a smoothing

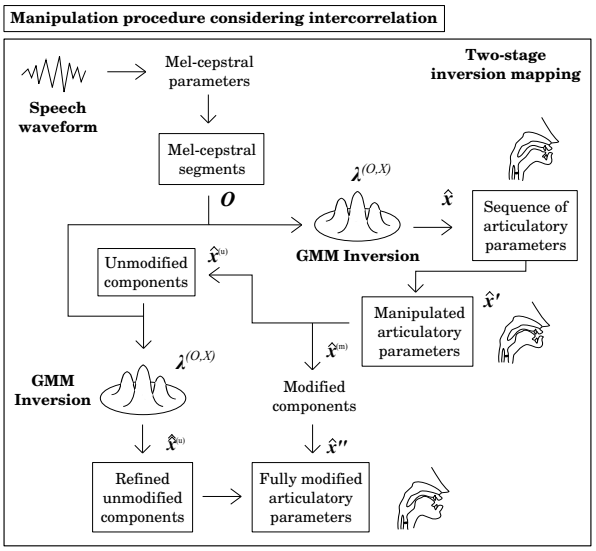


Fig. 4. Flow of the articulatory manipulation procedure that considers the inter-correlation of articulatory parameters by performing a two-stage inversion mapping.

method to take into account the intercorrelation of articulatory parameters.

1) Simple manipulation method: Let $\hat{\boldsymbol{x}}_t$ be the estimated articulatory feature vector at frame t. A manipulated articulatory feature vector $\hat{\boldsymbol{x}}'_t$ is then given by the following linear transformation:

$$\hat{\boldsymbol{x}}'_t = \boldsymbol{\Lambda}_t \hat{\boldsymbol{x}}_t + \boldsymbol{\psi}_t, \tag{8}$$

where the scaling matrix $\boldsymbol{\Lambda}_t$ and the shifting vector $\boldsymbol{\psi}_t$ are respectively written as

$$\boldsymbol{\Lambda}_t = \mathrm{diag}\left[\Lambda_t(1), \ldots, \Lambda_t(d), \ldots, \Lambda_t(D_x)\right], \tag{9}$$

$$\boldsymbol{\psi}_t = [\psi_t(1), \ldots, \psi_t(d), \ldots, \psi_t(D_x)]^\top. \tag{10}$$

Through the use of scaling factors in $\boldsymbol{\Lambda}_t$ at each frame t, the dilation or contraction of articulatory movements can be managed since Z-scores are used, i.e., the mean of the articulatory trajectory is not changed. On the other hand, the positions of individual articulators can be conveniently altered by using the shifting factors in $\boldsymbol{\psi}_t$ at each frame t.

By using the above linear transformation, the unobserved articulatory movements can be modified by manipulating the parameters of individual articulators with a set of scaling and shifting factors. However, it is known that these articulators have a certain degree of correlation between each other [48], for example, the movements of the tongue tip strongly affect those of the tongue body. Therefore, considering this fact, the manual manipulation of particular articulators must be compensated by the other articulators [17] by considering the degree of their correlation. Hence, unnatural articulatory movements are likely to be generated from this simple manipulation method.

2) Manipulation procedure considering intercorrelation of articulatory parameters: To generate more natural trajectories of modified articulatory movements, we present a manipulation procedure that takes into account the intercorrelation of articulatory parameters. To achieve this, specifically, we use a scheme that can be called a two-stage inversion mapping strategy. In the first stage, a sequence of articulatory parameters is estimated given the corresponding sequence of mel-cepstral segments. Then, a simple linear transformation is performed to manipulate these articulatory parameters as described in Section IV-B1. In the second stage, the modified components of the articulatory parameters are appended onto the input mel-cepstral segments. Then, a set of refined parameters corresponding to the unmodified articulatory components is estimated by utilizing the intercorrelation of articulatory parameters embedded within the GMM of the inversion mapping. Finally, a set of fully modified articulatory parameters is constructed from the modified components and the refined unmodified components. The flow of this manipulation procedure is depicted in Fig. 4.

Let $\hat{\boldsymbol{x}}_t^{(d)}$ and $\hat{\boldsymbol{x}}_t^{(u)}$ be the articulatory feature vectors of the modified components and the unmodified components, respectively, at frame t. Their joint static and dynamic feature vectors are then respectively given by $\hat{\boldsymbol{X}}_t^{(d)} = [\hat{\boldsymbol{x}}_t^{(d)^\top}, \Delta\hat{\boldsymbol{x}}_t^{(d)^\top}]^\top$ and $\hat{\boldsymbol{X}}_t^{(u)} = [\hat{\boldsymbol{x}}_t^{(u)^\top}, \Delta\hat{\boldsymbol{x}}_t^{(u)^\top}]^\top$ at frame t. Thus, by using the second stage of the inversion mapping, a sequence of refined unmodified components $\hat{\hat{\boldsymbol{x}}}^{(u)}$ can be estimated as follows:

$$\hat{\hat{\boldsymbol{x}}}^{(u)} = \arg\max_{\hat{\boldsymbol{x}}^{(u)}} P\left(\hat{\boldsymbol{X}}^{(u)} | \boldsymbol{O}, \hat{\boldsymbol{X}}^{(d)}, \hat{\boldsymbol{m}}^{(O)}, \boldsymbol{\lambda}^{(O,X)}\right), \quad \text{subject to } \hat{\boldsymbol{X}}^{(u)} = \boldsymbol{W}^{(x^{(u)})}\hat{\boldsymbol{x}}^{(u)}, \tag{11}$$

where the suboptimum mixture component sequence $\hat{\boldsymbol{m}}^{(O)}$ is determined by Eq. (3). The transformation matrix used to expand the dynamic features of the unmodified components is denoted as $\boldsymbol{W}^{(x^{(u)})}$. The corresponding sequences of articulatory feature vectors are written as $\hat{\boldsymbol{X}}^{(d)} = [\hat{\boldsymbol{X}}_1^{(d)^\top}, \ldots, \hat{\boldsymbol{X}}_T^{(d)^\top}]^\top$, $\hat{\boldsymbol{X}}^{(u)} = [\hat{\boldsymbol{X}}_1^{(u)^\top}, \ldots, \hat{\boldsymbol{X}}_T^{(u)^\top}]^\top$, and $\hat{\hat{\boldsymbol{x}}}^{(u)} = [\hat{\hat{\boldsymbol{x}}}_1^{(u)^\top}, \ldots, \hat{\hat{\boldsymbol{x}}}_T^{(u)^\top}]^\top$.

To capture the intercorrelation of articulatory parameters, first, the interdimensional correlation is taken into account with the use of mixture-dependent full-covariance matrices within the conditional pdf of the inversion mapping. Thus, the modified components of the articulatory parameters implicitly govern the change in the unmodified components through their statistical correspondence. Second, the interframe correlation of articulatory parameters is also considered owing to the use of a trajectory-based conversion framework [36], which employs an explicit relationship between the static and dynamic features [49]. Therefore, this manipulation procedure should be capable of generating natural movements of the articulatory parameters.

V. Articulatory Controllable Speech Modification without Vocoder-Based Waveform Generation

A. Problem in Terms of Speech Quality

In Section IV, the articulatory controllable speech modification system was introduced, where an input speech signal can be conveniently modified through manipulation of the unobserved articulatory movements. In this system, to generate a modified speech waveform, a vocoder-based framework is

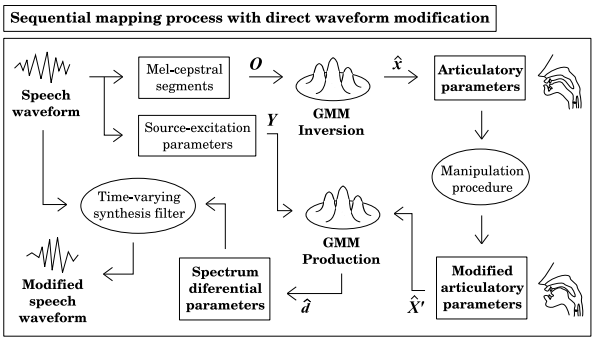


Fig. 5. Flow of the articulatory controllable speech modification system using direct waveform modification with spectrum differential parameters to generate a modified speech waveform.

employed, where the modified spectral envelope parameters and the source excitation parameters are used to generate the speech signal. However, it is well known that a vocoder-based procedure tends to degrade the quality of synthetic speech signals. Combined with its sensitivity to errors in speech parameterization, vocoder-based waveform generation would lead to a significant degradation in the quality of the modified speech waveform. In this paper, to alleviate this problem, we present several implementations of a direct waveform modification procedure [44] capable of avoiding the use of a vocoder-based excitation generation scheme by using spectrum differential parameters to directly filter an input speech signal. In this case, the spectrum differential parameters refer to the differences between modified and unmodified spectral envelope parameters. The flow of the sequential inversion and production mappings using the spectrum differential parameters is shown in Fig. 5.

By implementing a direct waveform modification procedure with spectrum differential parameters, we can alleviate the quality degradation caused by the use of a vocoder-based excitation generation process. However, in this framework, considering that the spectrum differential parameters are computed by using converted parameters, i.e., the converted spectral envelope parameters of modified speech, the quality of the modified speech waveform is still not optimized owing to the oversmoothed characteristics inherited from the trajectory-based conversion process [47]. One way to address the oversmoothing problem is by taking into account the global variance (GV) [47] and/or the modulation spectrum (MS) [50]. Nevertheless, the statistics of the GV or MS, which are obtained from the training data, do not exactly address the issue in new data. In this paper, to exactly address the oversmoothing problem, we present two other implementations of a direct waveform modification method that can preserve the fine structure of the spectral envelope from the input speech waveform.

B. Implementations of Direct Waveform Modification Method using Spectrum Differential Parameters

In a direct waveform modification method, an input speech signal is modified using the spectrum differential parameters by utilizing a time-varying synthesis filter, such as an MLSA filter [51]. To determine the best way of generating the spectrum differential parameters, we present three different methods: a basic method (DiffBM), a refined method (DiffRM), and a refined method with differential GMM (DiffGMM).

1) Basic method (DiffBM): In the basic method of calculating the spectrum differential parameters (DiffBM), extracted spectral envelope parameters of the input speech waveform and oversmoothed spectral envelope parameters of the modified speech waveform are employed. Let $\boldsymbol{c}$ be the time sequence of the spectral envelope parameters extracted from the input speech waveform and $\hat{\boldsymbol{c}}' = [\hat{\boldsymbol{c}}_1'^\top, \ldots, \hat{\boldsymbol{c}}_T'^\top]^\top$ be that of the oversmoothed spectral envelope parameters for the modified speech waveform. The time sequence of the DiffBM spectrum differential parameters $\boldsymbol{d}_{BM}$ is then given by

$$\boldsymbol{d}_{BM} = \hat{\boldsymbol{c}}' - \boldsymbol{c} = [[\hat{\boldsymbol{c}}_1' - \boldsymbol{c}_1]^\top, \ldots, [\hat{\boldsymbol{c}}_T' - \boldsymbol{c}_T]^\top]^\top. \quad (12)$$

A modified speech waveform is then generated by filtering the input speech waveform using the $\boldsymbol{d}_{BM}$ spectrum differential parameters. Therefore, the modified speech waveform can be characterized by a time sequence of DiffBM spectral envelope parameters $\boldsymbol{c}_{BM}$, which is given by

$$\begin{aligned}\boldsymbol{c}_{BM} &= \boldsymbol{c} + \boldsymbol{d}_{BM} \\ &= [[\boldsymbol{c}_1 + (\hat{\boldsymbol{c}}_1' - \boldsymbol{c}_1)]^\top, \ldots, [\boldsymbol{c}_T + (\hat{\boldsymbol{c}}_T' - \boldsymbol{c}_T)]^\top]^\top \\ &= [\hat{\boldsymbol{c}}_1'^\top, \ldots, \hat{\boldsymbol{c}}_T'^\top]^\top. \end{aligned} \quad (13)$$

Thus, the speech waveform modified by this basic method (DiffBM) is represented by a time sequence of the oversmoothed modified spectral envelope parameters $\hat{\boldsymbol{c}}'$. However, this sequence is completely different from that of the conventional vocoder-based system in terms of the excitation signal. This is because the direct filtering procedure of the input speech waveform avoids the use of vocoder-based excitation generation. The DiffBM scheme is shown on the left panel in Fig. 6.

2) Refined method to alleviate oversmoothing (DiffRM): In the refined method for calculating the spectrum differential parameters (DiffRM), the oversmoothing problem, which still appears in the basic method DiffBM, is alleviated by preserving the fine structure of the input speech waveform by employing oversmoothed spectral envelope parameters of both modified speech and unmodified speech waveforms. Let $\hat{\boldsymbol{c}}'$ be the time sequence of the oversmoothed spectral envelope parameters of the modified speech waveform and $\hat{\boldsymbol{c}} = [\hat{\boldsymbol{c}}_1^\top, \ldots, \hat{\boldsymbol{c}}_T^\top]^\top$ be that of the unmodified speech waveform. The time sequence of the DiffRM spectrum differential parameters $\boldsymbol{d}_{RM}$ is given by

$$\boldsymbol{d}_{RM} = \hat{\boldsymbol{c}}' - \hat{\boldsymbol{c}} = [[\hat{\boldsymbol{c}}_1' - \hat{\boldsymbol{c}}_1]^\top, \ldots, [\hat{\boldsymbol{c}}_T' - \hat{\boldsymbol{c}}_T]^\top]^\top, \quad (14)$$

where $\hat{\boldsymbol{c}}$ is given in Eq. (7).

Similarly to in the basic method, the modified speech waveform is generated by filtering the input speech waveform using the $\boldsymbol{d}_{RM}$ spectrum differential parameters. Thus, this modified speech waveform can be characterized by a time

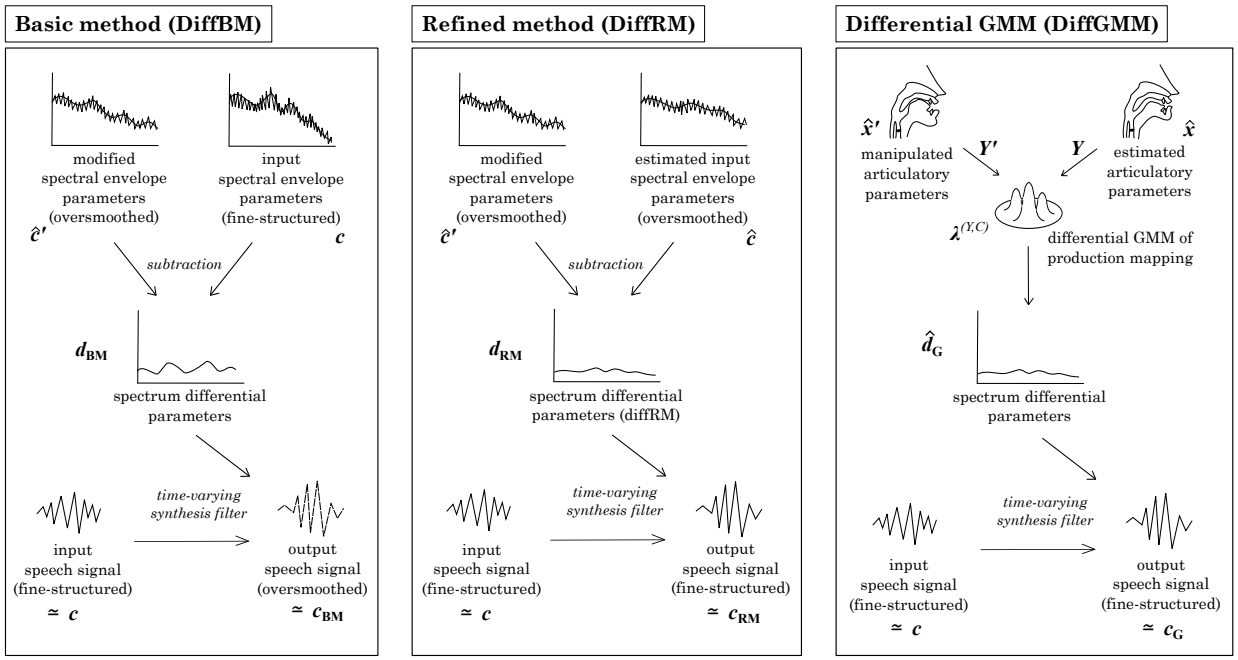


Fig. 6. Three different flows for the implementation of a direct waveform modification procedure in the articulatory controllable speech modification system according to the calculation scheme for the spectrum differential parameters.

sequence of DiffRM spectral envelope parameters $\boldsymbol{c}_{RM}$, which is given by

$$\begin{aligned}\boldsymbol{c}_{RM} &= \boldsymbol{c} + \boldsymbol{d}_{RM} \\ &= [[\boldsymbol{c}_1 + (\hat{\boldsymbol{c}}'_1 - \hat{\boldsymbol{c}}_1)]^\top, \ldots, [\boldsymbol{c}_T + (\hat{\boldsymbol{c}}'_T - \hat{\boldsymbol{c}}_T)]^\top]^\top \\ &= [[\hat{\boldsymbol{c}}'_1 + \boldsymbol{\epsilon}_1]^\top, \ldots, [\hat{\boldsymbol{c}}'_T + \boldsymbol{\epsilon}_T]^\top]^\top, \end{aligned} \tag{15}$$

where the refining factors are $\boldsymbol{\epsilon}_t = \boldsymbol{c}_1 - \hat{\boldsymbol{c}}_1$ at frame t. Hence, the modified speech waveform of the refined method (DiffRM) is represented not only by a time sequence of the oversmoothed modified spectral envelope parameters $\hat{\boldsymbol{c}}'$ but also by the residuals given in a time sequence of the refining factors $\boldsymbol{\epsilon} = [\boldsymbol{\epsilon}_1^\top, \ldots, \boldsymbol{\epsilon}_T^\top]^\top$ to preserve the fine structure of the spectral envelope. Consequently, the oversmoothed characteristic of the modified speech waveform is alleviated. The DiffRM scheme is shown in the middle panel in Fig. 6.

3) Refined method with differential GMM (DiffGMM): Finally, we present a method that works in a similar way to in the refined method DiffRM but in a more sophisticated manner by utilizing a differential GMM (DiffGMM). In this method, rather than generating the spectral envelope parameters twice, as in the DiffRM method, they are generated only once using the differential GMM of the production mapping. Let $\hat{\boldsymbol{Y}}' = [\hat{\boldsymbol{Y}}'^\top_1, \ldots, \hat{\boldsymbol{Y}}'^\top_T]^\top$ be the time sequence of the source excitation parameters and the modified articulatory parameters, and $\hat{\boldsymbol{Y}} = [\hat{\boldsymbol{Y}}^\top_1, \ldots, \hat{\boldsymbol{Y}}^\top_T]^\top$ be that of the unmodified articulatory parameters. At frame t, their corresponding feature vectors are respectively given by $\hat{\boldsymbol{Y}}'_t = [\hat{\boldsymbol{x}}'^\top_t, \boldsymbol{s}_t^\top, \Delta\hat{\boldsymbol{x}}'^\top_t, \Delta\boldsymbol{s}_t^\top]^\top$ and $\hat{\boldsymbol{Y}}_t = [\hat{\boldsymbol{x}}^\top_t, \boldsymbol{s}_t^\top, \Delta\hat{\boldsymbol{x}}^\top_t, \Delta\boldsymbol{s}_t^\top]^\top$. Then, the time sequence of DiffGMM spectrum differential parameters $\hat{\boldsymbol{d}}_G$ is estimated as follows:

$$\hat{\boldsymbol{d}}_G = \arg\max_{\boldsymbol{d}_G} P\left(\boldsymbol{D}_G | \boldsymbol{Y}', \boldsymbol{Y}, \boldsymbol{\lambda}^{(Y,C)}\right),$$
$$\text{subject to } \boldsymbol{D}_G = \boldsymbol{C}' - \boldsymbol{C}$$
$$\text{and } \boldsymbol{D}_G = \boldsymbol{W}^{(c)}\boldsymbol{d}_G. \tag{16}$$

The derivation of the DiffGMM parameters in Eq. 16 is given in the Appendix.

Then, a modified speech waveform of the DiffGMM method is generated by filtering the input speech waveform using the time sequence of the estimated spectrum differential parameters $\hat{\boldsymbol{d}}_G = [\hat{\boldsymbol{d}}^\top_{G_1}, \ldots, \hat{\boldsymbol{d}}^\top_{G_T}]^\top$ as follows:

$$\boldsymbol{c}_G = \boldsymbol{c} + \hat{\boldsymbol{d}}_G. \tag{17}$$

Therefore, the corresponding modified speech waveform is characterized by the time sequence of spectral envelope parameters $\boldsymbol{c}_G$, where the oversmoothed structure has been alleviated by preserving the fine structure of the input speech waveform because $\hat{\boldsymbol{d}}_{G_t} = \hat{\boldsymbol{c}}'_t - \hat{\boldsymbol{c}}_t$. However, the procedure is different from that of the refined method DiffRM because the parameters are only generated once using the differential GMM of the production mapping. Furthermore, it would also be straightforward to apply additional techniques, such as GV [52] or MS modeling [50]. The DiffGMM scheme is shown on the right panel in Fig. 6.

VI. Experimental Evaluation

A. Experimental Conditions

We used the parallel acoustic-articulatory data provided in MOCHA, described in Section II. As the spectral envelope parameters, we used the first through 24^{th} mel-cepstral coefficients converted from the spectral envelope, which was extracted frame-by-frame by STRAIGHT analysis [53]. As the source excitation parameters, we used log-scaled F_0 values also including an unvoiced/voiced binary decision feature

and log-scaled power values extracted from the STRAIGHT spectrum. The fixed-point analysis [54] in STRAIGHT was used to extract the F_0 values. As the articulatory parameters, we used 14-dimensional EMA data, elaborated in Section II, which were converted into Z-scores. The speech data were sampled at 16 kHz. The frame shift was set to 5 ms. The contextual frame length in Eq. (1) was set to ± 10 frames.

We performed both objective and subjective evaluations to assess the performance of the articulatory controllable speech modification system. In the objective evaluation, first, the accuracy of the inversion mapping, described in Section III-A, was measured by comparing the estimated articulatory parameters with the measured values. Then, the accuracy of the production mapping, described in Section III-B, was measured by comparing the estimated spectral envelope parameters, converted from the measured articulatory parameters, with the extracted spectral envelope parameters. Finally, the accuracy of the sequential procedure of inversion and production mappings, described in Section IV-A, was measured by comparing the estimated spectral envelope parameters, converted from the estimated articulatory parameters, with the extracted spectral envelope parameters. On the other hand, in the subjective evaluation, we evaluated both the quality of the generated speech sounds and the controllability of the system. In the first subjective evaluation of the speech quality, we compared the performance of the methods for manipulating articulatory parameters, described in Section IV-B, in terms of the naturalness of the modified speech sounds. Then, in the second subjective evaluation, we compared the performance of the implementations of the direct waveform modification method, described in Section V, which avoids the use of a vocoder-based procedure, by examining the quality of modified speech waveforms under several speaking conditions. Finally, the controllability of the system was evaluated by a categorical perception evaluation in which several vowel sounds were modified by manipulating the articulatory positions.

B. Objective Evaluation

1) Accuracy of inversion mapping: To measure the accuracy of the acoustic-to-articulatory inversion mapping described in Section III-A, first, we calculated the root-mean-square (RMS) error of the estimated articulatory parameters relative to the measured values as follows:

$$\mathrm{RMSE}(d) = \sqrt{\frac{\sum_{t=1}^{T}(a_t^{(o)}(d) - a_t^{(e)}(d))^2}{T}}, \qquad (18)$$

where $\mathrm{RMSE}(d)$ is the RMS error for the dth dimension of the articulatory parameters. The measured and estimated dth dimension articulatory parameters are respectively denoted as $a_t^{(o)}(d)$ and $a_t^{(e)}(d)$ at frame t. The lowest errors of 1.42 mm and 1.41 mm were achieved by using 128 mixture components for both male and female speakers, respectively. This result is consistent with the related work in [36].

Secondly, we measured the correlation coefficient, which was also calculated between the estimated and measured articulatory parameters as follows:

$$r(d) = \frac{\sum_{t=1}^{T}(a_t^{(o)}(d) - \hat{a}_t^{(o)}(d))(a_t^{(e)}(d) - \hat{a}_t^{(e)}(d))}{\sqrt{\sum_{t=1}^{T}(a_t^{(o)}(d) - \hat{a}_t^{(o)}(d))^2 \sum_{t=1}^{T}(a_t^{(e)}(d) - \hat{a}_t^{(e)}(d))^2}}, \qquad (19)$$

where $r(d)$ is the correlation coefficient for the dth dimension of the articulatory parameters. The mean values of the measured and estimated dth dimension articulatory parameters are respectively denoted as $\hat{a}_t^{(o)}(d)$ and $\hat{a}_t^{(e)}(d)$ at frame t. The highest correlation coefficients of 0.79 and 0.80 were yielded by using 128 mixture components for both male and female speakers, respectively. This result is also consistent with [36].

2) Accuracy of production mapping: To measure the accuracy of the articulatory-to-acoustic production mapping, described in Section III-B, we calculated the mel-cepstral distortion between the estimated mel-cepstral parameters and the extracted values as follows:

$$\text{Mel-CD[dB]} = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{24}(c^{(o)}(d) - c^{(e)}(d))^2}, \qquad (20)$$

where $c^{(o)}(d)$ and $c^{(e)}(d)$ denote the dth dimension of the extracted and estimated mel-cepstral parameters, respectively. The final result was averaged over all samples of training data and over all 24 dimensions of mel-cepstral parameters. The lowest mel-cepstral distortion values of 4.70 dB and 4.94 dB were achieved by using 64 mixture components for both male and female speakers, respectively, which are also comparable results to those in [36].

3) Accuracy of sequential procedure of inversion and production mappings: To assess the effectiveness of the sequential inversion and production mappings, described in Section IV-A, we also measured the mel-cepstral distortion between estimated mel-cepstral parameters and extracted mel-cepstral parameters. The estimated mel-cepstral parameters were converted from the estimated articulatory parameters with the sequential inversion-production mapping. Furthermore, we also measured the mel-cepstral distortion results yielded by using the GMM of the production mapping trained with the estimated articulatory training data instead of the measured articulatory training data.

The lowest distortion values of 4.38 dB and 4.65 dB were achieved by using 128 mixture components for both male and female speakers, respectively. This improvement was achieved because the estimated articulatory parameters are the most likely ones to be converted from the input mel-cepstral parameter sequence. Therefore, by estimating the input mel-cepstral parameters, which is performed in the sequential inversion-production mapping, one can evaluate the appropriateness of this mapping procedure. On the other hand, by using the estimated articulatory training data to train the GMM production used in the sequential mapping, the lowest values of 3.99 dB and 4.20 dB were achieved by using 64 mixture components for both male and female speakers, respectively. In the following experiments for subjective evaluation, we used the GMMs for the sequential inversion and production mappings with 128 mixture components.

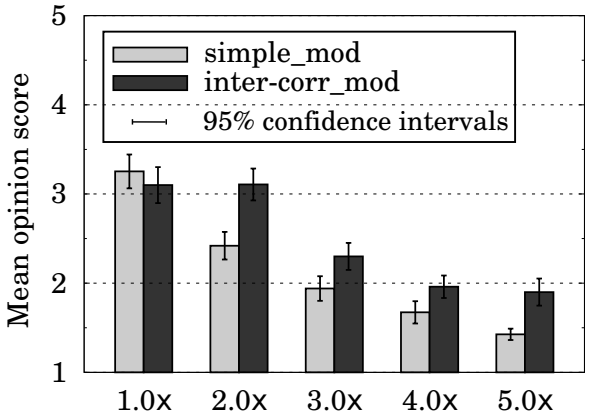


Fig. 7. Mean opinion score (MOS) results of male speaker (msak0) for the evaluation of modified speech quality by scaling the tongue tip movements in y-coordinate with two different methods of articulatory parameter manipulation.

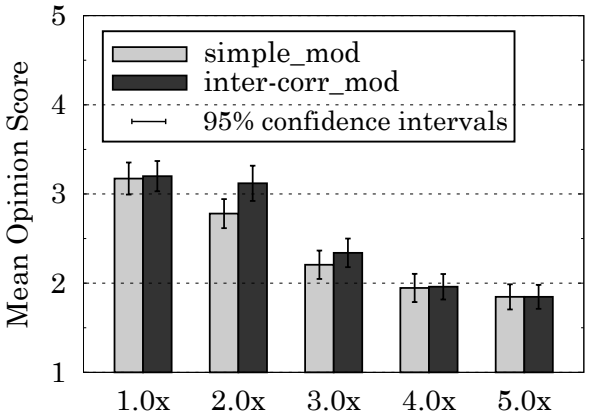


Fig. 8. Mean opinion score (MOS) results of female speaker (fsew0) for the evaluation of modified speech quality by scaling the tongue tip movements in y-coordinate with two different methods of articulatory parameter manipulation.

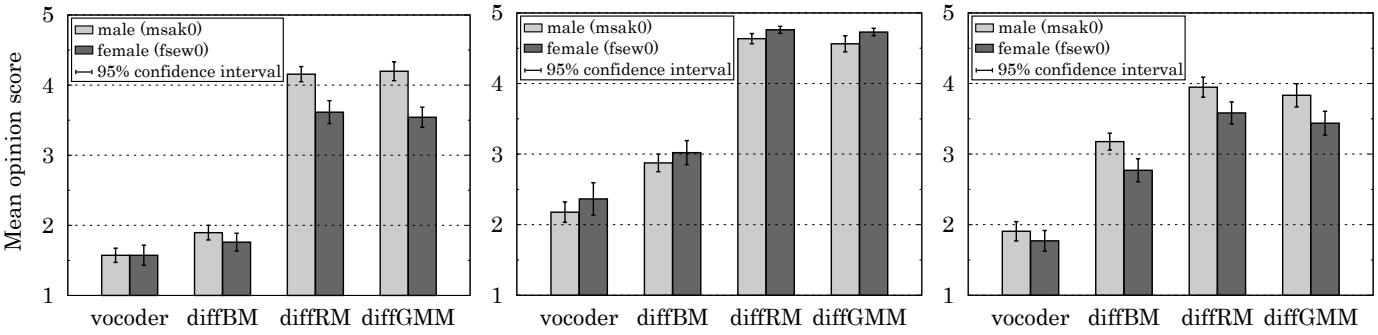
C. Subjective Evaluation of Speech Quality

1) Comparison of articulatory manipulation methods: In the first subjective evaluation of the speech quality, we compared the performance of the methods for manipulating articulatory parameters in Section IV-B. To do this, we modified the scaling factors of the tongue tip movements on the y-axis using five scaling values from 1.0-fold to 5.0-fold (hyperarticulation can be achieved by exaggerating the articulatory motions with value more than 1.0, while hypoarticulation by dimishing their range with value less than 1.0). Listeners were asked to evaluate the quality of the modified speech sounds in a mean opinion score (MOS) evaluation using a range of scores from 1.0 to 5.0, where 5.0 was the highest. The number of listeners was 10. The number of distinct utterances per listener was 15, which were randomly taken from the 110 evaluation data.

The MOS results for the different methods of manipulating the articulatory parameters are shown in Figs. 7 and 8 for the male and female speaker, respectively. The results show that the method considering the intercorrelation of articulatory parameters, described in Section IV-B2, gives higher scores than the simple linear transformation method for both of the male and female speaker data. It can also be observed that this method still preserves the quality of the modified speech up to a scaling value of 2.0-fold, which implies that higher values would lead to possible abnormalities caused by the physical constraints within the vocal tract being exceeded.

2) Comparison of spectrum differential calculation methods: In the second subjective evaluation of speech quality, we compared the performance of the implementations of the direct waveform modification method in Section V, which avoids the use of a vocoder-based speech generation framework to alleviate the quality degradation of synthetic speech. To do this, we emulated three speaking conditions by scaling the trajectories of articulatory movements: normal articulation, hypoarticulation, and hyperarticulation. Naturally, prosodic elements, such as the speaking rate and glottal stops, are included in the characterization of speaking conditions [55]. However, because the spectral characteristic is the characteristic most closely related to the phonetic quality, it can be relatively easily modified by manipulating the articulatory movements. To accomplish this, we used 1.0-fold scaling to emulate the normal articulation condition, 0.5-fold scaling for hypoarticulation, and 2.0-fold scaling for hyperarticulation. A MOS evaluation was conducted to assess the quality of modified speech sounds, with a range of scores from 1.0 to 5.0. Four different speech generation procedures were compared: the vocoder-based method, the basic method of direct waveform modification (DiffBM), the refined method (DiffRM), and the refined method with a differential GMM (DiffGMM). The number of listeners was 12. The number of distinct utterances was 8.

The MOS results of the male and female speakers are shown in Fig. 9. These results demonstrate that the implementations of the direct waveform modification method, particularly the refined method (DiffRM) and the refined method with the differential GMM (DiffGMM), significantly improve the quality of the modified speech over all three speaking conditions for both male and female speakers. On the other hand, the basic method (DiffBM) yields only a small improvement from the conventional vocoder-based method compared with the DiffRM and DiffGMM. This is because, even though the vocoder-based excitation generation procedure is avoided, the overall structure of the generated speech waveform still inherits the oversmoothed characteristics, which are alleviated in the DiffRM and DiffGMM methods. Higher scores yielded from the male speaker may have been caused by the higher accuracy in the estimation of speech spectrum for the male speaker than the female speaker. Spectrograms of a sample utterance, i.e., "Dolphins are intelligent marine mammals", from the male speaker generated using all four speech generation procedures in the hyperarticulation speaking condition (2.0-fold scaling) are shown in Fig. 10. It can be observed that at higher frequency bands, a strong periodic structure is generated by the vocoder-based method, while the DiffBM method is capable of preserving the more natural aperiodic structure, which is further refined by using either DiffRM or DiffGMM.



(a) Hypoarticulation speaking condition (0.5-fold scaling) (b) Normal articulation speaking condition (1.0-fold scaling) (c) Hyperarticulation speaking condition (2.0-fold scaling)

Fig. 9. MOS results for the evaluation of the modified speech quality under three speaking conditions using the vocoder-based framework and three implementations of the direct waveform modification method.

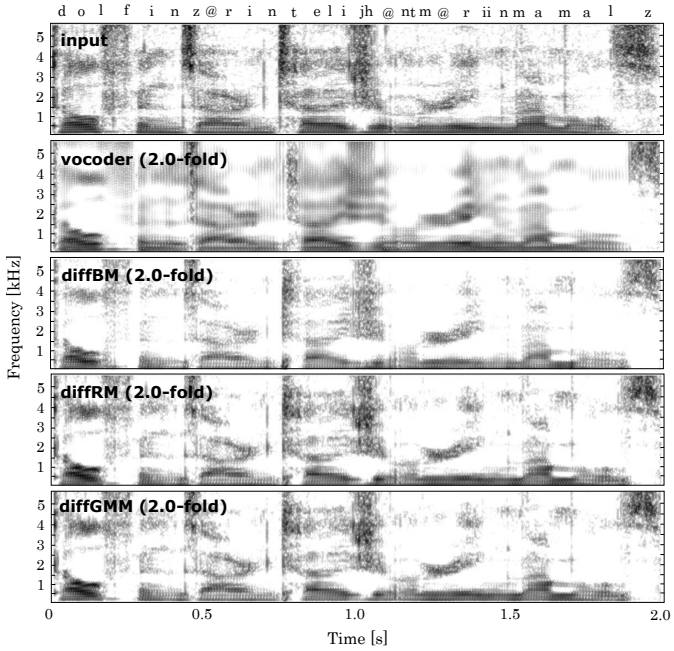


Fig. 10. Spectrograms of modified speech waveform under hyperarticulation (2.0-fold) condition using vocoder-based procedure (second top) and the three direct waveform modification schemes for the utterance "Dolphins are intelligent marine mammals" from the male speaker, while the spectrogram of input speech waveform is shown at the top.

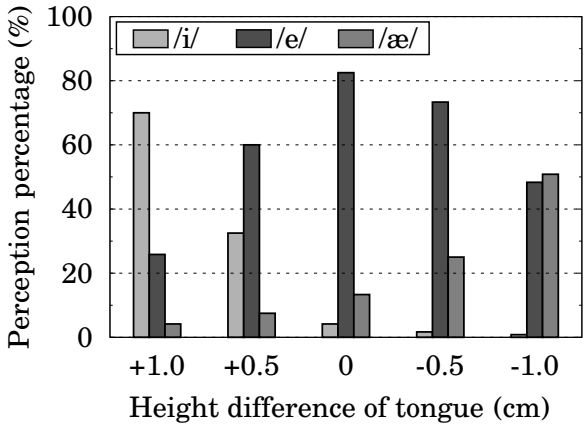


Fig. 11. Categorical perception results for evaluation of controllability of the system by modification of vowel sounds for male speaker.

D. Subjective Evaluation of Controllability

In the final subjective evaluation, we assessed the controllability of the articulatory controllable speech modification system by controlling several vowel sounds through the manipulation of articulatory positions. Specifically, we modified three front vowels in English, i.e., /ɪ/, /ɛ/, and /æ/. The prominent difference between these vowels in terms of articulation is in the height of the tongue during their pronunciation. For vowel /ɪ/, the tongue is located at the highest position among the three vowels, that for vowel /æ/ is located at the lowest position, and that for vowel /ɛ/ is located at an intermediate position. To simulate these conditions, we set the tongue position at the middle frame of vowel /ɛ/ at five different positions relative to the original position: $+1.0$ cm, $+0.5$ cm, 0 cm, -0.5 cm, and -1.0 cm. A more positive value means that the position is higher. The modified middle-frame position of the tongue height was interpolated to the middle-frame configurations of its surrounding left and right phonemes by using cubic-spline interpolation to ensure a smooth trajectory for the modified articulatory positions. In total, we chose 10 distinct words containing the vowel /ɛ/ excerpted from the evaluation data. The evaluation involved a categorical perception procedure. Each of the modified speech samples of the chosen words was presented to the listeners, along with a label showing its written word including the modified vowel, which was written with a question mark. Each of the listeners was asked to guess the missing vowel, either /ɪ/, /ɛ/, or /æ/. The total number of listeners was 10 none of which were native English speakers. The refined direct waveform modification method with the differential GMM (DiffGMM) was used to generate the speech sounds. Frames corresponding to the target vowels, i.e., /ɪ/ and /æ/, were removed when training the GMMs.

Results of the categorical perception evaluation for both male and female speakers are respectively shown in Fig. 11 and Fig. 12. It can be observed that the differences in the articulatory position configuration indeed lead to a change in the perception of the vowel sounds, as has also been observed in [38]. This is shown by the high gradient of the

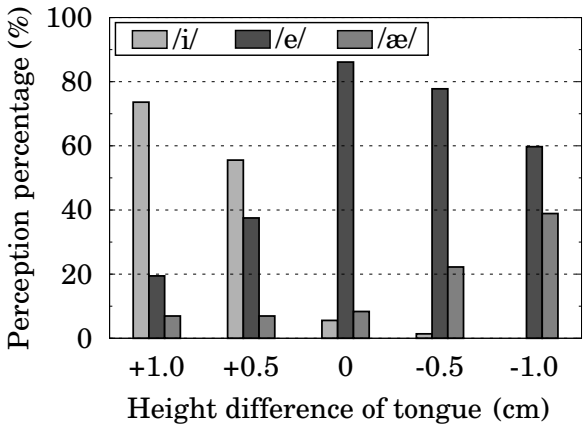


Fig. 12. Categorical perception results for evaluation of controllability of the system by modification of vowel sounds for female speaker.

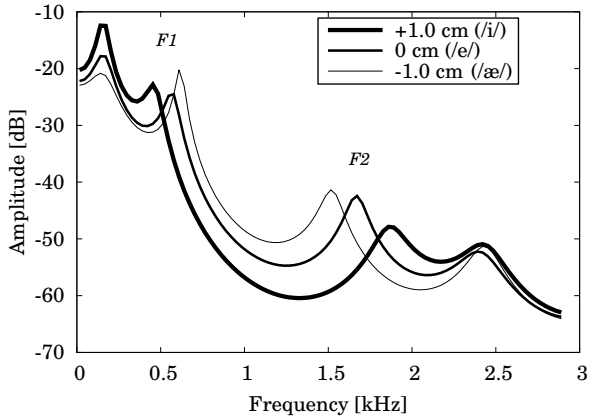


Fig. 13. LPC spectrums containing the first three formants from the speech section of the modified vowel for the word "stems" from the male speaker with three different relative heights of the tongue.

perception for vowel /ɪ/ as the position of the tongue becomes higher and the moderate gradient for vowel /æ/ as the position becomes lower. This difference is most likely caused by the relatively similar spectral characteristics of vowels /ɪ/ and /æ/, considering that none of the listeners were native English speakers. Samples of linear predictive coding (LPC) spectrums containing the first three formants of the modified vowel for the word "stems" from the male speaker are shown in Fig. 13. These spectrums suggest the consistency of the formant characteristics of these three vowels, where vowel /ɪ/ has the lowest F1 and highest F2, vowel /æ/ has the highest F1 and lowest F2, and vowel /ɛ/ has intermediate values for both formants. Moreover, comparison of global variance (GV) of the mel-cepstral parameters, over all utterances containing the modified vowels, generated with the vocoder-based and the all three spectrum differential system, is given in Fig. 14. The graph shows that the GVs computed from the diffRM and diffGMM methods are very close to those of the original waveform, in which high-quality modified speech sounds can be confirmed. Note that because of the removal of the frames corresponding to the target vowels in the training procedure, this implies that alterations in articulatory configurations that are not yet known/learned can still lead to the appropriate production of intended speech sounds. Applications in which articulatory parameters are deliberately modified for effect, e.g. for language learning or speech therapy, rely on the underlying waveform generation to be well behaved when pushed into configurations outside of that found in the training data. We consider this experiment an important test of the robustness of our approach.

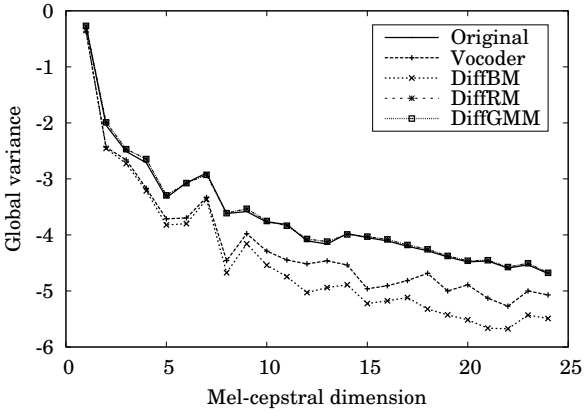


Fig. 14. Global variance of each mel-cepstral dimensions from all utterances containing modified vowel /ɪ/ onto /ɛ/ by leveraging the tongue height by 1 cm using the vocoder-based system and the three proposed spectrum differential systems. Global variances of the original waveform are also presented.

VII. Discussion

To make our proposed system more viable, we realize that it should not rely on only a limited amount of speaker characteristics. However, it is to be understood that the development of articulatory data is not a straightforward procedure. Therefore, we need to consider an approach that could take advantage of the available speech and articulatory data to adapt with arbitrary speaker characteristics. Such approach would alleviate the need in collecting articulatory data of a desired new speaker. One effective way of doing it is to use the vocal tract length normalization (VTLN) technique [56]. In this method, to compensate the difference in vocal tract lengths, a frequency warping function is employed in warping the frequency spectrum before computing the cepstral parameters. Hence, the acoustic space of available speakers can be adapted into the acoustic space of the desired speaker. By using the trained mapping models with the adapted acoustic space, given an unseen speech data of the target speaker, its corresponding articulatory movements can be estimated. Note that their movements would imitate the articulatory configurations of the trained speakers, as it has also been implemented with a similar idea in [57]. Then, in the production mapping, given the articulatory parameters, the estimated acoustic spectrum should be warped into the spectrum of the desired speaker. An alternative way to do speaker-independent mapping is by employing the idea of eigenvoice-based voice conversion [58]. In this framework, the desired speaker characteristics can be controlled with an optimum weight set that influence the eigenvoice parameters. Even with only a limited amount of adaptation speech data from the new speaker, the optimum

parameters can be easily obtained. In the implementation, the idea would be similar to that of the VTLN one, but instead of using the frequency warping procedure, we would use the voice conversion procedure.

VIII. Conclusion

In this paper, we presented a successful way of exploiting articulatory parameters through an articulatory controllable speech modification system with a sequential procedure of inversion and production mappings. GMM-based statistical feature mapping technique is employed for each of the acoustic-to-articulatory inversion mapping and the articulatory-to-acoustic production mapping. An input speech signal is modified through manipulation of the unobserved articulatory movements with the use of sequential inversion and production mappings. In controlling the movements of the articulators, a method that considers the intercorrelation of articulatory parameters is deployed after a simple linear transformation is performed. Furthermore, to alleviate the quality degradation of modified speech sounds, we apply several implementations of a direct waveform modification method that avoids the use of a vocoder-based excitation generation procedure. The experimental results demonstrate the following. 1) The sequential inversion and production mappings yield higher accuracy in estimating spectral envelope parameters, i.e., of 4.38 dB and 4.65 dB mel-cepstral distortions for male and female speakers, compared with 4.70 dB and 4.94 dB for a conventional production mapping that uses the measured articulatory parameters, respectively. 2) The method for manipulating articulatory parameters by considering their intercorrelation generates more natural results than a simple linear transformation method, i.e., an MOS score of 3.1 compared with a score of 2.1 for a twofold scaling value. 3) The implementations of the direct waveform modification method significantly improve the quality of the modified speech by avoiding the use of a vocoder-based excitation generation process and overcoming the oversmoothing problem by yielding MOS scores of above 3.5, while for the vocoder-based process, the MOS score was usually below 2.0. 4) The controllability of the system is ensured by its capability of allowing the modification of vowel sounds through a certain manipulation of articulatory configurations, giving averages of about 45% for correct perceptions of vowel /æ/, 85% for vowel /ɛ/, and 70% for vowel /ɪ/. In the future, we would like to develop a corresponding speaker-independent system [56]–[58] for employment in speech applications with interactive operation.

Appendix
Derivation of DiffGMM parameters in Eq. (16)

The likelihood function in Eq. (16) can be rewritten as

$$P(\boldsymbol{G}|\boldsymbol{V},\boldsymbol{\lambda}^{(Y,C)}) = \sum_{\text{all } \boldsymbol{m}^{(V)}} P(\boldsymbol{G},\boldsymbol{m}^{(V)}|\boldsymbol{V},\boldsymbol{\lambda}^{(Y,C)}),$$
$$\text{subject to } \boldsymbol{G} = \boldsymbol{C}' - \boldsymbol{C} = \boldsymbol{W}^{(c)}\boldsymbol{g},$$
$$\boldsymbol{V} = [\boldsymbol{Y}'^{\top}, \boldsymbol{Y}^{\top}]^{\top},$$
$$\text{and } \boldsymbol{m}^{(V)} = \{\{m_1, n_1\}, \ldots, \{m_T, n_T\}\}, \quad (21)$$

where

$$\begin{aligned} P(\boldsymbol{G},\boldsymbol{m}^{(V)}|\boldsymbol{V},\boldsymbol{\lambda}^{(Y,C)}) &= \prod_{t=1}^{T}\prod_{m=1}^{M}\prod_{n=1}^{M} P(\boldsymbol{G}_t, m, n|\boldsymbol{Y}'_t, \boldsymbol{Y}_t, \boldsymbol{\lambda}^{(Y,C)}) \\ &= \prod_{t=1}^{T}\prod_{m=1}^{M}\prod_{n=1}^{M} \mathcal{N}(\boldsymbol{G}_t; \boldsymbol{E}^{(G)}_{m,n,t}, \boldsymbol{D}^{(G)}_{m,n}) P(m|\boldsymbol{Y}', \boldsymbol{\lambda}^{(Y,C)}) \\ &\qquad P(n|\boldsymbol{Y}, \boldsymbol{\lambda}^{(Y,C)}). \end{aligned} \quad (22)$$

The conditional mean vector $\boldsymbol{E}^{(G)}_{m,n,t}$ and conditional covariance matrix $\boldsymbol{D}^{(G)}_{m,n}$ are written as

$$\boldsymbol{E}^{(G)}_{m,n,t} = \boldsymbol{E}^{(C)}_{m,t} - \boldsymbol{E}^{(C)}_{n,t} \quad (23)$$
$$\boldsymbol{D}^{(G)}_{m,n} = \boldsymbol{D}^{(C)}_{m} + \boldsymbol{D}^{(C)}_{n}, \quad (24)$$

where

$$\boldsymbol{E}^{(C)}_{m,t} = \boldsymbol{\mu}^{(C)}_m + \boldsymbol{\Sigma}^{(CY)}_m \boldsymbol{\Sigma}^{(YY)^{-1}}_m (\boldsymbol{Y}' - \boldsymbol{\mu}^{(Y)}_m) \quad (25)$$
$$\boldsymbol{E}^{(C)}_{n,t} = \boldsymbol{\mu}^{(C)}_n + \boldsymbol{\Sigma}^{(CY)}_n \boldsymbol{\Sigma}^{(YY)^{-1}}_n (\boldsymbol{Y} - \boldsymbol{\mu}^{(Y)}_n) \quad (26)$$
$$\boldsymbol{D}^{(C)}_{m} = \boldsymbol{\Sigma}^{(CC)}_m + \boldsymbol{\Sigma}^{(CY)}_m \boldsymbol{\Sigma}^{(YY)^{-1}}_m \boldsymbol{\Sigma}^{(YC)}_m. \quad (27)$$

Acknowledgment

Part of this work was supported by JST, PRESTO Grant Number JPMJPR1657, JSPS KAKENHI Grant Number 17H01763 and Grant-in-Aid for JSPS Research Fellow Number 16J10726.

References

[1] H. Dudley, "Remaking speech," *J. Acoust. Soc. Am.*, vol. 11, no. 2, pp. 169–177, 1939.
[2] G. Fant, *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*. Walter de Gruyter, 1971, vol. 2.
[3] M. M. Sondhi and J. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 35, no. 7, pp. 955–967, 1987.
[4] S. Parthasarathy, J. Schroeter, C. H. Coker, and M. M. Sondhi, "Articulatory analysis and synthesis of speech," in *TENCON '89. 4th IEEE Region 10 Int. Conf.*, Bombay, India, Nov. 1989, pp. 760–764.
[5] J. Schroeter and M. M. Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 1, pp. 133–150, 1994.
[6] J. Schroeter and M. M. Sondhi, "Speech coding based on physiological models of speech production," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds. New York: Marcel Dekker, 1992, pp. 231–267.
[7] M. M. Sondhi, "Articulatory modeling: a possible role in concatenative text-to-speech synthesis," in *IEEE 2002 Workshop on Speech Synthesis*, Santa Monica, USA, Sep. 2002, pp. 73–78.
[8] B. Bollepali, A. W. Black, and K. Prahallad, "Modeling a noisy-channel for voice conversion using articulatory features," in *Proc. INTERSPEECH*, Portland, USA, Sep. 2012, pp. 2202–2205.
[9] A. A. Wrench and K. Richmond, "Continuous speech recognition using articulatory data," in *Proc. ICSLP*, Beijing, China, Oct. 2000, pp. 145–148.
[10] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Commun.*, vol. 37, no. 3, pp. 303–319, 2002.
[11] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 121, no. 2, pp. 723–742, 2007.
[12] L. Revéret, G. Bailly, and P. Badin, "MOTHER: A new generation of talking heads providing a flexible articulatory control for video-realistic speech animation," in *Proc. ICSLP*, Beijing, China, Oct. 2000, pp. 755–758.

[13] P. Badin, F. Elisei, G. Bailly, and Y. Tarabalka, "An audiovisual talking head for augmented speech generation: Models and animations based on a real speakers articulatory data," in *Proc. Articulated Motion and Deformable Objects (AMDO)*. Mallorca, Spain: Berlin, Heidelberg: Springer, Jul. 2008, pp. 132–143.

[14] P. Mermelstein, "Determination of the vocal tract-shape from measured formant frequencies," *J. Acoust. Soc. Am.*, vol. 41, no. 5, pp. 1283–1294, 1967.

[15] H. Wakita, "Estimation of vocal-tract shapes from acoustical analysis of the speech wave: The state of the art," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 27, no. 3, pp. 281–285, 1979.

[16] B. S. Atal, J. J. Chang, M. V. Matthews, and J. W. Tukey, "Articulatory compensation: A study of ambiguities in the acoustic-articulatory mapping," *J. Acoust. Soc. Am.*, vol. 60, no. S1, p. S77, 1976.

[17] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marchal, Eds. Netherlands: Springer, 1990, pp. 131–149.

[18] P. Mermelstein, "Articulatory model for the study of speech production," *J. Acoust. Soc. Am.*, vol. 53, no. 4, pp. 1070–1082, 1973.

[19] B. S. Atal, J. J. Chang, M. V. Matthews, and J. W. Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," *J. Acoust. Soc. Am.*, vol. 63, no. 5, pp. 1535–1555, 1978.

[20] P. W. Schönle, K. Gräbe, P. Wenig, J. Höhne, J. Schrader, and B. Conrad, "Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," *Brain Lang.*, vol. 31, no. 1, pp. 26–35, 1987.

[21] A. A. Wrench and W. J. Hardcastle, "A multichannel articulatory database and its application for automatic speech recognition," in *Proc. 5th Seminar of Speech Prod.*, Kloster Seeon, Bavaria, Germany, May 2000, pp. 305–308.

[22] K. Richmond, P. Hoole, and S. King, "Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus," in *Proc. INTERSPEECH*, Florence, Italy, Aug. 2011, pp. 1505–1508.

[23] S. Narayanan, E. Bresch, P. K. Ghosh, L. Goldstein, A. Katsamanis, Y. Kim, A. C. Lammert, M. I. Proctor, V. Ramanarayanan, and Y. Zhu, "A multimodal real-time MRI articulatory corpus for speech research," in *Proc. INTERSPEECH*, Florence, Italy, Aug. 2011, pp. 837–840.

[24] J. Hogden, A. Lofqvist, V. Gracco, I. Zlokarnik, P. Rubin, and E. Saltzman, "Accurate recovery of articulator positions from acoustics: New conclusions based on human data," *J. Acoust. Soc. Am.*, vol. 100, no. 3, pp. 1819–1834, 1996.

[25] S. Suzuki, T. Okadome, and M. Honda, "Determination of articulatory positions from speech acoustics by applying dynamic articulatory constraints," in *Proc. ICSLP*, Sydney, Australia, Dec. 1998, pp. 2251–2254.

[26] K. Richmond, S. King, and P. Taylor, "Modelling the uncertainty in recovering articulation from acoustics," *Comput. Speech Lang.*, vol. 17, no. 2–3, pp. 153–172, 2003.

[27] K. Richmond, "Trajectory mixture density networks with multiple mixtures for acoustic-articulatory inversion," in *Proc. Int. Conf. Non-Linear Speech Process. (NOLISP)*, Paris, France, May 2007, pp. 263–272.

[28] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an HMM-based speech production model," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 2, pp. 175–185, 2004.

[29] A. B. Youssef, G. B. Pierre Badin, and P. Heracleous, "Acoustic-to-articulatory inversion using speech recognition and trajectory formation based on phoneme hidden Markov models," in *Proc. INTERSPEECH*, Brighton, United Kingdom, Sep. 2009, pp. 2255–2258.

[30] T. Toda, A. W. Black, and K. Tokuda, "Acoustic-to-articulatory inversion mapping with Gaussian mixture model," in *Proc. INTERSPEECH*, Jeju, Korea, Oct. 2004, pp. 1129–1132.

[31] T. Kaburagi and M. Honda, "Determination of the vocal tract spectrum from the articulatory movements based on the search of an articulatory-acoustic database," in *Proc. ICSLP*, Sydney, Australia, Dec. 1998, pp. 433–436.

[32] C. T. Kello and D. C. Plaut, "A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters," *J. Acoust. Soc. Am.*, vol. 116, no. 4, pp. 2354–2364, 2004.

[33] S. Aryal and R. Gutierrez-Osuna, "Data driven articulatory synthesis with deep neural networks," *Comput. Speech Lang.*, vol. 36, pp. 260–273, Mar. 2016.

[34] K. Nakamura, T. Toda, Y. Nankaku, and K. Tokuda, "On the use of phonetic information for mapping from articulatory movements to vocal tract spectrum," in *Proc. ICASSP*, Toulouse, France, May 2006, pp. 93–96.

[35] T. Toda, A. W. Black, and K. Tokuda, "Mapping from articulatory movements to vocal tract spectrum with Gaussian mixture model for articulatory speech synthesis," in *5th ISCA Tutorial and Research Workshop on Speech Synthesis*, Pittsburgh, USA, Jun. 2004, pp. 31–36.

[36] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Commun.*, vol. 50, no. 3, pp. 215–227, 2008.

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

[38] Z.-H. Ling, K. Richmond, and J. Yamagishi, "Articulatory control of HMM-based parametric speech synthesis using feature-space-switched multiple regression," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 21, no. 1, pp. 207–219, 2013.

[39] B. J. Kröger, J. Gotto, S. Albert, and C. Neuschaefer-Rube, "A visual articulatory model and its application to therapy of speech disorders: a pilot study," *Speech production and perception: Experimental analyses and models. ZAS Papers in Linguistics*, vol. 40, pp. 79–94, 2005.

[40] B. J. Kröger, V. Graf-Borttscheller, and A. Lowit, "Two- and three-dimensional visual articulatory models for pronunciation training and for treatment of speech disorders," in *Proc. INTERSPEECH*, Brisbane, Australia, 2008, pp. 2639–2642.

[41] D. W. Massaro, "The psychology and technology of talking heads: Applications in language learning," in *Advances in Natural Multimodal Dialogue Systems*. Netherlands: Springer, 2005, vol. 30, pp. 183–214.

[42] B. J. Kröger, P. Birkholz, R. Hoffmann, and H. Meng, "Audiovisual tools for phonetic and articulatory visualization in computer-aided pronunciation training," in *Development of Multimodal Interfaces: Active Listening and Synchrony*. Berlin, Heidelberg: Springer, 2010, pp. 337–345.

[43] P. L. Tobing, T. Toda, G. Neubig, S. Sakti, S. Nakamura, and A. Purwarianti, "Articulatory controllable speech modification based on statistical feature mapping with Gaussian mixture models," in *Proc. INTERSPEECH*, Singapore, Sep. 2014, pp. 2298–2302.

[44] 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 *Proc. INTERSPEECH*, Singapore, Sep. 2014, pp. 2514–2518.

[45] P. L. Tobing, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Articulatory controllable speech modification based on Gaussian mixture models with direct waveform modification using spectrum differential," in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 3350–3354.

[46] A. Wrench. (1999) The MOCHA-TIMIT articulatory database. Queen Margaret University College. [Online]. Available: http://www.cstr.ed.ac.uk/artic/mocha.html

[47] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, 2007.

[48] B. Youssef, A. Badin, and G. Bailly, "Can tongue be recovered from face? The answer of data-driven statistical models," in *Proc. INTERSPEECH*, Makuhari, Japan, Sep. 2010, pp. 2002–2005.

[49] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. ICASSP*, Detroit, USA, Sep. 1995, pp. 660–663.

[50] S. Takamichi, T. Toda, A. W. Black, G. Neubig, S. Sakti, and S. Nakamura, "Postfilters to modify the modulation spectrum for statistical parametric speech synthesis," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 4, pp. 755–767, 2016.

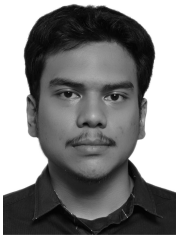
[51] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Electron. Commun. Jpn. (Part I: Commun.)*, vol. 66, no. 2, pp. 10–18, 1983.

[52] K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Statistical singing voice conversion based on direct waveform modification with global variance," in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 2754–2758.

[53] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representation using a pitch-adaptive time frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, no. 3–4, pp. 187–207, 1999.

[54] H. Kawahara, H. Katayose, A. de Cheveigné, and R. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," in *Proc. EUROSPEECH*, Budapest, Hungary, Sep. 1999, pp. 2781–2784.

[55] B. Picart, T. Drugman, and T. Dutoit, "Analysis and synthesis of hypo- and hyperarticulated speech," in *7th ISCA Tutorial and Research Workshop on Speech Synthesis*, Kyoto, Japan, Sep. 2010, pp. 270–275.
[56] G. Sivaraman, V. Mitra, H. Nam, M. K. Tiede, and C. Y. Espy-Wilson, "Vocal tract length normalization for speaker independent acoustic-to-articulatory speech inversion," in *INTERSPEECH*, San Francisco, USA, 2016, pp. 455–459.
[57] P. K. Ghosh and S. S. Narayanan, "A subject-independent acoustic-to-articulatory inversion," in *Proc. ICASSP*, Prague, Czech Republic, 2011, pp. 4624–4627.
[58] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on Gaussian mixture model," in *INTERSPEECH*, Pittsburgh, USA, 2006, pp. 2446–2449.



Patrick Lumban Tobing received his B.E. degree from Bandung Institute of Technology (ITB), Indonesia, in 2014 and his M.E degree from Nara Institute of Science and Technology (NAIST), Japan, in 2016. He is currently enrolled in a Ph.D. course at the Graduate School of Information Science, Nagoya University, Japan. He received a Best Student Presentation Award from the Acoustical Society of Japan (ASJ). He is a student member of IEEE, ISCA, and ASJ.

Kazuhiro Kobayashi received his B.E. degree from the Department of Electrical and Electronic Engineering, Faculty of Engineering Science, Kansai University, Japan, in 2012, and his M.E. and Ph.D. degree from Nara Institute of Science and Technology (NAIST), Japan, in 2014 and 2017, respectively. He is currently working as a postdoctoral researcher at the Graduate School of Information Science, Nagoya University, Japan. He has received a few awards including a Best Presentation Award from the Acoustical Society of Japan (ASJ). He is a student member of IEEE, ISCA, and ASJ.

Tomoki Toda received his B.E. degree from Nagoya University, Japan, in 1999 and his M.E. and D.E. degrees from Nara Institute of Science and Technology (NAIST), Japan, in 2001 and 2003, respectively. He was a Research Fellow of the Japan Society for the Promotion of Science from 2003 to 2005. He was then an Assistant Professor (2005-2011) and an Associate Professor (2011-2015) at NAIST. Since 2015, he has been a Professor in the Information Technology Center at Nagoya University. His research interests include statistical approaches to speech and audio processing. He has received more than 10 paper/achievement awards including the IEEE SPS 2009 Young Author Best Paper Award and the 2013 EURASIP-ISCA Best Paper Award (Speech Communication Journal).