NAIST-IS-MT1451207

**Master's Thesis**

# Articulatory Controllable Speech Modification using Statistical Feature Mapping Techniques

Patrick Lumban Tobing

August 8, 2016

Graduate School of Information Science
Nara Institute of Science and Technology

A Master's Thesis
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
MASTER of ENGINEERING

Patrick Lumban Tobing

Thesis Committee:

| | |
|---|---|
| Professor Satoshi Nakamura | (Supervisor) |
| Professor Kenji Sugimoto | (Co-supervisor) |
| Professor Tomoki Toda | (Co-supervisor/Nagoya University) |
| Assistant Professor Sakriani Sakti | (Co-supervisor) |

# Articulatory Controllable Speech Modification using Statistical Feature Mapping Techniques*

Patrick Lumban Tobing

**Abstract**

Speech is one of the most universal way for people to communicate with each other. In the creation of speech sounds, our articulators, such as tongue and lips, play an essential role in determining the resonance characteristics of the vocal tract. The traits of these speech organs, in fact, are easy to be perceived in comprehending the speech production process. Therefore, by imposing the use of articulators, one may develop intuitive and perceptive speech applications, e.g. for assisting speech-disabled people, for supplementary tools in language-learning sessions, and many others. In this thesis, in order to lay a groundwork towards the development of such applications, we propose an articulatory controllable speech modification method based on statistical feature mapping techniques, such as the Gaussian mixture model (GMM)-based acoustic-to-articulatory inversion mapping and the GMM-based articulatory-to-acoustic production mapping. These GMM-based statistical feature mapping techniques possess invaluable attributes by having low-cost development resources, clear and convenient training–conversion schemes, and independency of any language specification textual features. To maximize their potentials, we propose a sequential mapping procedure which enables a speech modification mechanism by manipulation of the unobserved articulatory movements from an input speech sound. We also deploy a method for controlling movements of the articulators by considering their inter-correlations to produce more natural modification outcomes. Additionally, to alleviate the degradation of speech quality in a vocoder-based

*Master's Thesis, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-MT1451207, August 8, 2016.

excitation generation process, we apply a direct waveform modification process that directly filters an input speech signal according to the differences of spectrum between modified speech and the original one. The experimental results demonstrate that: 1) the proposed sequential mapping between acoustic spectrum and articulatory movements yields higher production mapping accuracy than the conventional production procedure using the measured articulatory parameters, 2) the proposed method for controlling articulatory movements makes it possible to generate more natural modified speech sounds, 3) the proposed speech generation schema based on direct waveform modification significantly improves the quality of modified speech sounds, and 4) the controllability of the proposed system has been affirmed by its capability of producing different sounds of vowel by means of handling the configuration of particular articulatory positions.

**Keywords:**

articulatory control, speech modification, Gaussian mixture model, statistical feature mapping, inter-correlations of articulators, direct waveform modification

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1. Background

Speech is the most common way for people in communicating with each other. Indeed, through speech, almost every contents of our mind can be expressed and conveyed naturally. The proficiency of producing proper speech, thus, be one of the fundamental facets in the human relation.

To produce speech sounds, first, our lungs build up air pressure, where the resulting airflow would be vibrated by our vocal folds. Then, our speech organs act in such a way, so that the resulting excitation sounds are modulated to produce particular speech sounds. These speech organs or the so called articulators, such as tongue and lips, are the ones that determine the resonance characteristics of the vocal tract which in turn dictate the phonetical quality of the resulting speech sounds. An illustration of the mechanism of speech production is shown in Fig. 1.1.

Moving a little bit further, in this fast-paced computing era, technology has been permeating into practically every aspects of human life. This encompasses also the aspect of human communication, especially in terms of speech technology. It can be denied that, right now, the use of speech technology is unavoidable in our life. Henceforth, effective and efficient way of speech-related technologies development becomes one of the crucial feature to keeping up with the advancements of automation system in our life.

Employment of speech technology often suggests a kind of sophisticated and

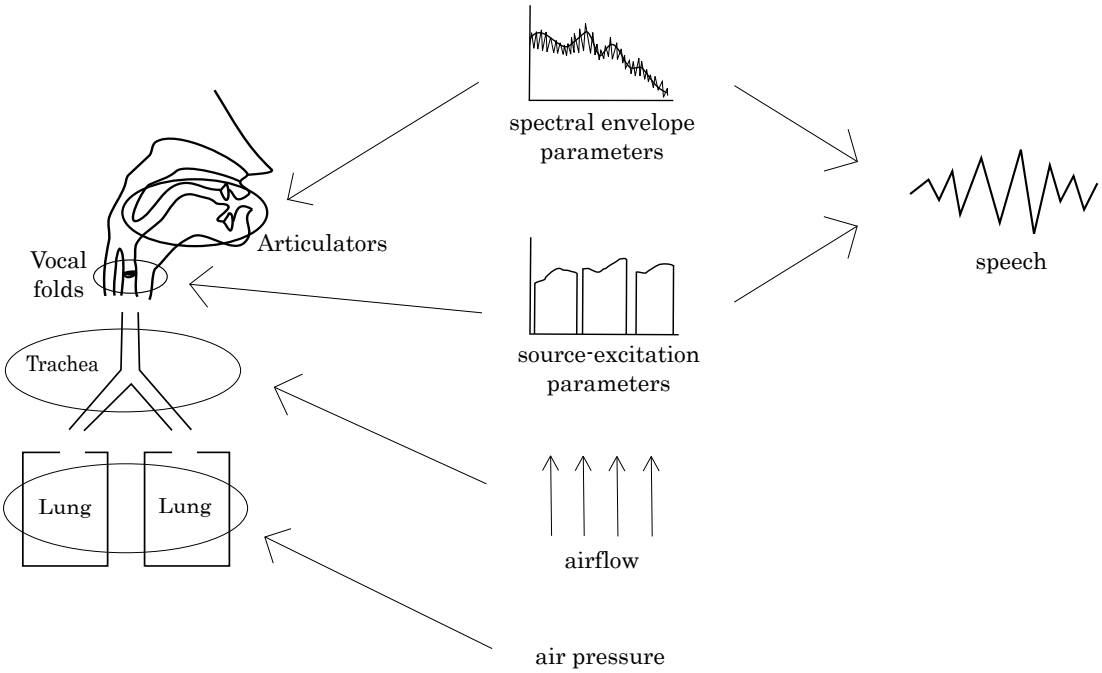



Figure 1.1. Illustration of human speech production mechanism.

cryptic system. However, if we take a look at the natural process of speech production, we can easily attempt to make use of the dependency between the creation of speech sounds and the traits of our speech organs. To be able to automatically perform computation of speech signal, it is parameterized generally by the utilization of the spectrum of the vocal tract. Though, it is also possible, by knowing the fact that our articulators take a vital part in the determination of vocal tract spectrum, to use the more slowly varying parameters, i.e. the articulatory parameters [1]. This, of course, would lead to an overally more intuitive and perceptive speech technology. In fact, the relationship between speech signal and the articulators has been studied in many works, such as in speech synthesis [2, 3], speech recognition [4, 5], and speech coding [6].

Development of a system that is capable of taking advantage of the intuitiveness of articulatory representation is one of the most vital contribution in the advancements of speech technology. This is due to the nature of the speech production which cannot be separated from the quality of the articulation. As a matter of fact, the visualization of articulatory movements has been showing

promising potentials in the creation of a system to help speech-disabled people in speech therapy sessions [7, 8], as supplementary tools for language learning program [9, 10, 11], and as a vital component in helping non-native speaker to correct their accent in pronunciation learning procedure [12, 13, 14]. Overall, it has been summarized in [15] that currently, the use of articulatory-aid in speech technology is one of the fundamental needs. However, a system which enables us to flexibly tinker our articulatory movements from our own speech generation procedure has not yet been developed. Such kind of mechanism would be very useful not only for the use of assistances in daily life, as have been elaborated, but also, of course, as an invaluable tool in research work.

## 1.2. Related Work

In accomplishing the automation of the employment of articulatory elements in speech applications, first, we need to define the procedure for relationship mapping between acoustic and articulatory parameters. The mapping process from acoustic to articulatory parameters is well known as the inversion mapping. On the other hand, the opposite procedure, i.e. mapping from articulatory to acoustic parameters is often named as the production mapping. To establish these mapping relationships, formerly, a mathematical production model was used [6, 16]. However, the speech production process cannot be modeled without some approximations. Lately, thanks to advancements of recording devices which enable us to simultaneously record speech sounds and articulatory data, the availability of parallel-data of speech and articulatory movements has been growing rapidly. This contributes to a lot of recent works in data-driven based mapping techniques which is capable of capturing statistical traits between the acoustic and articulatory data rather than mathematically modeling them.

Some works on statistical data-driven feature mapping for inversion mapping have been studied and published. In [17], a codebook based mapping approach has been proposed as the initial work on data-driven inversion mapping. In [18], the incorporation of dynamic features capturing the temporal patterns of acoustic spectrum significantly improves the accuracy in estimating articulatory positions with a codebook-based approach. Then, in [19], an inversion mapping method us-

Table 1.1. Comparison of several statistical feature mapping techniques.

| trait \ method | Neural network | HMM | GMM |
|---|---|---|---|
| accuracy | excellent | very good | very good |
| computation load | very large | small | small |
| data requirement | very large | small | small |
| language dependent | no | yes | no |
| parameter characteristics | non-interpretable | interpretable | interpretable |

ing neural network based on mixture density estimation has been examined showing the importance of multiple representation of articulatory probability density. Hidden Markov model (HMM)-based inversion mapping has also been studied in [20, 21], by using the inclusion of language specification features. And in [22], the Gaussian mixture model (GMM)-based inversion mapping has been proposed enabling efficient and effective inversion mapping procedure without the use of phonetic features.

Several noted works on the articulatory-to-acoustic production mapping have also been proposed in a similar chronological fashion as with the inversion mapping. In [23], the vocal tract spectrum can be estimated from the articulatory data with a searching procedure through acoustic-articulatory data with the inclusion of also phonetical categorization. Another neural network based method has also been studied for the production mapping task, in [24]. In [25], harmonic features in multiple frames have been used to statistically performing better estimation of acoustic spectrum from articulatory parameters. The HMM-based method has also been studied in [26], to incorporate not only the phonetical information but also temporal patterns in spectrum estimation from articulatory configurations. Then, in [27], the GMM-based approach is used to alleviate the needs of any textual input in generating speech spectra from articulatory movements. The trade-off between the related techniques that have been studied is shown in Table 1.1.

In this thesis, we focus on the statistical feature mapping approach with Gaussian mixture model (GMM), which has been studied and summarized for both the acoustic-to-articulatory inversion mapping and the articulatory-to-acoustic

production mapping in [28]. The GMM mapping technique is widely known for its pioneering use in the voice conversion mechanism [29]. The employment of this method has been used widely, as in the more refined voice conversion system [30], speech enhancement for laryngectomee patients [31], statistical singing voice conversion [32], and non-audible murmur to speech conversion [33]. These are the examples of the extensive use of GMM-based statistical feature mapping. Talking back again about a system for speech and articulators, an HMM-based system has recently been proposed to enabling a control of articulatory parameters in a text-to-speech (TTS) system [34], but without any comprehensive manipulation method of the articulatory features. Moreover, reviewing over the motivation in developing a system based on speech and articulatory movements, it cannot be denied that the independency of any language characteristics would be one of the fundamental element to create a flexible and widely applicable applications.

The GMM-based method is known widely as a generative model, as also with the HMM-based, whereas the NN (neural-net)-based is categorized as a discriminative model. This, supported by the fact that generative models can elegantly capture the statistical traits between joint input and output data compared with discriminative ones that try to create predictive model of output given input data, [35, 36] would prove to be a significant advantage in the case of flexibility for adapting the model into various tasks. Moreover, due to the characteristic of NN-based mapping which abstractly captures the contribution between features of data, through their weights, without providing meaningful interpretations of correlations between each features [37], it has a very conflictive trait with the spirit in developing speech-articulatory related applications. Therefore, deployment of the GMM-based statistical feature mapping technique for efficient utilization of articulatory movements in a speech-related system, such as a robust and flexible speech modification system, would be an essential contribution for speech-technology development.

## 1.3. Thesis Scope

In this thesis, we propose an articulatory controllable speech modification system using statistical feature mappings. Specifically, we employ the Gaussian mixture

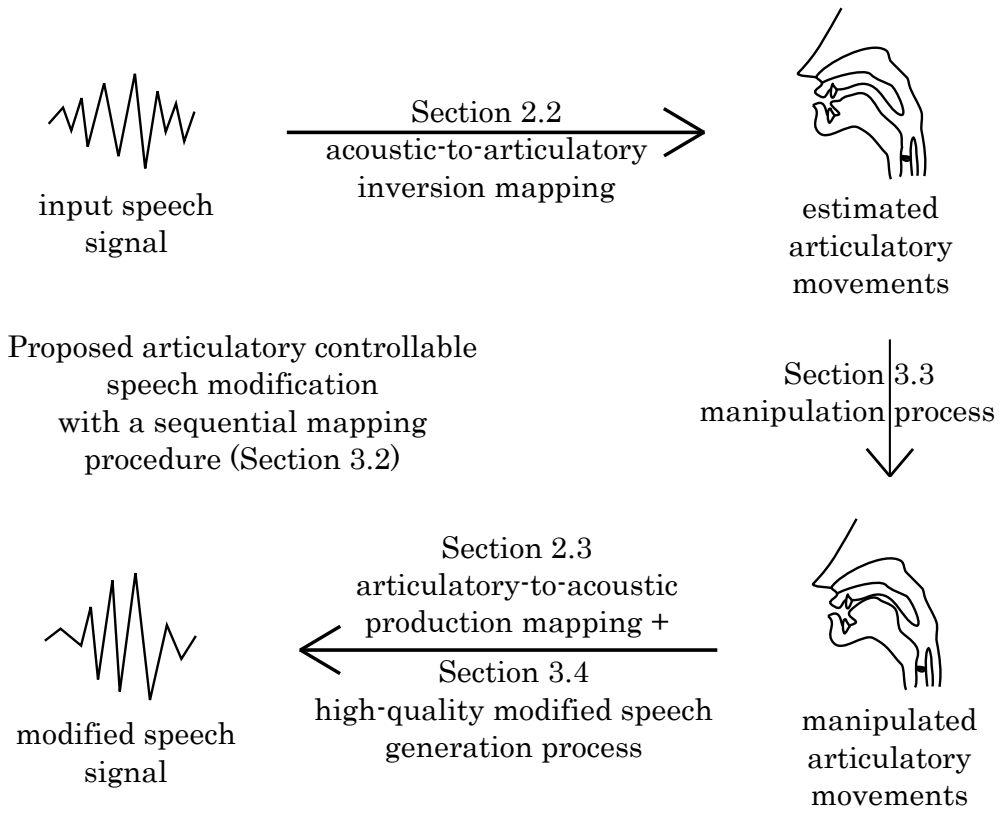



Figure 1.2. Flow of the proposed articulatory controllable speech modification system by sequentially integrating the acoustic-articulatory mappings with inclusion of articulatory control and high-quality modified speech generation process.

model (GMM)-based acoustic-to-articulatory inversion mapping and the GMM-based articulatory-to-acoustic production mapping. The overview of the proposed system is illustrated in Fig. 1.2. These inversion and production mappings are integrated in a sequential mapping scheme to enable an intuitive speech modification procedure via manipulation of the unobserved articulatory movements of an input speech signal. In the proposed system, in order to administer fine results from the manipulation of articulatory movements, we deploy a method for controlling the articulatory parameters through consideration of their inter-correlations. In addition, to guarantee high-quality modified speech sounds, we bypass the use of vocoder-based excitation generation process, in synthesizing

the speech signal, through the use of direct waveform modification method which directly filters an input speech waveform according to the spectrum differences between modified speech and original one. The experimental results demonstrate that the proposed system is capable of producing high-quality modified speech sounds through the use of convenient manipulation practice of the articulatory organs.

## 1.4. Thesis Overview

This thesis is organized as follows. In chapter 2, the GMM-based statistical feature mappings between acoustic and articulatory parameters are explained. In chapter 3, the proposed articulatory controllable speech modification system is described. In chapter 4, the experimental evaluation results are given. Finally, chapter 5 presents summary of this thesis and the future work.

# Chapter 2

# Statistical Feature Mapping between Acoustic and Articulatory Parameters with Gaussian Mixture Model

## 2.1. Introduction

In this thesis, two main mapping systems using Gaussian mixture model (GMM) are deployed, i.e. acoustic-to-articulatory inversion mapping and articulatory-to-acoustic production mapping. In order to develop these systems, two processes are required to be established in each of the mapping system, i.e. training and conversion process. In the training process, the joint probability density function of source and target features is modeled with a GMM. In the conversion process, given the source features and trained model parameters, target features are determined by utilizing the conditional probability function derived from the joint GMM.

Regarding the acoustic and articulatory parameters, in this thesis, the mel-cepstrum parameters representing the spectral envelope are used for the acoustic parameters. On the other hand, the EMA data representing the articulatory movements are utilized as the articulatory parameters. These features are ex-

plained in details in the section 4.1

This chapter is organized as follows. The GMM-based acoustic-to-articulatory inversion mapping is described in section 2.2. Specifically, its training and conversion process are given in sections 2.2.1 and 2.2.2, respectively. In the section 2.3, the GMM-based articulatory-to-acoustic inversion mapping is presented, along with its training and conversion process in sections 2.3.1 and 2.3.2. Finally, this chapter is summarized in section 2.5.

## 2.2. GMM-based Acoustic-to-Articulatory Inversion Mapping

In the GMM-based acoustic-to-articulatory inversion mapping system, the source features are composed by the acoustic parameters, while the target features are composed by the articulatory parameters. First, let us define a $D_c$-dimensional feature vector of the acoustic parameters, i.e. the mel-cepstrum, as $\boldsymbol{c}_t$ and a $D_x$-dimensional feature vector of the articulatory parameters as $\boldsymbol{x}_t$ at frame $t$. Then, the time sequence vector of acoustic parameters and that of the articulatory parameters are respectively written as

$$\boldsymbol{c} = [\boldsymbol{c}_1^\top, \boldsymbol{c}_2^\top, \ldots, \boldsymbol{c}_t^\top, \ldots, \boldsymbol{c}_T^\top]^\top, \tag{2.1}$$

$$\boldsymbol{x} = [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \ldots, \boldsymbol{x}_t^\top, \ldots, \boldsymbol{x}_T^\top]^\top, \tag{2.2}$$

where the superscript $^\top$ indicates transposition.

As the source features, mel-cepstral segment feature vectors are developed by extracting the mel-cepstrum parameters at multiple frames around the current frame. Let us define the time sequence of mel-cepstral segment feature vectors $\boldsymbol{O}$, which is written as

$$\boldsymbol{O} = [\boldsymbol{O}_1^\top, \boldsymbol{O}_2^\top, \ldots, \boldsymbol{O}_t^\top, \ldots, \boldsymbol{O}_T^\top]^\top. \tag{2.3}$$

At frame $t$, the mel-cepstral segment feature vector $\boldsymbol{O}_t$ is then given by

$$\boldsymbol{O}_t = \boldsymbol{A}[\boldsymbol{c}_{t-L}^\top, \ldots, \boldsymbol{c}_t^\top, \ldots, \boldsymbol{c}_{t+L}^\top]^\top + \boldsymbol{b}, \tag{2.4}$$

where $L$ is the length of the context-window of a segment. The principal component analysis (PCA) method [38] is used to determine the transformation matrix $\boldsymbol{A}$ and $\boldsymbol{b}$ with training data beforehand.

As the target features, let us define the time sequence of joint static and dynamic feature vectors of articulatory parameters $\boldsymbol{X}$, which is written as

$$\boldsymbol{X} = [\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top, \ldots, \boldsymbol{X}_t^\top, \ldots, \boldsymbol{X}_T^\top]^\top. \tag{2.5}$$

Here, the $2D_x$-dimensional joint static and dynamic feature vector of articulatory parameters is denoted as $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$ at frame $t$. The dynamic feature vector $\Delta\boldsymbol{x}_t$ is computed from the static feature vector $\boldsymbol{x}_t$, which is given by

$$\Delta\boldsymbol{x}_t = \sum_{\tau=-L_-^{(1)}}^{L_+^{(1)}} w^{(1)}(\tau)\boldsymbol{x}_{t+\tau}, \tag{2.6}$$

where $w^{(1)}(\tau)$, $L_-^{(1)}$, and $L_+^{(1)}$ are the 1-st order weight coefficients and the frame lengths for computing the dynamic feature vector.

### 2.2.1 Training process

Schematic flow of the training process for inversion mapping is shown in Fig. 2.1. From the training data, a joint source and target feature vector $[\boldsymbol{O}_t^\top, \boldsymbol{X}_t^\top]^\top$ is developed at each frame $t$. Then, its joint probability density function is modeled with the GMM for the inversion mapping as follows:

$$\begin{aligned} P(\boldsymbol{O}_t, \boldsymbol{X}_t|\boldsymbol{\lambda}^{(O,X)}) &= \sum_{m=1}^{M} P(\boldsymbol{O}_t, \boldsymbol{X}_t|m, \boldsymbol{\lambda}^{(O,X)}) \\ &= \sum_{m=1}^{M} \alpha_m^{(O,X)} \mathcal{N}([\boldsymbol{O}_t^\top, \boldsymbol{X}_t^\top]^\top; \boldsymbol{\mu}_m^{(O,X)}, \boldsymbol{\Sigma}_m^{(O,X)}), \end{aligned} \tag{2.7}$$

where the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ is denoted as $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Here, the mixture component index is denoted as $m$ and the total number of mixture components is $M$. The set of the GMM parameters for inversion mapping is denoted as $\boldsymbol{\lambda}^{(O,X)}$, which consists of weights $\alpha_m^{(O,X)}$, mean vectors $\boldsymbol{\mu}_m^{(O,X)}$ and covariance matrices $\boldsymbol{\Sigma}_m^{(O,X)}$ of individual mixture components. The mean vector $\boldsymbol{\mu}_m^{(O,X)}$ and covariance matrix $\boldsymbol{\Sigma}_m^{(O,X)}$ of the $m$th mixture component are written as

$$\boldsymbol{\mu}_m^{(O,X)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(O)} \\ \boldsymbol{\mu}_m^{(X)} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(O,X)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(OO)} & \boldsymbol{\Sigma}_m^{(OX)} \\ \boldsymbol{\Sigma}_m^{(XO)} & \boldsymbol{\Sigma}_m^{(XX)} \end{bmatrix}, \tag{2.8}$$

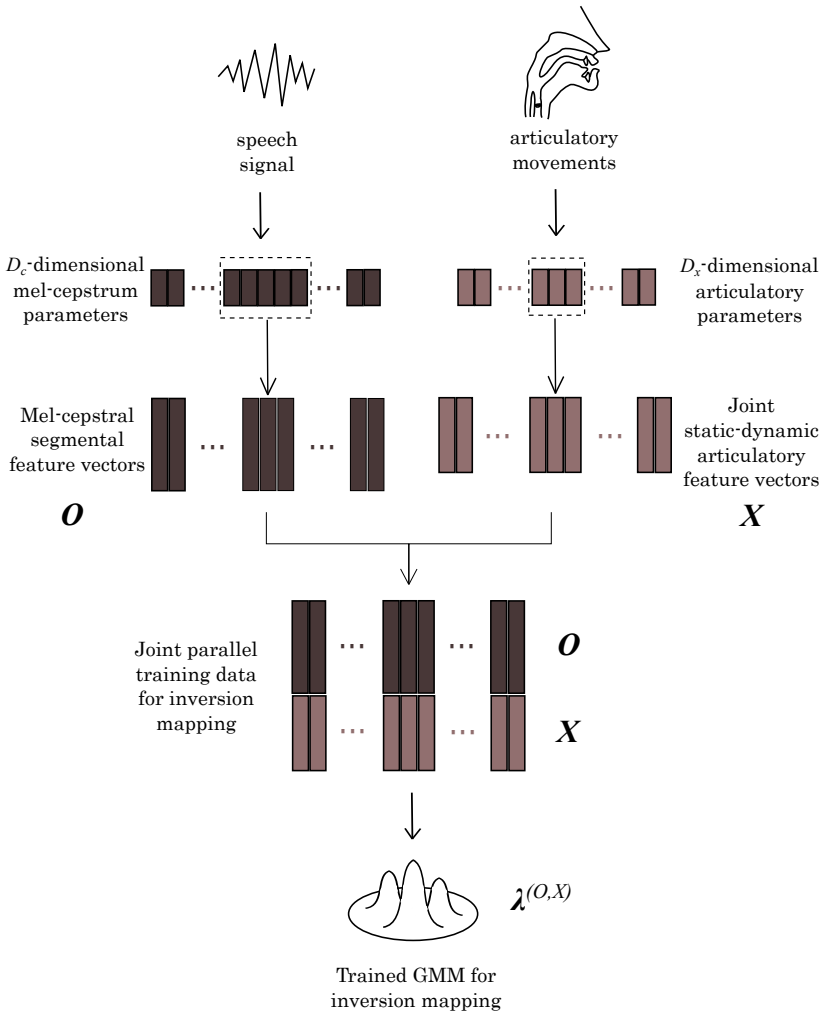



Figure 2.1. Schematic flow of the training process for GMM-based inversion mapping.

where the mean vectors of the acoustic parameters and that of the articulatory parameters for the $m$th mixture component are denoted as $\boldsymbol{\mu}_m^{(O)}$ and $\boldsymbol{\mu}_m^{(X)}$, respectively. The covariance matrix of the acoustic parameters is denoted as $\boldsymbol{\Sigma}_m^{(OO)}$ and that of the articulatory parameters is denoted as $\boldsymbol{\Sigma}_m^{(XX)}$ for the $m$th mixture component. The cross covariance matrices of the acoustic and articulatory parameters for the $m$th mixture components are denoted as $\boldsymbol{\Sigma}_m^{(OX)}$ and $\boldsymbol{\Sigma}_m^{(XO)}$. All of these covariance matrices are full.

In order to train the GMM parameters of the inversion mapping $\boldsymbol{\lambda}^{(O,X)}$, the

following likelihood function is to be maximized

$$P(\boldsymbol{O}, \boldsymbol{X}|\boldsymbol{\lambda}^{(O,X)}) = \sum_{\text{all } \boldsymbol{m}} P(\boldsymbol{m}|\boldsymbol{O}, \boldsymbol{X}, \boldsymbol{\lambda}^{(O,X)}) P(\boldsymbol{O}, \boldsymbol{X}, |\boldsymbol{m}, \boldsymbol{\lambda}^{(O,X)})$$
$$= \prod_{t=1}^{T} \sum_{m=1}^{M} P(m|\boldsymbol{O}_t, \boldsymbol{X}_t, \boldsymbol{\lambda}^{(O,X)}) P(\boldsymbol{O}_t, \boldsymbol{X}_t|m, \boldsymbol{\lambda}^{(O,X)}), \quad (2.9)$$

where $\boldsymbol{m} = \{m_1, m_2, \ldots, m_t, \ldots, m_T\}$ is a mixture component sequence.

Utilizing the Expectation-Maximization (EM) algorithm [39], given a set of initial parameters $\boldsymbol{\lambda}^{(O,X)}$, a set of updated parameters $\hat{\boldsymbol{\lambda}}^{(O,X)}$ is estimated with respect to the following auxiliary function

$$Q(\hat{\boldsymbol{\lambda}}^{(O,X)}, \boldsymbol{\lambda}^{(O,X)}) = \sum_{\text{all } \boldsymbol{m}} P(\boldsymbol{m}|\boldsymbol{O}, \boldsymbol{X}, \boldsymbol{\lambda}^{(O,X)}) \log P(\boldsymbol{O}, \boldsymbol{X}, |\boldsymbol{m}, \hat{\boldsymbol{\lambda}}^{(O,X)}). \quad (2.10)$$

In the expectation step (E-step), the occupancies $\gamma_{m,t}^{(O,X)}$ of each mixture components for each frames and the total number of samples belonging to each mixture components $N_m^{(O,X)}$ are given by

$$\gamma_{m,t}^{(O,X)} = P(m|\boldsymbol{O}_t, \boldsymbol{X}_t, \boldsymbol{\lambda}^{(O,X)})$$
$$= \frac{\alpha_m \mathcal{N}([\boldsymbol{O}_t^\top, \boldsymbol{X}_t^\top]^\top; \boldsymbol{\mu}_m^{(O,X)}, \boldsymbol{\Sigma}_m^{(O,X)})}{\sum_{n=1}^{M} \alpha_n \mathcal{N}([\boldsymbol{O}_t^\top, \boldsymbol{X}_t^\top]^\top; \boldsymbol{\mu}_n^{(O,X)}, \boldsymbol{\Sigma}_n^{(O,X)})}, \quad (2.11)$$

$$N_m^{(O,X)} = \sum_{t=1}^{T} \gamma_{m,t}^{(O,X)}. \quad (2.12)$$

Then, in the maximization step (M-step), the updated parameters in the set $\hat{\boldsymbol{\lambda}}^{(O,X)}$ would be finally given by

$$\hat{\alpha}_m^{(O,X)} = \frac{N_m^{(O,X)}}{T} \quad (2.13)$$

$$\hat{\boldsymbol{\mu}}_m^{(O,X)} = \frac{1}{N_m^{(O,X)}} \sum_{t=1}^{T} \gamma_{m,t}^{(O,X)} [\boldsymbol{O}_t^\top, \boldsymbol{X}_t^\top]^\top \quad (2.14)$$

$$\hat{\boldsymbol{\Sigma}}_m^{(O,X)} = \frac{1}{N_m^{(O,X)}} \sum_{t=1}^{T} \gamma_{m,t}^{(O,X)} ([\boldsymbol{O}_t^\top, \boldsymbol{X}_t^\top]^\top - \hat{\boldsymbol{\mu}}_m^{(O,X)})([\boldsymbol{O}_t^\top, \boldsymbol{X}_t^\top]^\top - \hat{\boldsymbol{\mu}}_m^{(O,X)})^\top. \quad (2.15)$$

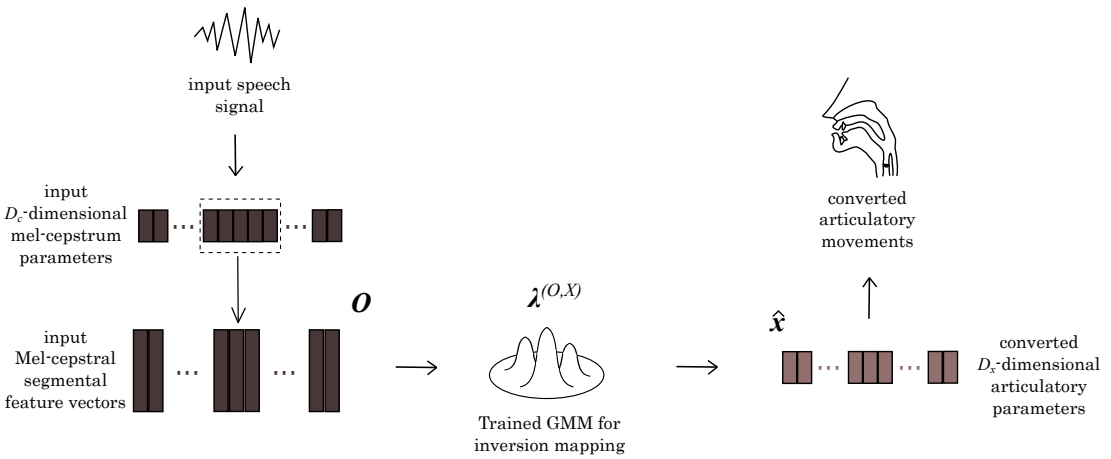



Figure 2.2. Schematic flow of the conversion process for GMM-based inversion mapping.

### 2.2.2 Conversion process

Schematic flow of the conversion process for inversion mapping is shown in Fig. 2.2. Given a time sequence of mel-cepstral segment feature vectors $\boldsymbol{O}$ and the trained GMM parameters of the inversion mapping $\boldsymbol{\lambda}^{(O,X)}$, a time sequence vector of the corresponding articulatory parameters $\hat{\boldsymbol{x}}$ is determined by

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} P(\boldsymbol{X}|\boldsymbol{O}, \boldsymbol{\lambda}^{(O,X)}), \tag{2.16}$$

where

$$\begin{aligned} P(\boldsymbol{X}|\boldsymbol{O}, \boldsymbol{\lambda}^{(O,X)}) &= \sum_{\text{all } \boldsymbol{m}} P(\boldsymbol{m}|\boldsymbol{O}, \boldsymbol{\lambda}^{(O,X)}) P(\boldsymbol{X}|\boldsymbol{O}, \boldsymbol{m}, \boldsymbol{\lambda}^{(O,X)}) \\ &= \prod_{t=1}^{T} \sum_{m=1}^{M} P(m|\boldsymbol{O}_t, \boldsymbol{\lambda}^{(O,X)}) P(\boldsymbol{X}_t|\boldsymbol{O}_t, m, \boldsymbol{\lambda}^{(O,X)}). \end{aligned} \tag{2.17}$$

The $m$-th posterior probability density $P(m|\boldsymbol{O}_t, \boldsymbol{\lambda}^{(O,X)})$ and the $m$-th conditional probability density $P(\boldsymbol{X}_t|\boldsymbol{O}_t, m, \boldsymbol{\lambda}^{(O,X)})$ at each frame $t$ are given by

$$P(m|\boldsymbol{O}_t, \boldsymbol{\lambda}^{(O,X)}) = \frac{\alpha_m \mathcal{N}(\boldsymbol{O}_t; \boldsymbol{\mu}_m^{(O)}, \boldsymbol{\Sigma}_m^{(OO)})}{\sum_{n=1}^{M} \alpha_n \mathcal{N}(\boldsymbol{O}_t; \boldsymbol{\mu}_n^{(O)}, \boldsymbol{\Sigma}_n^{(OO)})}, \tag{2.18}$$

$$P(\boldsymbol{X}_t|\boldsymbol{O}_t, m, \boldsymbol{\lambda}^{(O,X)}) = \mathcal{N}(\boldsymbol{X}_t; \boldsymbol{E}_{m,t}^{(X)}, \boldsymbol{D}_m^{(X)}), \tag{2.19}$$

where the conditional mean vector $\boldsymbol{E}_{m,t}^{(X)}$ and the conditional covariance matrix $\boldsymbol{D}_m^{(X)}$ are written as

$$\boldsymbol{E}_{m,t}^{(X)} = \boldsymbol{\mu}_m^{(X)} + \boldsymbol{\Sigma}^{(XO)}\boldsymbol{\Sigma}_m^{(OO)^{-1}}(\boldsymbol{O}_t - \boldsymbol{\mu}_m^{(O)}), \tag{2.20}$$

$$\boldsymbol{D}_m^{(X)} = \boldsymbol{\Sigma}_m^{(XX)} - \boldsymbol{\Sigma}_m^{(XO)}\boldsymbol{\Sigma}_m^{(OO)^{-1}}\boldsymbol{\Sigma}_m^{(OX)}. \tag{2.21}$$

In the same way as the parameter generation algorithm in the HMM [40, 41], a time sequence vector of the articulatory parameters is determined under an explicit relationship between the time sequence of static feature vectors $\boldsymbol{x}$ and that of the joint static and dynamic feature vectors $\boldsymbol{X}$, which is given by

$$\boldsymbol{X} = \boldsymbol{W}_x\boldsymbol{x}, \tag{2.22}$$

where $\boldsymbol{W}_x$ is a $2D_xT$-by-$D_xT$ linear transformation matrix to append the dynamic feature vectors from the Eq. (2.6), which is written as

$$\boldsymbol{W}_x = \left[\boldsymbol{W}_{x_1}, \boldsymbol{W}_{x_2}, \ldots, \boldsymbol{W}_{x_t}, \ldots, \boldsymbol{W}_{x_T}\right]^\top \otimes \boldsymbol{I}_{D_x \times D_x}, \tag{2.23}$$

$$\boldsymbol{W}_{x_t} = \left[\boldsymbol{w}_{x_t}^{(0)}, \boldsymbol{w}_{x_t}^{(1)}\right], \tag{2.24}$$

$$\boldsymbol{w}_{x_t}^{(n)} = \Big[\overbrace{0, \ldots, 0}^{1st}, \overbrace{w_x^{(n)}(-L_-^{(n)})}^{(t-L_-^{(n)})-th}, \ldots, \overbrace{w_x^{(n)}(0)}^{(t)-th}, \overbrace{\ldots, w_x^{(n)}(L_+^{(n)})}^{(t+L_+^{(n)})-th}, \overbrace{0, \ldots, 0}^{T-th}\Big]^\top, n = 0, 1, \tag{2.25}$$

$L_-^{(0)} = L_+^{(0)} = 0$ and $w_x^{(0)}(0) = 1$.

In order to maximize the likelihood function $P(\boldsymbol{X}|\boldsymbol{O}, \boldsymbol{\lambda}^{(O,X)})$ in Eq. (2.16), the EM algorithm can be deployed. Explicitly, by using the constraint in Eq. (2.22), an auxiliary function of a current parameter $\boldsymbol{X}$ and an updated parameter $\hat{\boldsymbol{X}}$ is maximized as follows:

$$\begin{aligned} Q(\boldsymbol{X}, \hat{\boldsymbol{X}}) &= \sum_{\text{all } \boldsymbol{m}} P(\boldsymbol{m}|\boldsymbol{O}, \boldsymbol{X}, \boldsymbol{\lambda}^{(O,X)}) \log P(\hat{\boldsymbol{X}}|\boldsymbol{O}, \boldsymbol{m}, \boldsymbol{\lambda}^{(O,X)}) \\ &= -\frac{1}{2}\hat{\boldsymbol{x}}^\top \boldsymbol{W}_x^\top \overline{\boldsymbol{D}^{(X)^{-1}}} \boldsymbol{W}_x \hat{\boldsymbol{x}} + \hat{\boldsymbol{x}}^\top \boldsymbol{W}_x^\top \overline{\boldsymbol{D}^{(X)^{-1}}\boldsymbol{E}^{(X)}} + \overline{K'}, \end{aligned} \tag{2.26}$$

where

$$\overline{\boldsymbol{D}^{(X)^{-1}}} = \mathrm{diag}\left[\overline{\boldsymbol{D}_1^{(X)^{-1}}}, \overline{\boldsymbol{D}_2^{(X)^{-1}}}, \ldots, \overline{\boldsymbol{D}_t^{(X)^{-1}}}, \ldots, \overline{\boldsymbol{D}_T^{(X)^{-1}}}\right], \tag{2.27}$$

$$\begin{aligned}\overline{\boldsymbol{D}^{(X)^{-1}}\boldsymbol{E}^{(X)}} = \Big[&\overline{\boldsymbol{D}_1^{(X)^{-1}}\boldsymbol{E}_1^{(X)^\top}}, \overline{\boldsymbol{D}_2^{(X)^{-1}}\boldsymbol{E}_2^{(X)^\top}} \ldots, \\ &\overline{\boldsymbol{D}_t^{(X)^{-1}}\boldsymbol{E}_t^{(X)^\top}}, \ldots, \overline{\boldsymbol{D}_T^{(X)^{-1}}\boldsymbol{E}_T^{(X)^\top}}\Big]^\top,\end{aligned} \tag{2.28}$$

$$\overline{\boldsymbol{D}_t^{(X)^{-1}}} = \sum_{m=1}^{M} \gamma_{m,t}^{(O,X)} \boldsymbol{D}_m^{(X)^{-1}}, \tag{2.29}$$

$$\overline{\boldsymbol{D}_t^{(X)^{-1}}\boldsymbol{E}_t^{(X)}} = \sum_{m=1}^{M} \gamma_{m,t}^{(O,X)} \boldsymbol{D}_m^{(X)^{-1}}\boldsymbol{E}_{m,t}^{(X)}, \tag{2.30}$$

$$\gamma_{m,t}^{(O,X)} = P(m|\boldsymbol{O}_t, \boldsymbol{X}_t, \boldsymbol{\lambda}^{(O,X)}), \tag{2.31}$$

and the constant $\overline{K'}$ is independent of $\hat{\boldsymbol{X}}$. Then, the time sequence vector of the estimated articulatory parameters $\hat{\boldsymbol{x}}$ would be given by

$$\hat{\boldsymbol{x}} = (\boldsymbol{W}_x^\top \overline{\boldsymbol{D}^{(X)^{-1}}} \boldsymbol{W}_x)^{-1} \boldsymbol{W}_x^\top \overline{\boldsymbol{D}^{(X)^{-1}}\boldsymbol{E}^{(X)}}. \tag{2.32}$$

In this thesis, the likelihood function in Eq. (2.16) is approximated with a single mixture component sequence as follows:

$$P(\boldsymbol{X}|\boldsymbol{O}, \boldsymbol{\lambda}^{(O,X)}) \simeq P(\boldsymbol{m}|\boldsymbol{O}, \boldsymbol{\lambda}^{(O,X)}) P(\boldsymbol{X}|\boldsymbol{O}, \boldsymbol{m}, \boldsymbol{\lambda}^{(O,X)}). \tag{2.33}$$

First, the sub-optimum mixture component sequence $\hat{\boldsymbol{m}}^{(O)} = \{\hat{m}_1^{(O)}, \hat{m}_2^{(O)}, \ldots, \hat{m}_t^{(O)}, \ldots, \hat{m}_T^{(O)}\}$ is determined by

$$\hat{\boldsymbol{m}}^{(O)} = \arg\max_{\boldsymbol{m}^{(O)}} P(\boldsymbol{m}^{(O)}|\boldsymbol{O}, \boldsymbol{\lambda}^{(O,X)}). \tag{2.34}$$

Then, the maximization of the auxiliary function is approximated as follows:

$$\begin{aligned}Q(\boldsymbol{X}, \hat{\boldsymbol{X}}) &\simeq \log P(\hat{\boldsymbol{m}}^{(O)}|\boldsymbol{O}, \boldsymbol{\lambda}^{(O,X)}) P(\hat{\boldsymbol{X}}|\boldsymbol{O}, \hat{\boldsymbol{m}}^{(O)}, \boldsymbol{\lambda}^{(O,X)}) \\ &= -\frac{1}{2}\hat{\boldsymbol{x}}^\top \boldsymbol{W}_x^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}^{(O)}}^{(X)^{-1}} \boldsymbol{W}_x \hat{\boldsymbol{x}} + \hat{\boldsymbol{x}}^\top \boldsymbol{W}_x^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}^{(O)}}^{(X)^{-1}} \boldsymbol{E}_{\hat{\boldsymbol{m}}^{(O)}}^{(X)} + K',\end{aligned} \tag{2.35}$$

where

$$\boldsymbol{E}_{\hat{\boldsymbol{m}}^{(O)}}^{(X)} = \left[\boldsymbol{E}_{\hat{m}_1^{(O)},1}^{(X)^\top}, \boldsymbol{E}_{\hat{m}_2^{(O)},2}^{(X)^\top}, \ldots, \boldsymbol{E}_{\hat{m}_t^{(O)},t}^{(X)^\top}, \ldots, \boldsymbol{E}_{\hat{m}_T^{(O)},T}^{(X)^\top}\right]^\top, \tag{2.36}$$

$$\boldsymbol{D}_{\hat{\boldsymbol{m}}^{(O)}}^{(X)^{-1}} = \mathrm{diag}\left[\boldsymbol{D}_{\hat{m}_1^{(O)}}^{(X)^{-1}}, \boldsymbol{D}_{\hat{m}_2^{(O)}}^{(X)^{-1}}, \ldots, \boldsymbol{D}_{\hat{m}_t^{(O)}}^{(X)^{-1}}, \ldots, \boldsymbol{D}_{\hat{m}_T^{(O)}}^{(X)^{-1}}\right]. \tag{2.37}$$

So that, the time sequence vector of the estimated articulatory parameters $\hat{\boldsymbol{x}}$ is finally given by

$$\hat{\boldsymbol{x}} = (\boldsymbol{W}_x^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}^{(O)}}^{(X)^{-1}} \boldsymbol{W}_x)^{-1} \boldsymbol{W}_x^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}^{(O)}}^{(X)^{-1}} \boldsymbol{E}_{\hat{\boldsymbol{m}}^{(O)}}^{(X)}. \tag{2.38}$$

## 2.3. GMM-based Articulatory-to-Acoustic Production Mapping

In the GMM-based articulatory-to-acoustic production mapping, the source features are composed not only by the articulatory parameters, but also by the source-excitation parameters, i.e. log-scaled $F_0$ and log-scaled waveform power in this thesis, whereas the target features are composed by the acoustic parameters. Let us then define a $D_s$-dimensional feature vector of the source excitation parameters as $\boldsymbol{s}_t$ at frame $t$.

As the source features, a $2(D_x + D_s)$-dimensional feature vector denoted as $\boldsymbol{Y}_t = [\boldsymbol{x}_t^\top, \boldsymbol{s}_t^\top \Delta \boldsymbol{x}_t^\top, \Delta \boldsymbol{s}_t^\top]^\top$ is used at frame $t$. On the other hand, as the target features, a $2D_c$-dimensional feature vector denoted as $\boldsymbol{C}_t = [\boldsymbol{c}_t^\top, \Delta \boldsymbol{c}_t^\top]^\top$ is used at frame $t$. So that, the time sequence of the source feature vectors and that of the target feature vectors are respectively written as

$$\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top, \ldots, \boldsymbol{Y}_t^\top, \ldots, \boldsymbol{Y}_T^\top]^\top, \tag{2.39}$$

$$\boldsymbol{C} = [\boldsymbol{C}_1^\top, \boldsymbol{C}_2^\top, \ldots, \boldsymbol{C}_t^\top, \ldots, \boldsymbol{C}_T^\top]^\top. \tag{2.40}$$

As in the section 2.2, the dynamic feature vector $\Delta \boldsymbol{s}_t$ is computed from the static feature vector $\boldsymbol{s}_t$ at frame $t$, which is given by

$$\Delta \boldsymbol{s}_t = \sum_{\tau=-L_-^{(1)}}^{L_+^{(1)}} w^{(1)}(\tau) \boldsymbol{s}_{t+\tau}. \tag{2.41}$$

Similarly, the dynamic feature vector $\Delta \boldsymbol{c}_t$ is computed from the static feature vector $\boldsymbol{c}_t$ at frame $t$ as given by

$$\Delta \boldsymbol{c}_t = \sum_{\tau=-L_-^{(1)}}^{L_+^{(1)}} w^{(1)}(\tau) \boldsymbol{c}_{t+\tau}, \tag{2.42}$$

where $w^{(1)}(\tau)$, $L_{-}^{(1)}$, and $L_{+}^{(1)}$ are the 1-st order weight coefficients and the frame lengths for computing the dynamic feature vectors.

### 2.3.1 Training process

Schematic flow of the training process for production mapping is shown in Fig. 2.3. From the training data, a joint source and target feature vector $[\boldsymbol{Y}_t^\top, \boldsymbol{C}_t^\top]^\top$ is developed at each frame $t$. Then, its joint probability density function is modeled with the GMM for the production mapping as follows:

$$\begin{aligned} P(\boldsymbol{Y}_t, \boldsymbol{C}_t|\boldsymbol{\lambda}^{(Y,C)}) &= \sum_{m=1}^{M} P(\boldsymbol{Y}_t, \boldsymbol{C}_t|m, \boldsymbol{\lambda}^{(Y,C)}) \\ &= \sum_{m=1}^{M} \alpha_m^{(Y,C)} \mathcal{N}([\boldsymbol{Y}_t^\top, \boldsymbol{C}_t^\top]^\top; \boldsymbol{\mu}_m^{(Y,C)}, \boldsymbol{\Sigma}_m^{(Y,C)}), \end{aligned} \tag{2.43}$$

where the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ is denoted as $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Again, the mixture component index is denoted as $m$ and the total number of mixture components is $M$. The set of the GMM parameters for the production mapping is denoted as $\boldsymbol{\lambda}^{(Y,C)}$, which consists of weights $\alpha_m^{(Y,C)}$, mean vectors $\boldsymbol{\mu}_m^{(Y,C)}$ and covariance matrices $\boldsymbol{\Sigma}_m^{(Y,C)}$ of individual mixture components. The mean vector $\boldsymbol{\mu}_m^{(Y,C)}$ and covariance matrix $\boldsymbol{\Sigma}_m^{(Y,C)}$ of the $m$th mixture component are written as

$$\boldsymbol{\mu}_m^{(Y,C)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(Y)} \\ \boldsymbol{\mu}_m^{(C)} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(Y,C)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(YY)} & \boldsymbol{\Sigma}_m^{(YC)} \\ \boldsymbol{\Sigma}_m^{(CY)} & \boldsymbol{\Sigma}_m^{(CC)} \end{bmatrix}, \tag{2.44}$$

where the mean vectors of the source feature vectors and that of the target feature vectors for the $m$th mixture component are denoted as $\boldsymbol{\mu}_m^{(Y)}$ and $\boldsymbol{\mu}_m^{(C)}$, respectively. The covariance matrix of the source feature vectors is denoted as $\boldsymbol{\Sigma}_m^{(YY)}$ and that of the target feature vectors is denoted as $\boldsymbol{\Sigma}_m^{(CC)}$ for the $m$th mixture component. The cross covariance matrices of the source and target feature vectors for the $m$th mixture components are denoted as $\boldsymbol{\Sigma}_m^{(YC)}$ and $\boldsymbol{\Sigma}_m^{(CY)}$. All of these covariance matrices are full.

In order to train the GMM parameters of the production mapping $\boldsymbol{\lambda}^{(Y,C)}$, the

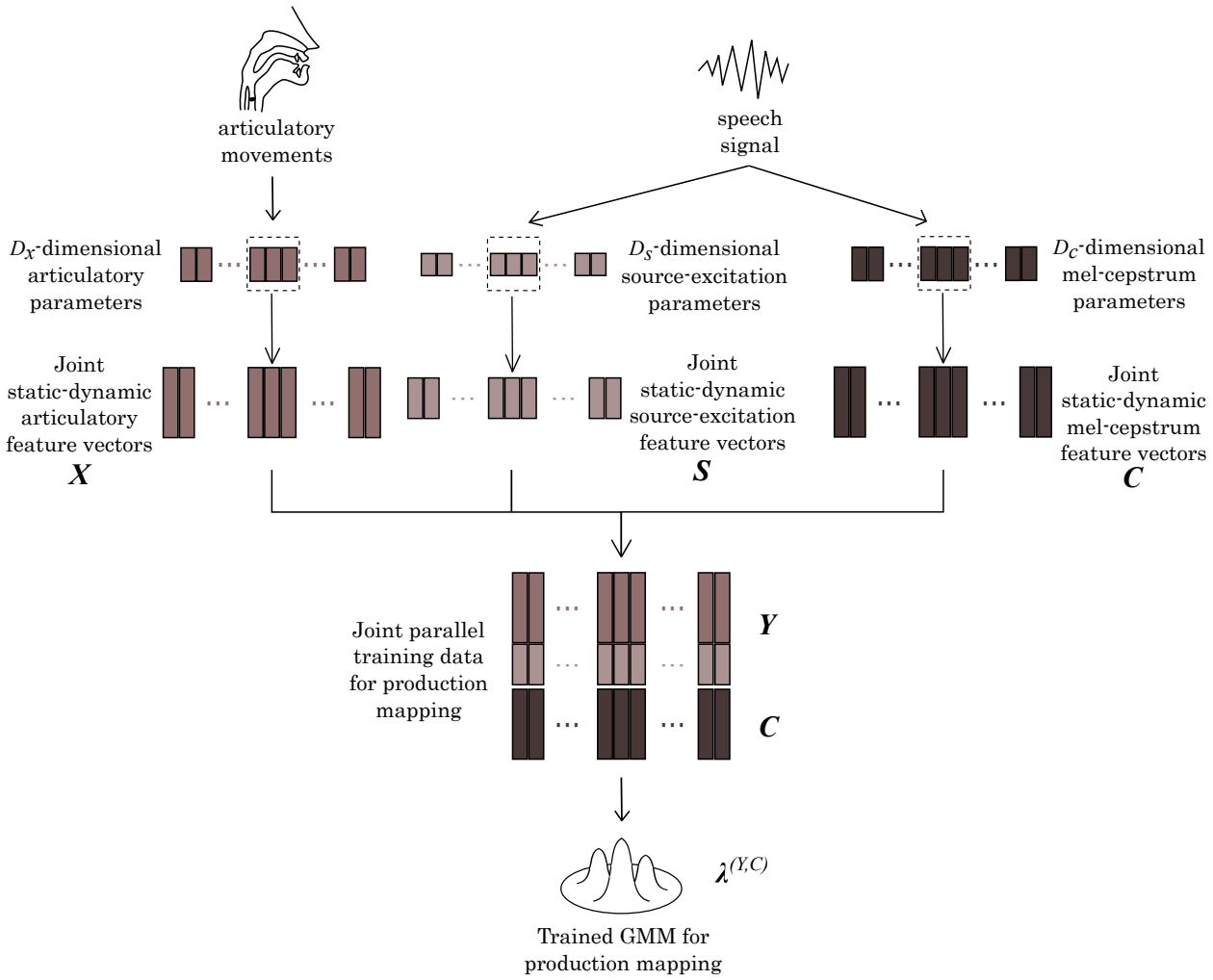



Figure 2.3. Schematic flow of the training process for GMM-based production mapping.

following likelihood function is to be maximized

$$P(\boldsymbol{Y}, \boldsymbol{C}|\boldsymbol{\lambda}^{(Y,C)}) = \sum_{\text{all } \boldsymbol{m}} P(\boldsymbol{m}|\boldsymbol{Y}, \boldsymbol{C}, \boldsymbol{\lambda}^{(Y,C)}) P(\boldsymbol{Y}, \boldsymbol{C}, |\boldsymbol{m}, \boldsymbol{\lambda}^{(Y,C)})$$
$$= \prod_{t=1}^{T} \sum_{m=1}^{M} P(m|\boldsymbol{Y}_t, \boldsymbol{C}_t, \boldsymbol{\lambda}^{(Y,C)}) P(\boldsymbol{Y}_t, \boldsymbol{C}_t|m, \boldsymbol{\lambda}^{(Y,C)}), \qquad (2.45)$$

where $\boldsymbol{m} = \{m_1, m_2, \ldots, m_t, \ldots, m_T\}$ is again a mixture component sequence.

Employing also the EM algorithm [39], given a set of initial GMM parameters for the production mapping $\boldsymbol{\lambda}^{(Y,C)}$, a set of updated parameters $\hat{\boldsymbol{\lambda}}^{(Y,C)}$ is estimated

with respect to the following auxiliary function

$$Q(\hat{\boldsymbol{\lambda}}^{(Y,C)}, \boldsymbol{\lambda}^{(Y,C)}) = \sum_{\text{all } \boldsymbol{m}} P(\boldsymbol{m}|\boldsymbol{Y}, \boldsymbol{C}, \boldsymbol{\lambda}^{(Y,C)}) \log P(\boldsymbol{Y}, \boldsymbol{C}, |\boldsymbol{m}, \hat{\boldsymbol{\lambda}}^{(Y,C)}). \tag{2.46}$$

In the E-step, the occupancies $\gamma_{m,t}^{(Y,C)}$ and the number of samples $N_m^{(Y,C)}$ are given by

$$\begin{aligned}\gamma_{m,t}^{(Y,C)} &= P(m|\boldsymbol{Y}_t, \boldsymbol{C}_t, \boldsymbol{\lambda}^{(Y,C)}) \\ &= \frac{\alpha_m \mathcal{N}([\boldsymbol{Y}_t^\top, \boldsymbol{C}_t^\top]^\top; \boldsymbol{\mu}_m^{(Y,C)}, \boldsymbol{\Sigma}_m^{(Y,C)})}{\sum_{n=1}^M \alpha_n \mathcal{N}([\boldsymbol{Y}_t^\top, \boldsymbol{C}_t^\top]^\top; \boldsymbol{\mu}_n^{(Y,C)}, \boldsymbol{\Sigma}_n^{(Y,C)})}, \end{aligned} \tag{2.47}$$

$$N_m^{(Y,C)} = \sum_{t=1}^T \gamma_{m,t}^{(Y,C)}. \tag{2.48}$$

Then, in the M-step, the updated parameters in the set $\hat{\boldsymbol{\lambda}}^{(Y,C)}$ would be given by

$$\hat{\alpha}_m^{(Y,C)} = \frac{N_m^{(Y,C)}}{T} \tag{2.49}$$

$$\hat{\boldsymbol{\mu}}_m^{(Y,C)} = \frac{1}{N_m^{(Y,C)}} \sum_{t=1}^T \gamma_{m,t}^{(Y,C)} [\boldsymbol{Y}_t^\top, \boldsymbol{C}_t^\top]^\top \tag{2.50}$$

$$\hat{\boldsymbol{\Sigma}}_m^{(Y,C)} = \frac{1}{N_m^{(Y,C)}} \sum_{t=1}^T \gamma_{m,t}^{(Y,C)} ([\boldsymbol{Y}_t^\top, \boldsymbol{C}_t^\top]^\top - \hat{\boldsymbol{\mu}}_m^{(Y,C)})([\boldsymbol{Y}_t^\top, \boldsymbol{C}_t^\top]^\top - \hat{\boldsymbol{\mu}}_m^{(Y,C)})^\top. \tag{2.51}$$

### 2.3.2 Conversion process

Schematic flow of the conversion process for production mapping is shown in Fig. 2.4. Given a time sequence of the articulatory feature vectors $\boldsymbol{Y}$ and the trained GMM parameters $\boldsymbol{\lambda}^{(Y,C)}$, a time sequence of the corresponding acoustic feature vectors $\hat{\boldsymbol{c}}$ is determined by

$$\hat{\boldsymbol{c}} = \arg\max_{\boldsymbol{c}} P(\boldsymbol{C}|\boldsymbol{Y}, \boldsymbol{\lambda}^{(Y,C)}), \tag{2.52}$$

where

$$\begin{aligned} P(\boldsymbol{C}|\boldsymbol{Y}, \boldsymbol{\lambda}^{(Y,C)}) &= \sum_{\text{all } \boldsymbol{m}} P(\boldsymbol{m}|\boldsymbol{Y}, \boldsymbol{\lambda}^{(Y,C)}) P(\boldsymbol{C}|\boldsymbol{Y}, \boldsymbol{m}, \boldsymbol{\lambda}^{(Y,C)}) \\ &= \prod_{t=1}^T \sum_{m=1}^M P(m|\boldsymbol{Y}_t, \boldsymbol{\lambda}^{(Y,C)}) P(\boldsymbol{C}_t|\boldsymbol{Y}_t, m, \boldsymbol{\lambda}^{(Y,C)}), \end{aligned} \tag{2.53}$$

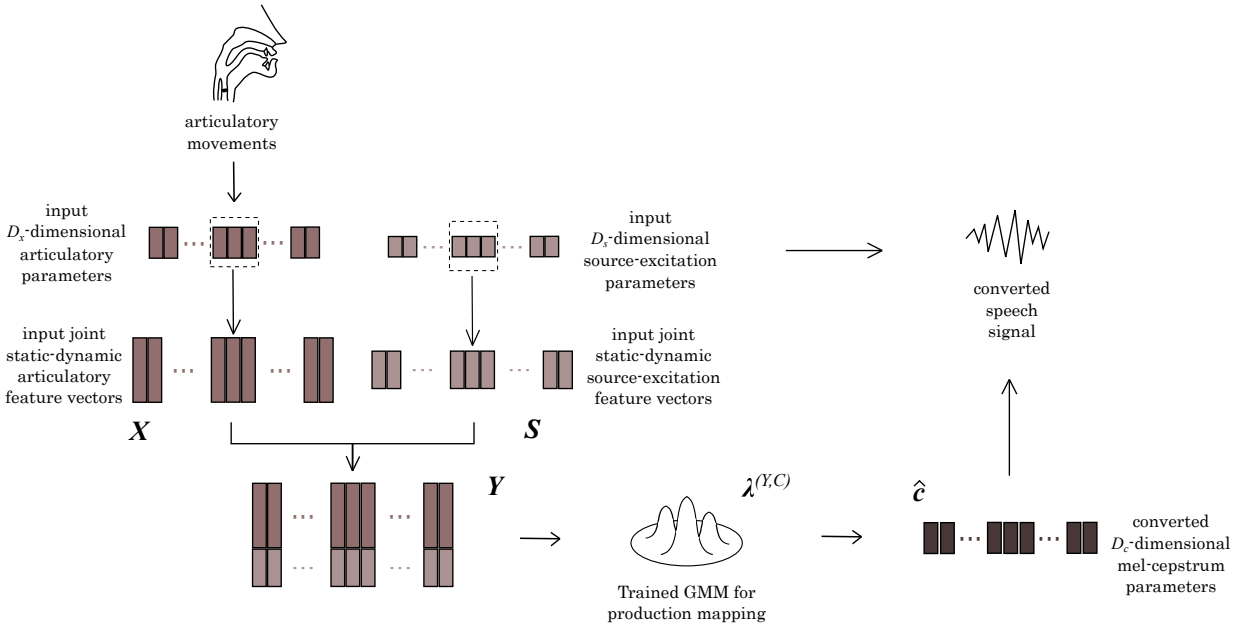



Figure 2.4. Schematic flow of the conversion process for GMM-based production mapping.

and $\boldsymbol{m} = \{m_1, m_2, \ldots, m_t, \ldots, m_T\}$ is again a mixture component sequence.

The $m$-th posterior probability density $P(m|\boldsymbol{Y}_t, \boldsymbol{\lambda}^{(Y,C)})$ and the $m$-th conditional probability density $P(\boldsymbol{C}_t|\boldsymbol{Y}_t, m, \boldsymbol{\lambda}^{(Y,C)})$ at each frame $t$ are given by:

$$P(m|\boldsymbol{Y}_t, \boldsymbol{\lambda}^{(Y,C)}) = \frac{\alpha_m \mathcal{N}(\boldsymbol{Y}_t; \boldsymbol{\mu}_m^{(Y)}, \boldsymbol{\Sigma}_m^{(YY)})}{\sum_{n=1}^{M} \alpha_n \mathcal{N}(\boldsymbol{Y}_t; \boldsymbol{\mu}_n^{(Y)}, \boldsymbol{\Sigma}_n^{(YY)})}, \tag{2.54}$$

$$P(\boldsymbol{C}_t|\boldsymbol{Y}_t, m, \boldsymbol{\lambda}^{(Y,C)}) = \mathcal{N}(\boldsymbol{C}_t; \boldsymbol{E}_{m,t}^{(C)}, \boldsymbol{D}_m^{(C)}), \tag{2.55}$$

where the conditional mean vector $\boldsymbol{E}_{m,t}^{(C)}$ and the conditional covariance matrix $\boldsymbol{D}_m^{(C)}$ are written as

$$\boldsymbol{E}_{m,t}^{(C)} = \boldsymbol{\mu}_m^{(C)} + \boldsymbol{\Sigma}^{(CY)} \boldsymbol{\Sigma}_m^{(YY)^{-1}} (\boldsymbol{Y}_t - \boldsymbol{\mu}_m^{(Y)}), \tag{2.56}$$

$$\boldsymbol{D}_m^{(C)} = \boldsymbol{\Sigma}_m^{(CC)} - \boldsymbol{\Sigma}_m^{(CY)} \boldsymbol{\Sigma}_m^{(YY)^{-1}} \boldsymbol{\Sigma}_m^{(YC)}. \tag{2.57}$$

As in the section 2.2.2, following the parameter generation algorithm of the HMM [40, 41], a time sequence vector of the acoustic parameters $\hat{\boldsymbol{c}}$ is determined under an explicit relationship between the time sequence of static feature vectors

$\boldsymbol{c}$ and that of the joint static and dynamic feature vectors $\boldsymbol{C}$, which is given by

$$\boldsymbol{C} = \boldsymbol{W}_c \boldsymbol{c}, \tag{2.58}$$

where $\boldsymbol{W}_c$ is a $2D_cT$-by-$D_cT$ linear transformation matrix to append the dynamic feature vectors from the Eq. (2.42), which is written as

$$\boldsymbol{W}_c = \left[\boldsymbol{W}_{c_1}, \boldsymbol{W}_{c_2}, \ldots, \boldsymbol{W}_{c_t}, \ldots, \boldsymbol{W}_{c_T}\right]^\top \otimes \boldsymbol{I}_{D_c \times D_c}, \tag{2.59}$$

$$\boldsymbol{W}_{c_t} = \left[\boldsymbol{w}_{c_t}^{(0)}, \boldsymbol{w}_{c_t}^{(1)}\right], \tag{2.60}$$

$$\boldsymbol{w}_{c_t}^{(n)} = \Big[\overbrace{0, \ldots, 0}^{1st}, \overbrace{w_c^{(n)}(-L_-^{(n)}), \ldots, w_c^{(n)}(0)}^{(t-L_-^{(n)})-th}, \overbrace{\ldots, w_c^{(n)}(L_+^{(n)}), 0, \ldots, 0}^{(t+L_+^{(n)})-th}\overbrace{\Big]^\top}^{T-th}, n = 0, 1, \tag{2.61}$$

$L_-^{(0)} = L_+^{(0)} = 0$ and $w_c^{(0)}(0) = 1$.

Then, in order to maximize the likelihood function $P(\boldsymbol{C}|\boldsymbol{Y}, \boldsymbol{\lambda}^{(Y,C)})$ in Eq. (2.52), the EM algorithm can be deployed. Specifically, by using the constraint in Eq. (2.58), an auxiliary function of a current parameter $\boldsymbol{Y}$ and an updated parameter $\hat{\boldsymbol{Y}}$ is maximized as follows:

$$\begin{aligned} Q(\boldsymbol{C}, \hat{\boldsymbol{C}}) &= \sum_{\text{all } \boldsymbol{m}} P(\boldsymbol{m}|\boldsymbol{Y}, \boldsymbol{C}, \boldsymbol{\lambda}^{(Y,C)}) \log P(\hat{\boldsymbol{C}}|\boldsymbol{Y}, \boldsymbol{m}, \boldsymbol{\lambda}^{(Y,C)}) \\ &= -\frac{1}{2}\hat{\boldsymbol{c}}^\top \boldsymbol{W}_c^\top \overline{\boldsymbol{D}^{(C)^{-1}}} \boldsymbol{W}_c \hat{\boldsymbol{c}} + \hat{\boldsymbol{c}}^\top \boldsymbol{W}_c^\top \overline{\boldsymbol{D}^{(C)^{-1}} \boldsymbol{E}^{(C)}} + \overline{K'}, \end{aligned} \tag{2.62}$$

where

$$\overline{\boldsymbol{D}^{(C)^{-1}}} = \operatorname{diag}\left[\overline{\boldsymbol{D}_1^{(C)^{-1}}}, \overline{\boldsymbol{D}_2^{(C)^{-1}}}, \ldots, \overline{\boldsymbol{D}_t^{(C)^{-1}}}, \ldots, \overline{\boldsymbol{D}_T^{(C)^{-1}}}\right], \tag{2.63}$$

$$\begin{aligned} \overline{\boldsymbol{D}^{(C)^{-1}} \boldsymbol{E}^{(C)}} = \Big[&\overline{\boldsymbol{D}_1^{(C)^{-1}} \boldsymbol{E}_1^{(C)^\top}}, \overline{\boldsymbol{D}_1^{(C)^{-1}} \boldsymbol{E}_2^{(C)^\top}}, \ldots, \\ &\overline{\boldsymbol{D}_t^{(C)^{-1}} \boldsymbol{E}_t^{(C)^\top}}, \ldots, \overline{\boldsymbol{D}_T^{(C)^{-1}} \boldsymbol{E}_T^{(C)^\top}}\Big]^\top, \end{aligned} \tag{2.64}$$

$$\overline{\boldsymbol{D}_t^{(C)^{-1}}} = \sum_{m=1}^{M} \gamma_{m,t}^{(Y,C)} \boldsymbol{D}_m^{(C)^{-1}}, \tag{2.65}$$

$$\overline{\boldsymbol{D}_t^{(C)^{-1}}, \boldsymbol{E}_t^{(C)}} = \sum_{m=1}^{M} \gamma_{m,t}^{(Y,C)} \boldsymbol{D}_m^{(C)^{-1}} \boldsymbol{E}_{m,t}^{(C)}, \tag{2.66}$$

$$\gamma_{m,t}^{(Y,C)} = P(m|\boldsymbol{Y}_t, \boldsymbol{C}_t, \boldsymbol{\lambda}^{(Y,C)}), \tag{2.67}$$

and the constant $\overline{K'}$ is again independent of $\hat{\boldsymbol{C}}$. Thus, the time sequence vector of the estimated acoustic parameters $\hat{\boldsymbol{c}}$ would be given by

$$\hat{\boldsymbol{c}} = (\boldsymbol{W}_c^\top \overline{\boldsymbol{D}^{(C)^{-1}}} \boldsymbol{W}_c)^{-1} \boldsymbol{W}_c^\top \overline{\boldsymbol{D}^{(C)^{-1}} \boldsymbol{E}^{(C)}}. \tag{2.68}$$

In this thesis, the likelihood function in Eq. (2.52) is again approximated with a single mixture component sequence as follows:

$$P(\boldsymbol{C}|\boldsymbol{Y}, \boldsymbol{\lambda}^{(Y,C)}) \simeq P(\boldsymbol{m}|\boldsymbol{Y}, \boldsymbol{\lambda}^{(Y,C)}) P(\boldsymbol{C}|\boldsymbol{Y}, \boldsymbol{m}, \boldsymbol{\lambda}^{(Y,C)}). \tag{2.69}$$

First, the sub-optimum mixture component sequence $\hat{\boldsymbol{m}}^{(Y)} = \{\hat{m}_1^{(Y)}, \hat{m}_2^{(Y)}, \ldots, \hat{m}_t^{(Y)}, \ldots, \hat{m}_T^{(Y)}\}$ is determined by

$$\hat{\boldsymbol{m}}^{(Y)} = \arg\max_{\boldsymbol{m}^{(Y)}} P(\boldsymbol{m}^{(Y)}|\boldsymbol{Y}, \boldsymbol{\lambda}^{(Y,C)}). \tag{2.70}$$

The maximization of the auxiliary function is then approximated as follows:

$$\begin{aligned} Q(\boldsymbol{C}, \hat{\boldsymbol{C}}) &\simeq \log P(\hat{\boldsymbol{m}}^{(Y)}|\boldsymbol{Y}, \boldsymbol{\lambda}^{(O,X)}) P(\hat{\boldsymbol{C}}|\boldsymbol{Y}, \hat{\boldsymbol{m}}^{(Y)}, \boldsymbol{\lambda}^{(Y,C)}) \\ &= -\frac{1}{2} \hat{\boldsymbol{c}}^\top \boldsymbol{W}_c^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}^{(Y)}}^{(C)^{-1}} \boldsymbol{W}_c \hat{\boldsymbol{c}} + \hat{\boldsymbol{c}}^\top \boldsymbol{W}_c^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}^{(Y)}}^{(C)^{-1}} \boldsymbol{E}_{\hat{\boldsymbol{m}}^{(Y)}}^{(C)} + K', \end{aligned} \tag{2.71}$$

where

$$\boldsymbol{E}_{\hat{\boldsymbol{m}}^{(Y)}}^{(C)} = \left[\boldsymbol{E}_{\hat{m}_1^{(Y)},1}^{(C)^\top}, \boldsymbol{E}_{\hat{m}_2^{(Y)},1}^{(C)^\top}, \ldots, \boldsymbol{E}_{\hat{m}_t^{(Y)},t}^{(C)^\top}, \ldots, \boldsymbol{E}_{\hat{m}_T^{(Y)},T}^{(C)^\top}\right]^\top, \tag{2.72}$$

$$\boldsymbol{D}_{\hat{\boldsymbol{m}}^{(Y)}}^{(C)^{-1}} = \mathrm{diag}\left[\boldsymbol{D}_{\hat{m}_1^{(Y)}}^{(C)^{-1}}, \boldsymbol{D}_{\hat{m}_2^{(Y)}}^{(C)^{-1}}, \ldots, \boldsymbol{D}_{\hat{m}_t^{(Y)}}^{(C)^{-1}}, \ldots, \boldsymbol{D}_{\hat{m}_T^{(Y)}}^{(C)^{-1}}\right]. \tag{2.73}$$

So that, the time sequence vector of the estimated acoustic parameters $\hat{\boldsymbol{c}}$ would be finally given by

$$\hat{\boldsymbol{c}} = (\boldsymbol{W}_c^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}^{(Y)}}^{(C)^{-1}} \boldsymbol{W}_c)^{-1} \boldsymbol{W}_c^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}^{(Y)}}^{(C)^{-1}} \boldsymbol{E}_{\hat{\boldsymbol{m}}^{(Y)}}^{(C)}. \tag{2.74}$$

## 2.4. Oversmoothing Problem of the Converted Trajectory

One of the main problem of the conversion process in the GMM-based mapping technique is the oversmoothed characteristic of the converted trajectory. This

is because the trajectory-based conversion elaborated in both sections 2.2.2 and 2.3.2 forces the generated parameters to be as close as possible to the mean vector sequence of the conditional probability density function. Moreover, in training each of the mixture components, multiple contexts encompassing various characteristics of features are considered. In this case, the variance features are considered to be noise in modeling the joint probability density function. Therefore, reduction of the global variance (GV) [30] is often observed and oversmoothes the converted trajectory.

In order to address this oversmoothing problem, a parameter generation process considering GV has been studied in [30]. Its effectiveness has been confirmed by the capability of the method in significantly improving both speech quality and conversion accuracy in a voice conversion task. Moreover, in [42], the GV criterion has been included in the training process to address the oversmoothing issue while preserving the consistency between training and conversion process. In [43], another constraint in a parameter generation procedure has been proposed to alleviate the oversmoothing effect by considering the modulation spectrum (MS), i.e. in this case MS is the log-scaled power spectrum of the parameter sequence. In [44], the MS constraint has been also taken into account within the training procedure. Both of these works have shown very promising results in addresing the oversmoothing issue.

In this thesis, we are observing the oversmoothing problem from another point of view by considering the generation process of the speech waveform. In the conventional method, i.e. using vocoder-based speech generation process, a speech waveform is synthesized from the spectral envelope parameters and the excitation parameters. However, this vocoder-based procedure is very sensitive to the errors from the extracted speech parameters, such as $F_0$ extraction errors and spectral parameterization errors. Thus, the oversmoothing effects often observed in the synthesized speech waveform, which significantly degrade the synthesized speech quality, where one may call it as the "vocoded speech" quality. In order to address this issue, one possible solution is to avoid the use of the vocoder-based process in synthesizing the speech waveform. In [32], it has been studied that by directly filtering an input speech waveform according to the spectrum differences between modified and input speech, which avoid the use of vocoder-based excita-

tion generation process, a significant improvement of the synthetic speech quality can be achieved. This method will be elaborated further in the section 3.4.

## 2.5. Summary

In this chapter, two GMM-based statistical feature mapping methods between acoustic and articulatory parameters have been elaborated, i.e. the GMM-based acoustic-to-articulatory inversion mapping and the GMM-based articulatory-to-acoustic production mapping. The advantages of using the GMM-based statistical feature mapping are its independency of input textual features and its flexibility in terms of parameters modification. This flexibility is shown by the clear and convenient process of model training and parameters conversion, which mainly utilize the sophisticated EM algorithm. In this thesis, these traits will prove to be significant benefits allowing the development of a high-quality speech modification system through the manipulation of articulatory movements.

# Chapter 3

# Proposed Articulatory Controllable Speech Modification Method using GMM-based Statistical Feature Mappings

## 3.1. Introduction

In this chapter, based on the statistical feature mappings between acoustic and articulatory parameters explained in the chapter 2, we propose the articulatory controllable speech modification method based on Gaussian mixture models (GMMs). In the proposed method, both the GMM-based acoustic-to-articulatory inversion mapping and the GMM-based articulatory-to-acoustic production mapping are integrated in a sequential mapping process. The proposed method allows us to perform a speech modification process through a manipulation procedure for controlling the unobserved articulatory movements corresponding to the input speech signal. Moreover, high-quality modified speech sounds can be generated from the manipulated articulatory movements by utilizing the spectrum differences between the original spectrum and the modified spectrum to avoid the use of vocoder-based speech generation process [32].

This chapter is organized as follows. In section 3.2, the proposed sequential

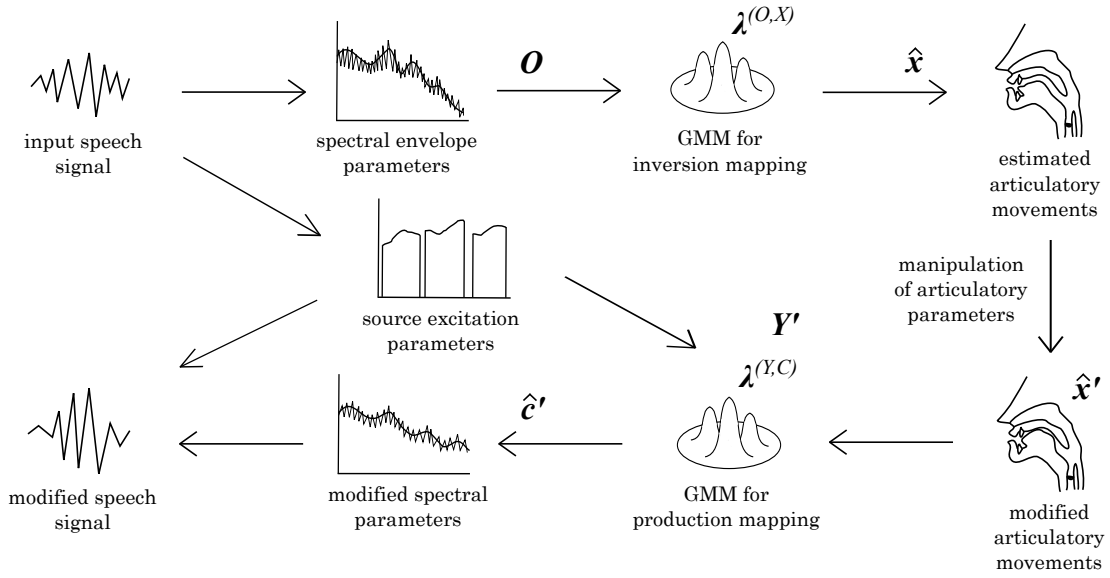



Figure 3.1. Schematic flow of the proposed sequential mapping for articulatory controllable speech modification system.

mapping system between acoustic and articulatory parameters is described. In section 3.3, the proposed methods for manipulating the articulatory movements are described. In section 3.4, the proposed direct waveform modification methods for generating modified speech sounds using spectrum differential are described. Finally, we summarize this chapter in section 3.5.

## 3.2. Sequential Mapping System between Acoustic and Articulatory Parameters

In the proposed system, the GMM-based acoustic-to-articulatory inversion mapping, desribed in section 2.2, and the GMM-based articulatory-to-acoustic production mapping, described in section 2.3, are sequentially integrated. First, given an input speech signal, the mel-cepstrum parameters and source excitation parameters are extracted. Then, using the trained GMM parameters of inversion mapping $\boldsymbol{\lambda}^{(O,X)}$, given the time sequence vector mel-cepstral segment features

$\boldsymbol{O}$, a time sequence vector of articulatory parameters $\hat{\boldsymbol{x}}$ is estimated. In order to perform a speech modification, the estimated articulatory parameters can be manually manipulated, which would be denoted as $\hat{\boldsymbol{x}}'$. The methods for manipulating these parameters are explained within the section 3.3. So that, given the time sequence vector of manipulated articulatory parameters and source excitation parameters $\boldsymbol{Y}'$, by using the trained GMM parameters of production mapping $\boldsymbol{\lambda}^{(Y,C)}$, a time sequence vector of the modified mel-cepstrum parameters $\hat{\boldsymbol{c}}'$ is estimated. Finally, the modified speech signal can be generated by using a direct waveform modification method utilizing the spectrum differences between the original spectrum and the modified one. The proposed direct waveform modification methods are further explained in the section 3.4. Figure 3.1 illustrate the proposed sequential mapping system.

In the proposed system, there are two main characteristics that serve as the fundamental reasons for the development, i.e. its intuitive approach in performing speech modification procedure and its flexibility for convenient parameter modification for various purposes. The proposed system allows us to manipulate the unobserved articulatory movements from a given input speech signal to perform the speech modification. This process indeed benefits from the trait of the articulatory movements, which are more understandable compared to the spectrum of the vocal tract. Then, the methods for manipulating the articulatory movements and the direct waveform modification methods using spectrum differential [32] can also be easily applied, thanks to the use of the GMM-based statistical feature mappings, which have sophisticated and convenient procedures. Moreover, thanks to its independency of any input textual features, this system can then be easily adapted into any languages, which raises the chances to be implemented in various speech applications, such as language-learning or speech-therapy system.

## 3.3. Manipulation Methods for Controlling the Articulatory Movements

In order to perform manipulation of the articulatory movements, it is more convenient to manually control the movements of a limited number of articulators, e.g. only the movement of the tongue tip, rather than to manually control all

articulators simultaneously. In this section, in order to perform such modification, two methods of manipulation for controlling the articulatory movements are described, i.e. simple manipulation method and manipulation method by considering inter-correlations of articulatory parameters.

### 3.3.1 Simple manipulation method

From the inversion mapping described in the section 2.2, at frame $t$, a $D_x$-dimensional feature vector of the estimated articulatory parameters $\hat{\boldsymbol{x}}_t$ is written as

$$\hat{\boldsymbol{x}}_t = [\hat{x}_t(1), \hat{x}_t(2), \ldots, \hat{x}_t(d), \ldots, \hat{x}_t(D_x)]^\top. \tag{3.1}$$

By performing scaling and/or translation, the feature vector of the manipulated articulatory parameters $\hat{\boldsymbol{x}}'_t$ will be given with the following simple linear transformation

$$\hat{\boldsymbol{x}}'_t = \boldsymbol{\Lambda}_t \hat{\boldsymbol{x}}_t + \boldsymbol{\psi}_t, \tag{3.2}$$

where

$$\boldsymbol{\Lambda}_t = \mathrm{diag}[\Lambda_t(1), \Lambda_t(2), \ldots, \Lambda_t(d), \ldots, \Lambda_t(D_x)], \tag{3.3}$$

$$\boldsymbol{\psi}_t = [\psi_t(1), \psi_t(2), \ldots, \psi_t(d), \ldots, \psi_t(D_x)]^\top. \tag{3.4}$$

Here, the scaling factor and the translation factor to manipulate the $d$-th dimension articulatory parameter at frame $t$ are denoted as $\Lambda_t(d)$ and $\psi_t(d)$, respectively. By default, their values respectively are 1 and 0. So then, given a time sequence vector of the estimated articulatory parameters $\hat{\boldsymbol{x}}$, which is written as

$$\hat{\boldsymbol{x}} = [\hat{\boldsymbol{x}}_1^\top, \hat{\boldsymbol{x}}_2^\top, \ldots, \hat{\boldsymbol{x}}_t^\top, \ldots, \hat{\boldsymbol{x}}_T^\top]^\top, \tag{3.5}$$

a time sequence vector of the manipulated articulatory parameters $\hat{\boldsymbol{x}}'$ would then be simply given by

$$\hat{\boldsymbol{x}}' = \boldsymbol{\Lambda}\hat{\boldsymbol{x}} + \boldsymbol{\psi}, \tag{3.6}$$

where

$$\boldsymbol{\Lambda} = [\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2, \ldots, \boldsymbol{\Lambda}_t, \ldots, \boldsymbol{\Lambda}_T]^\top, \tag{3.7}$$

$$\boldsymbol{\psi} = [\boldsymbol{\psi}_1^\top, \boldsymbol{\psi}_2^\top, \ldots, \boldsymbol{\psi}_t^\top, \ldots, \boldsymbol{\psi}_T^\top]^\top. \tag{3.8}$$

By using this method, it would then allow one to perform a convenient manipulation process of particular articulatory movements. Moreover, it provides also two important factors, i.e. for scaling and for translation, making it possible to do any kind of alterations of the articulatory movements. That said, let us then consider that if from total $D_x$ dimensions of articulatory parameters, only 2 of them are manipulated at frame $t$, say the 1-st and the 2-nd ones. Thus, the feature vector of the manipulated articulatory parameters $\hat{\boldsymbol{x}}_t$ would be written as

$$\hat{\boldsymbol{x}}'_t = [\hat{x}'_t(1), \hat{x}'_t(2), \ldots, \hat{x}_t(d), \ldots, \hat{x}_t(D_x)]^\top \tag{3.9}$$

at frame $t$. This procedure would easily permit the manipulation of those first 2 dimensions of articulatory parameters. However, because movements of some articulators are strongly correlated to each other [45], e.g. the movements of the tip area of the tongue affects also the middle and back areas of the tongue, this method would possibly cause unnatural movements of the articulators.

### 3.3.2 Manipulation method considering inter-correlations of articulatory parameters

In order to prevent possible unnatural movements of the articulators, in this section, a manipulation method by considering inter-correlations of articulatory parameters is described. The basic idea in this second method of manipulation is by utilization of both the GMM parameters for inversion mapping and the trajectory-based conversion framework [28] that respectively capture the inter-dimensional correlation and the inter-frame correlation of the articulatory parameters. To be able to do so, the conversion process of inversion mapping, described in section 2.2.2, needs to be performed in a two stage inversion procedure. In the first stage, after the first inversion mapping, the estimated articulatory parameters are manipulated by using the simple manipulation method described in the section 3.3.1. Then, the modified components of the articulatory parameters are appended to the source features, i.e. the mel-cepstral segment features. After that, in the second stage, the second inversion mapping is performed to refine the un-modified components of the articulatory parameters using the conditional probability density function derived from the GMM of the inversion mapping.

Finally, the modified components and the refined-un-modified components of articulatory parameters are re-unified.

First, let us define a feature vector of the modified components of articulatory parameters $\hat{\boldsymbol{x}}_t^{(\omega)}$, which is written as

$$\hat{\boldsymbol{x}}_t^{(\omega)} = [\hat{x}'_t(1), \hat{x}'_t(2), \ldots, \hat{x}'_t(d), \ldots, \hat{x}'_t(D_{x^{(\omega)}})]^\top, \tag{3.10}$$

and that of the un-modified components of articulatory parameters $\hat{\boldsymbol{x}}_t^{(u)}$, which is written as

$$\hat{\boldsymbol{x}}_t^{(u)} = [\hat{x}_t(1), \hat{x}_t(2), \ldots, \hat{x}_t(d), \ldots, \hat{x}_t(D_{x^{(u)}})]^\top, \tag{3.11}$$

at frame $t$, where

$$D_{x^{(\omega)}} + D_{x^{(u)}} = D_x. \tag{3.12}$$

So that, the $2D_{x^{(\omega)}}$-dimensional and the $2D_{x^{(u)}}$-dimensional feature vectors of the joint static and dynamic modified components and un-modified components of articulatory parameters are respectively denoted as $\hat{\boldsymbol{X}}_t^{(\omega)} = [\hat{\boldsymbol{x}}_t^{(\omega)^\top}, \Delta\hat{\boldsymbol{x}}_t^{(\omega)^\top}]^\top$ and $\hat{\boldsymbol{X}}_t^{(u)} = [\hat{\boldsymbol{x}}_t^{(u)^\top}, \Delta\hat{\boldsymbol{x}}_t^{(u)^\top}]^\top$ at frame $t$. Here, similarly as in the section 2.2, the dynamic feature vectors $\Delta\hat{\boldsymbol{x}}_t^{(\omega)}$ and $\Delta\hat{\boldsymbol{x}}_t^{(u)}$ are respectively given by

$$\Delta\hat{\boldsymbol{x}}_t^{(\omega)} = \sum_{\tau=-L_-^{(1)}}^{L_+^{(1)}} w^{(1)}(\tau)\hat{\boldsymbol{x}}_{t+\tau}^{(\omega)}, \tag{3.13}$$

$$\Delta\hat{\boldsymbol{x}}_t^{(u)} = \sum_{\tau=-L_-^{(1)}}^{L_+^{(1)}} w^{(1)}(\tau)\hat{\boldsymbol{x}}_{t+\tau}^{(u)}, \tag{3.14}$$

where $w^{(1)}(\tau)$, $L_-^{(1)}$, and $L_+^{(1)}$ are the 1-st order weight coefficients and the frame lengths for computing the dynamic feature vectors. Then, the time sequence vector of the modified components of articulatory parameters $\hat{\boldsymbol{X}}^{(\omega)}$ and that of the un-modified components of articulatory parameters $\hat{\boldsymbol{X}}^{(u)}$ are respectively written as

$$\hat{\boldsymbol{X}}^{(\omega)} = \left[\hat{\boldsymbol{X}}_1^{(\omega)^\top}, \hat{\boldsymbol{X}}_2^{(\omega)^\top}, \ldots, \hat{\boldsymbol{X}}_t^{(\omega)^\top}, \ldots, \hat{\boldsymbol{X}}_T^{(\omega)^\top}\right]^\top, \tag{3.15}$$

$$\hat{\boldsymbol{X}}^{(u)} = \left[\hat{\boldsymbol{X}}_1^{(u)^\top}, \hat{\boldsymbol{X}}_2^{(u)^\top}, \ldots, \hat{\boldsymbol{X}}_t^{(u)^\top}, \ldots, \hat{\boldsymbol{X}}_T^{(u)^\top}\right]^\top. \tag{3.16}$$

Next, by performing the second stage inversion mapping, with a similar manner as in the section 2.2.2, the refined time sequence vector of the un-modified components of articulatory parameters $\hat{\hat{\boldsymbol{x}}}^{(u)}$ can be determined by

$$\hat{\hat{\boldsymbol{x}}}^{(u)} = \arg\max_{\hat{\boldsymbol{x}}^{(u)}} P(\hat{\boldsymbol{X}}^{(u)}|\boldsymbol{V}, \boldsymbol{\lambda}^{(O,X)}), \tag{3.17}$$

where

$$\begin{aligned}\boldsymbol{V} &= \left[\boldsymbol{V}_1^\top, \boldsymbol{V}_2^\top, \ldots, \boldsymbol{V}_t^\top, \ldots, \boldsymbol{V}_T^\top\right]^\top \\ &= \left[\left[\boldsymbol{O}_1, \hat{\boldsymbol{X}}_1^{(\omega)}\right]^\top, \left[\boldsymbol{O}_2, \hat{\boldsymbol{X}}_2^{(\omega)}\right]^\top, \ldots, \left[\boldsymbol{O}_t, \hat{\boldsymbol{X}}_t^{(\omega)}\right]^\top, \ldots, \left[\boldsymbol{O}_T, \hat{\boldsymbol{X}}_T^{(\omega)}\right]^\top\right]^\top. \end{aligned} \tag{3.18}$$

The above likelihood function is then written as follows:

$$\begin{aligned} P(\hat{\boldsymbol{X}}^{(u)}|\boldsymbol{V}, \boldsymbol{\lambda}^{(O,X)}) &= \sum_{\text{all } \boldsymbol{m}} P(\boldsymbol{m}|\boldsymbol{O}, \boldsymbol{\lambda}^{(O,X)}) P(\hat{\boldsymbol{X}}^{(u)}|\boldsymbol{V}, \boldsymbol{m}, \boldsymbol{\lambda}^{(O,X)}) \\ &= \prod_{t=1}^{T} \sum_{m=1}^{M} P(m|\boldsymbol{O}_t, \boldsymbol{\lambda}^{(O,X)}) P(\hat{\boldsymbol{X}}_t^{(u)}|\boldsymbol{V}_t, m, \boldsymbol{\lambda}^{(O,X)}), \end{aligned} \tag{3.19}$$

where

$$P(\hat{\boldsymbol{X}}_t^{(u)}|\boldsymbol{V}_t, m, \boldsymbol{\lambda}^{(O,X)}) = \mathcal{N}(\hat{\boldsymbol{X}}_t^{(u)}; \boldsymbol{E}_{m,t}^{(X^{(u)})}, \boldsymbol{D}_m^{(X^{(u)})}), \tag{3.20}$$

and

$$\boldsymbol{E}_{m,t}^{(X^{(u)})} = \boldsymbol{\mu}_m^{(X^{(u)})} + \boldsymbol{\Sigma}^{(X^{(u)}V)} \boldsymbol{\Sigma}_m^{(VV)^{-1}} (\boldsymbol{V}_t - \boldsymbol{\mu}_m^{(V)}), \tag{3.21}$$

$$\boldsymbol{D}_m^{(X^{(u)})} = \boldsymbol{\Sigma}_m^{(X^{(u)}X^{(u)})} - \boldsymbol{\Sigma}_m^{(X^{(u)}V)} \boldsymbol{\Sigma}_m^{(VV)^{-1}} \boldsymbol{\Sigma}_m^{(VX^{(u)})}. \tag{3.22}$$

Here, the relationship constraint between static and dynamic features of the un-modified components of articulatory parameters is given by

$$\hat{\boldsymbol{X}}^{(u)} = \boldsymbol{W}_{x^{(u)}} \hat{\boldsymbol{x}}^{(u)}, \tag{3.23}$$

where $\boldsymbol{W}_{x^{(u)}}$ is a $2D_{x^{(u)}}T$-by-$D_{x^{(u)}}T$ linear transformation matrix to append the dynamic feature vectors from the Eq. (3.14). And thus, given the initial parameter $\hat{\boldsymbol{X}}^{(u)}$, the updated parameter $\hat{\hat{\boldsymbol{X}}}^{(u)}$ can be estimated with the EM algorithm by maximizing the following auxiliary function

$$Q(\hat{\boldsymbol{X}}^{(u)}, \hat{\hat{\boldsymbol{X}}}^{(u)}) = \sum_{\text{all } \boldsymbol{m}} P(\boldsymbol{m}|\boldsymbol{O}, \hat{\boldsymbol{X}}', \boldsymbol{\lambda}^{(O,X)}) \log P(\hat{\hat{\boldsymbol{X}}}^{(u)}|\boldsymbol{V}, \boldsymbol{m}, \boldsymbol{\lambda}^{(O,X)}), \tag{3.24}$$

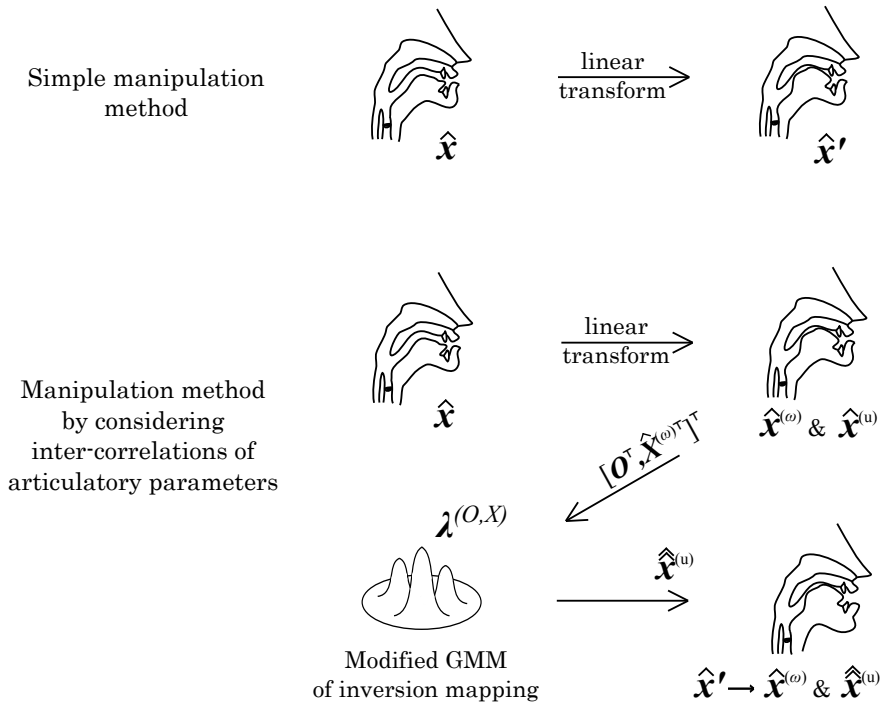



Figure 3.2. Schematic flow of the proposed methods for controlling the articulatory movements.

where $\hat{\boldsymbol{X}}'$ is composed by both the modified components of articulatory parameters $\hat{\boldsymbol{X}}^{(\omega)}$ and the un-modified ones $\hat{\boldsymbol{X}}^{(u)}$.

In this thesis, the likelihood function in Eq. (3.17) is again approximated with a single mixture component sequence as follows:

$$P(\hat{\boldsymbol{X}}^{(u)}|\boldsymbol{V},\boldsymbol{\lambda}^{(O,X)}) \simeq P(\boldsymbol{m}|\boldsymbol{O},\boldsymbol{\lambda}^{(O,X)})P(\hat{\boldsymbol{X}}^{(u)}|\boldsymbol{V},\boldsymbol{m},\boldsymbol{\lambda}^{(O,X)}). \tag{3.25}$$

First, the sub-optimum mixture component sequence $\hat{\boldsymbol{m}}^{(O)} = \{\hat{m}_1^{(O)}, \hat{m}_2^{(O)}, \ldots, \hat{m}_t^{(O)}, \ldots, \hat{m}_T^{(O)}\}$ is determined by the Eq. (2.34). The maximization of the auxiliary function is then approximated as follows:

$$\begin{aligned} Q(\hat{\boldsymbol{X}}^{(u)}, \hat{\hat{\boldsymbol{X}}}^{(u)}) &\simeq \log P(\hat{\boldsymbol{m}}^{(O)}|\boldsymbol{O},\boldsymbol{\lambda}^{(O,X)})P(\hat{\hat{\boldsymbol{X}}}^{(u)}|\boldsymbol{V},\hat{\boldsymbol{m}}^{(O)},\boldsymbol{\lambda}^{(O,X)}) \\ &= -\frac{1}{2}\hat{\hat{\boldsymbol{x}}}^{(u)\top}\boldsymbol{W}_{x^{(u)}}^{\top}\boldsymbol{D}_{\hat{\boldsymbol{m}}^{(O)}}^{(X^{(u)})^{-1}}\boldsymbol{W}_{x^{(u)}}\hat{\hat{\boldsymbol{x}}}^{(u)} + \hat{\hat{\boldsymbol{x}}}^{(u)\top}\boldsymbol{W}_{x^{(u)}}^{\top}\boldsymbol{D}_{\hat{\boldsymbol{m}}^{(O)}}^{(X^{(u)})^{-1}}\boldsymbol{E}_{\hat{\boldsymbol{m}}^{(O)}}^{(X)(u)} + K'. \end{aligned} \tag{3.26}$$

Then, the refined time sequence vector of the un-modified components of articulatory parameters $\hat{\tilde{\boldsymbol{x}}}^{(u)} = \left[\hat{\tilde{\boldsymbol{x}}}_1^{(u)\top}, \hat{\tilde{\boldsymbol{x}}}_2^{(u)\top}, \ldots, \hat{\tilde{\boldsymbol{x}}}_t^{(u)\top}, \ldots, \hat{\tilde{\boldsymbol{x}}}_T^{(u)\top}\right]^\top$ is given by

$$\hat{\tilde{\boldsymbol{x}}}^{(u)} = (\boldsymbol{W}_{x^{(u)}}^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}^{(O)}}^{(X^{(u)})^{-1}} \boldsymbol{W}_{x^{(u)}})^{-1} \boldsymbol{W}_{x^{(u)}}^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}^{(O)}}^{(X^{(u)})^{-1}} \boldsymbol{E}_{\hat{\boldsymbol{m}}^{(O)}}^{(X^{(u)})}. \tag{3.27}$$

So that, the time sequence vector of the manipulated articulatory parameters $\hat{\boldsymbol{x}}'$ would be written as

$$\hat{\boldsymbol{x}}' = [\hat{\boldsymbol{x}}'_1, \hat{\boldsymbol{x}}'_2, \ldots, \hat{\boldsymbol{x}}'_t, \ldots, \hat{\boldsymbol{x}}'_T]^\top, \tag{3.28}$$

where

$$\hat{\boldsymbol{x}}'_t = [\hat{x}'_t(1), \hat{x}'_t(2), \ldots, \hat{x}'_t(d), \ldots, \hat{x}'_t(D_x)]^\top, \tag{3.29}$$

and

$$\hat{x}'_t(d) \in \hat{\boldsymbol{x}}_t^{(\omega)} \text{ or } \hat{x}'_t(d) \in \hat{\tilde{\boldsymbol{x}}}_t^{(u)}. \tag{3.30}$$

Schematic flow of both simple manipulation method and the manipulation method by considering inter-correlations between articulatory parameters are shown in Fig. 3.2.

## 3.4. Modified Speech Generation Process with Direct Waveform Modification Methods using Spectrum Differential

In order to be able to generate modified speech signal, the articulatory-to-acoustic production mapping, described in the section 2.3.2 is performed. At frame $t$, given the feature vector of manipulated articulatory parameters $\hat{\boldsymbol{x}}'_t$ and that of the source-excitation parameters, $\boldsymbol{s}_t$, the $2(D_x + D_s)$-dimensional feature vector $\boldsymbol{Y}'_t = [\hat{\boldsymbol{x}}'^\top_t, \boldsymbol{s}_t^\top, \Delta\hat{\boldsymbol{x}}'^\top_t, \Delta\boldsymbol{s}_t^\top]^\top$ is developed. So that, the time sequence vector consisting of the manipulated articulatory parameters would be denoted as $\boldsymbol{Y}' = [\boldsymbol{Y}'^\top_1, \boldsymbol{Y}'^\top_2, \ldots, \boldsymbol{Y}'^\top_t, \ldots, \boldsymbol{Y}'^\top_T]^\top$. Then, similarly, as in the Eq. (2.52), the time sequence of the modified acoustic parameters $\hat{\boldsymbol{c}}'$ would be determined by

$$\hat{\boldsymbol{c}}' = \arg\max_{\boldsymbol{c}'} P(\boldsymbol{C}'|\boldsymbol{Y}', \boldsymbol{\lambda}^{(Y,C)}), \text{ s.t. } \boldsymbol{C}' = \boldsymbol{W}_c \boldsymbol{c}', \tag{3.31}$$

where

$$P(\boldsymbol{C}'|\boldsymbol{Y}', \boldsymbol{\lambda}^{(Y,C)}) = \sum_{\text{all } \boldsymbol{m}} P(\boldsymbol{m}|\boldsymbol{Y}', \boldsymbol{\lambda}^{(Y,C)}) P(\boldsymbol{C}'|\boldsymbol{Y}', \boldsymbol{m}, \boldsymbol{\lambda}^{(Y,C)})$$
$$= \prod_{t=1}^{T} \sum_{m=1}^{M} P(m|\boldsymbol{Y}'_t, \boldsymbol{\lambda}^{(Y,C)}) P(\boldsymbol{C}'_t|\boldsymbol{Y}'_t, m, \boldsymbol{\lambda}^{(Y,C)}), \tag{3.32}$$

$$P(m|\boldsymbol{Y}'_t, \boldsymbol{\lambda}^{(Y,C)}) = \frac{\alpha_m \mathcal{N}(\boldsymbol{Y}'_t; \boldsymbol{\mu}_m^{(Y)}, \boldsymbol{\Sigma}_m^{(YY)})}{\sum_{n=1}^{M} \alpha_n \mathcal{N}(\boldsymbol{Y}'_t; \boldsymbol{\mu}_n^{(Y)}, \boldsymbol{\Sigma}_n^{(YY)})}, \tag{3.33}$$

$$P(\boldsymbol{C}'_t|\boldsymbol{Y}'_t, m, \boldsymbol{\lambda}^{(Y,C)}) = \mathcal{N}(\boldsymbol{C}'_t; \boldsymbol{E}_{m,t}^{(C')}, \boldsymbol{D}_m^{(C)}), \tag{3.34}$$

and

$$\boldsymbol{E}_{m,t}^{(C')} = \boldsymbol{\mu}_m^{(C)} + \boldsymbol{\Sigma}_m^{(CY)} \boldsymbol{\Sigma}_m^{(YY)^{-1}} (\boldsymbol{Y}'_t - \boldsymbol{\mu}_m^{(Y)}). \tag{3.35}$$

So that, given the current parameter $\boldsymbol{C}'$, an updated parameter $\hat{\boldsymbol{C}}'$ can be estimated with EM algorithm by maximizing the following auxiliary function

$$Q(\boldsymbol{C}', \hat{\boldsymbol{C}}') = \sum_{\text{all } \boldsymbol{m}} P(\boldsymbol{m}|\boldsymbol{Y}', \boldsymbol{C}', \boldsymbol{\lambda}^{(Y,C)}) \log P(\hat{\boldsymbol{C}}'|\boldsymbol{Y}', \boldsymbol{m}, \boldsymbol{\lambda}^{(Y,C)}). \tag{3.36}$$

Again, in this thesis, an approximation of the likelihood function in Eq. (3.31) is deployed with a single mixture component sequence as follows:

$$P(\boldsymbol{C}'|\boldsymbol{Y}', \boldsymbol{\lambda}^{(Y,C)}) \simeq P(\boldsymbol{m}|\boldsymbol{Y}', \boldsymbol{\lambda}^{(Y,C)}) P(\boldsymbol{C}'|\boldsymbol{Y}', \boldsymbol{m}, \boldsymbol{\lambda}^{(Y,C)}). \tag{3.37}$$

First, the sub-optimum mixture component sequence $\hat{\boldsymbol{m}}^{(Y')} = \{\hat{m}_1^{(Y')}, \hat{m}_2^{(Y')}, \ldots, \hat{m}_t^{(Y')}, \ldots, \hat{m}_T^{(Y')}\}$ is determined by

$$\hat{\boldsymbol{m}}^{(Y')} = \arg\max_{\boldsymbol{m}^{(Y')}} P(\boldsymbol{m}^{(Y')}|\boldsymbol{Y}', \boldsymbol{\lambda}^{(Y,C)}). \tag{3.38}$$

Then the maximization of the approximated auxiliary function is defined by

$$Q(\boldsymbol{C}', \hat{\boldsymbol{C}}') \simeq \log P(\hat{\boldsymbol{m}}^{(Y')}|\boldsymbol{Y}', \boldsymbol{\lambda}^{(O,X)}) P(\hat{\boldsymbol{C}}'|\boldsymbol{Y}', \hat{\boldsymbol{m}}^{(Y')}, \boldsymbol{\lambda}^{(Y,C)})$$
$$= -\frac{1}{2} \hat{\boldsymbol{c}}'^{\top} \boldsymbol{W}_c^{\top} \boldsymbol{D}_{\hat{\boldsymbol{m}}^{(Y')}}^{(C)^{-1}} \boldsymbol{W}_c \hat{\boldsymbol{c}}' + \hat{\boldsymbol{c}}'^{\top} \boldsymbol{W}_c^{\top} \boldsymbol{D}_{\hat{\boldsymbol{m}}^{(Y')}}^{(C)^{-1}} \boldsymbol{E}_{\hat{\boldsymbol{m}}^{(Y')}}^{(C')} + K', \tag{3.39}$$

where

$$\boldsymbol{E}_{\hat{\boldsymbol{m}}^{(Y')}}^{(C')} = \left[\boldsymbol{E}_{\hat{m}_1^{(Y')},1}^{(C')^{\top}}, \boldsymbol{E}_{\hat{m}_2^{(Y')},2}^{(C')^{\top}}, \ldots, \boldsymbol{E}_{\hat{m}_t^{(Y')},t}^{(C')^{\top}}, \ldots, \boldsymbol{E}_{\hat{m}_T^{(Y')},T}^{(C')^{\top}}\right]^{\top}, \tag{3.40}$$

$$\boldsymbol{D}_{\hat{\boldsymbol{m}}^{(Y')}}^{(C)^{-1}} = \text{diag}\left[\boldsymbol{D}_{\hat{m}_1^{(Y')}}^{(C)^{-1}}, \boldsymbol{D}_{\hat{m}_2^{(Y')}}^{(C)^{-1}}, \ldots, \boldsymbol{D}_{\hat{m}_t^{(Y')}}^{(C)^{-1}}, \ldots, \boldsymbol{D}_{\hat{m}_T^{(Y')}}^{(C)^{-1}}\right]. \tag{3.41}$$

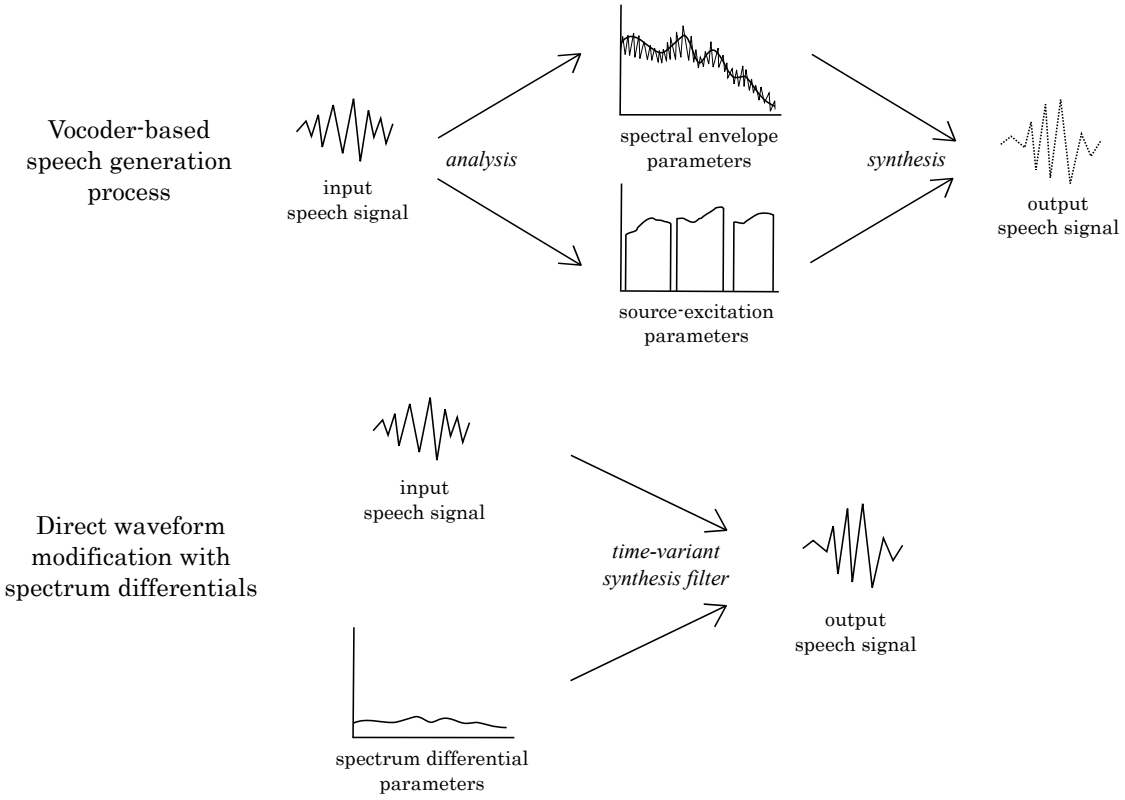



Figure 3.3. Schematic flow of the vocoder-based speech generation process and the direct waveform modification using spectrum differentials.

Finally, the time sequence vector of the modified acoustic parameters $\hat{\boldsymbol{c}}'$ is given by

$$\hat{\boldsymbol{c}}' = (\boldsymbol{W}_c^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}^{(Y')}}^{(C)^{-1}} \boldsymbol{W}_c)^{-1} \boldsymbol{W}_c^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}^{(Y')}}^{(C)^{-1}} \boldsymbol{E}_{\hat{\boldsymbol{m}}^{(Y')}}^{(C')}. \tag{3.42}$$

Given the modified acoustic features, i.e. modified mel-cepstrum, $\hat{\boldsymbol{c}}'$, the modified speech signal can then be generated by utilizing the waveform generation process based on a vocoder. In order to do that, the original excitation parameters, included in the feature sequence vector $\boldsymbol{s}$, need to be used. This process would then cause degradation of the quality of the generated speech signal. It is mainly due to the use of the vocoder-based excitation generation process, which is sensitive to the errors from the parameters extraction [32].

To alleviate the degradation of the quality, in this chapter, the direct waveform modification methods are described by avoiding the vocoder framework in generating the excitation signal through direct filtering of the original wave-

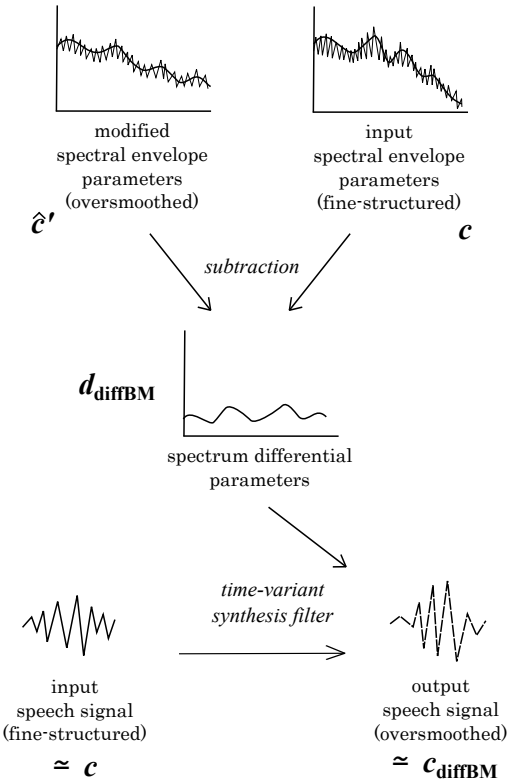



Figure 3.4. Process flow of the basic direct waveform modification method.

form with spectrum differential parameters. Figure 3.3 illustrates the difference between the vocoder-based speech generation process and the direct waveform modification-based using spectrum differential. Here, the spectrum differential parameters represent the value differences between the original spectrum and the modified one. Specifically, three direct waveform modification methods are proposed, which are different to one another in the sense of the procedure for producing the spectrum differential parameters.

### 3.4.1 Basic direct waveform modification method (diffBM)

In the basic direct waveform modification method (diffBM), the time sequence vector of spectrum differential parameters $\boldsymbol{d}_{\text{diffBM}}$ is written as

$$\begin{aligned}\boldsymbol{d}_{\text{diffBM}} &= \hat{\boldsymbol{c}}' - \boldsymbol{c} \\ &= \left[[\hat{\boldsymbol{c}}'_1 - \boldsymbol{c}_1]^\top, [\hat{\boldsymbol{c}}'_2 - \boldsymbol{c}_2]^\top, \ldots, [\hat{\boldsymbol{c}}'_t - \boldsymbol{c}_t]^\top, \ldots, [\hat{\boldsymbol{c}}'_T - \boldsymbol{c}_T]^\top\right]^\top, \end{aligned} \tag{3.43}$$

where feature vector of the original spectrum parameters and that of the modified ones are respectively denoted as $\boldsymbol{c}_t$ and $\hat{\boldsymbol{c}}'_t$ at frame $t$. Then, through a filtering procedure, the original speech waveform is modified according to the spectrum diferential parameters in $\boldsymbol{d}_{\text{diffBM}}$. The modified speech waveform of the diffBM method is then characterized with a time sequence vector of spectrum parameters $\boldsymbol{c}_{\text{diffBM}}$ as follows:

$$\begin{aligned}\boldsymbol{c}_{\text{diffBM}} &= \boldsymbol{c} + \boldsymbol{d}_{\text{diffBM}} \\ &= \big[[\boldsymbol{c}_1 + (\hat{\boldsymbol{c}}'_1 - \boldsymbol{c}_1)]^\top, [\boldsymbol{c}_2 + (\hat{\boldsymbol{c}}'_2 - \boldsymbol{c}_2)]^\top, \ldots, \\ &\qquad [\boldsymbol{c}_t + (\hat{\boldsymbol{c}}'_t - \boldsymbol{c}_t)]^\top, \ldots, [\boldsymbol{c}_T + (\hat{\boldsymbol{c}}'_T - \boldsymbol{c}_T)]^\top\big]^\top \\ &= [\hat{\boldsymbol{c}}'^\top_1, \hat{\boldsymbol{c}}'^\top_2, \ldots, \hat{\boldsymbol{c}}'^\top_t, \ldots, \hat{\boldsymbol{c}}'^\top_T]^\top. \end{aligned} \tag{3.44}$$

Therefore, the modified speech waveform of the diffBM method would still be defined by the oversmoothed modified spectrum parameters $\hat{\boldsymbol{c}}'$ as in the case with using the conventional system (vocoder-based process). However, it is completely different from that of the conventional system in terms of the excitation signal due to the direct filtering procedure of the original speech waveform without using the vocoder-based excitation generation. The process flow of the basic direct waveform modification method is shown in Fig. 3.4

### 3.4.2 Refined direct waveform modification method (diffRM)

In the refined direct waveform modification method (diffRM), the time sequence vector of spectrum differential parameters $\boldsymbol{d}_{\text{diffRM}}$ is written as

$$\begin{aligned}\boldsymbol{d}_{\text{diffRM}} &= \hat{\boldsymbol{c}}' - \hat{\boldsymbol{c}} \\ &= \big[[\hat{\boldsymbol{c}}'_1 - \hat{\boldsymbol{c}}_1]^\top, [\hat{\boldsymbol{c}}'_2 - \hat{\boldsymbol{c}}_2]^\top, \ldots, [\hat{\boldsymbol{c}}'_t - \hat{\boldsymbol{c}}_t]^\top, \ldots, [\hat{\boldsymbol{c}}'_T - \hat{\boldsymbol{c}}_T]^\top\big]^\top, \end{aligned} \tag{3.45}$$

where feature vector of the oversmoothed original spectrum parameters is written as $\hat{\boldsymbol{c}}_t$ at frame $t$. Then, again, through a filtering procedure, the original speech waveform is modified according to the spectrum diferential parameters in $\boldsymbol{d}_{\text{diffRM}}$. The modified speech waveform of the diffRM method is then characterized with

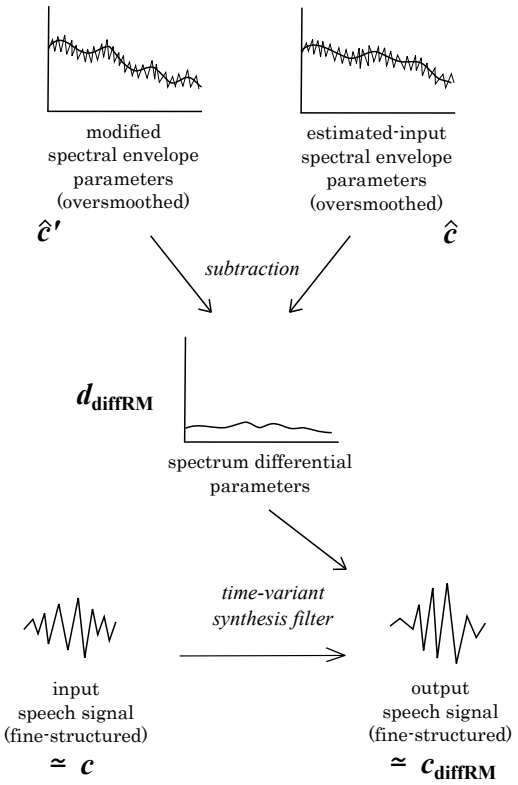



Figure 3.5. Process flow of the refined direct waveform modification method.

a time sequence vector of spectrum parameters $\boldsymbol{c}_{\text{diffRM}}$ as follows:

$$\begin{aligned}\boldsymbol{c}_{\text{diffRM}} &= \boldsymbol{c} + \boldsymbol{d}_{\text{diffRM}} \\ &= \big[[\boldsymbol{c}_1 + (\hat{\boldsymbol{c}}'_1 - \hat{\boldsymbol{c}}_1)]^\top, [\boldsymbol{c}_2 + (\hat{\boldsymbol{c}}'_2 - \hat{\boldsymbol{c}}_2)]^\top, \ldots, \\ &\qquad [\boldsymbol{c}_t + (\hat{\boldsymbol{c}}'_t - \hat{\boldsymbol{c}}_t)]^\top, \ldots, [\boldsymbol{c}_T + (\hat{\boldsymbol{c}}'_T - \hat{\boldsymbol{c}}_T)]^\top\big]^\top \\ &= \big[[\hat{\boldsymbol{c}}'_1 + \boldsymbol{\epsilon}_1]^\top, [\hat{\boldsymbol{c}}'_2 + \boldsymbol{\epsilon}_2]^\top, \ldots, [\hat{\boldsymbol{c}}'_t + \boldsymbol{\epsilon}_t]^\top, \ldots, [\hat{\boldsymbol{c}}'_T + \boldsymbol{\epsilon}_T]^\top\big]^\top, \end{aligned} \tag{3.46}$$

where the refining factor is denoted as $\boldsymbol{\epsilon}_t = \boldsymbol{c}_t - \hat{\boldsymbol{c}}_t$ at frame $t$.

Therefore, the modified speech waveform of the diffRM method is defined by not only the oversmoothed spectrum parameters $\hat{\boldsymbol{c}}'_t$, but also by the residual given by $\boldsymbol{\epsilon}_t$ at each frame $t$, that refine the overall structure of the spectrum. The process flow of the refined direct waveform modification method is shown in Fig. 3.5

### 3.4.3 Refined method using differential GMM (diffGMM)

In the previous section, the spectrum differential parameters in $\boldsymbol{d}_{\text{diffRM}}$ provides refining factors to alleviate the oversmoothing effect of the modified spectrum parameters. However, in order to generate these parameters, the acoustic-to-articulatory production mapping needs to be performed twice, i.e. to estimate the oversmoothed spectrum of the original speech $\hat{\boldsymbol{c}}$ and to estimated the modified spectrum $\hat{\boldsymbol{c}}'$. In this method, similar characteristics of the spectrum differential parameters can be achieved by performing the mapping procedure only once through the use of a differential GMM (diffGMM), which is analytically derived from the subtraction of two normally-distributed independent random variables to generate the spectrum differential parameters.

Then, let us define the time sequence vector of spectrum differential parameters with differential GMM as $\boldsymbol{g}$, which is determined by

$$\begin{aligned} \hat{\boldsymbol{g}} &= \arg\max_{\boldsymbol{g}} P(\boldsymbol{G}|\boldsymbol{Y}', \boldsymbol{Y}, \boldsymbol{\lambda}^{(Y,C)}), \\ \text{s.t. } \boldsymbol{G} &= \boldsymbol{C}' - \boldsymbol{C} \text{ and } \boldsymbol{G} = \boldsymbol{W}_c \boldsymbol{g}, \end{aligned} \tag{3.47}$$

where

$$\begin{aligned} P(\boldsymbol{G}|\boldsymbol{Y}', \boldsymbol{Y}, \boldsymbol{\lambda}^{(Y,C)}) = & \sum_{\text{all } \boldsymbol{m}^{(Y')}} \sum_{\text{all } \boldsymbol{m}^{(Y)}} P(\boldsymbol{m}^{(Y')}, \boldsymbol{m}^{(Y)}|\boldsymbol{Y}', \boldsymbol{Y}, \boldsymbol{\lambda}^{(Y,C)}) \\ & P(\boldsymbol{G}|\boldsymbol{Y}', \boldsymbol{Y}, \boldsymbol{m}^{(Y')}, \boldsymbol{m}^{(Y)}, \boldsymbol{\lambda}^{(Y,C)}) \\ = & \prod_{t=1}^{T} \sum_{m^{(Y')}=1}^{M} \sum_{m^{(Y)}=1}^{M} P(m^{(Y')}|\boldsymbol{Y}'_t, \boldsymbol{\lambda}^{(Y,C)}) P(m^{(Y)}|\boldsymbol{Y}_t, \boldsymbol{\lambda}^{(Y,C)}) \\ & P(\boldsymbol{G}_t|\boldsymbol{Y}'_t, \boldsymbol{Y}_t, m^{(Y')}, m^{(Y)}, \boldsymbol{\lambda}^{(Y,C)}), \end{aligned} \tag{3.48}$$

and

$$P(\boldsymbol{G}_t|\boldsymbol{Y}'_t, \boldsymbol{Y}_t, m^{(Y')}, m^{(Y)}, \boldsymbol{\lambda}^{(Y,C)}) = \mathcal{N}(\boldsymbol{G}_t; \boldsymbol{E}^{(G)}_{m^{(Y')},m^{(Y)},t}, \boldsymbol{D}^{(G)}_{m^{(Y')},m^{(Y)}}), \tag{3.49}$$

$$\boldsymbol{E}^{(G)}_{m^{(Y')},m^{(Y)},t} = \boldsymbol{E}^{(C')}_{m^{(Y')},t} - \boldsymbol{E}^{(C)}_{m^{(Y)},t}, \tag{3.50}$$

$$\boldsymbol{D}^{(G)}_{m^{(Y')},m^{(Y')}} = \boldsymbol{D}^{(C)}_{m^{(Y')}} + \boldsymbol{D}^{(C)}_{m^{(Y)}}. \tag{3.51}$$

So that, given the current parameter $\boldsymbol{G}$, an updated parameter $\hat{\boldsymbol{G}}$ can be esti-

mated with EM algorithm by maximizing the following auxiliary function

$$Q(\boldsymbol{G},\hat{\boldsymbol{G}}) = \sum_{\text{all } \boldsymbol{m}^{(Y')}} \sum_{\text{all } \boldsymbol{m}^{(Y)}} P(\boldsymbol{m}^{(Y')},\boldsymbol{m}^{(Y)}|\boldsymbol{Y}',\boldsymbol{Y},\boldsymbol{C}',\boldsymbol{\lambda}^{(Y,C)}) \log P(\hat{\boldsymbol{G}}|\boldsymbol{Y}',\boldsymbol{Y},\boldsymbol{m}^{(Y')},\boldsymbol{m}^{(Y)},\boldsymbol{\lambda}^{(Y,C)}). \tag{3.52}$$

In this thesis, an approximation of the likelihood function in Eq. (3.47) is deployed with single mixture component sequences as follows:

$$P(\boldsymbol{G}|\boldsymbol{Y}',\boldsymbol{Y},\boldsymbol{\lambda}^{(Y,C)}) \simeq P(\boldsymbol{m}^{(Y')},\boldsymbol{m}^{(Y)}|\boldsymbol{Y}',\boldsymbol{Y},\boldsymbol{\lambda}^{(Y,C)}) P(\boldsymbol{G}|\boldsymbol{Y}',\boldsymbol{Y},\boldsymbol{m}^{(Y')},\boldsymbol{m}^{(Y)},\boldsymbol{\lambda}^{(Y,C)}). \tag{3.53}$$

The sub-optimum mixture component sequences $\hat{\boldsymbol{m}}^{(Y')} = \{\hat{m}_1^{(Y')}, \hat{m}_2^{(Y')}, \ldots, \hat{m}_t^{(Y')}, \ldots, \hat{m}_T^{(Y')}\}$ and $\hat{\boldsymbol{m}}^{(Y)} = \{\hat{m}_1^{(Y)}, \hat{m}_2^{(Y)}, \ldots, \hat{m}_t^{(Y)}, \ldots, \hat{m}_T^{(Y)}\}$ are determined by the Eq. (3.38) and Eq. (2.70), respectively. Then, the maximization of the approximated auxiliary function is defined by

$$\begin{aligned} Q(\boldsymbol{G},\hat{\boldsymbol{G}}) &\simeq \log P(\hat{\boldsymbol{m}}^{(Y')},\hat{\boldsymbol{m}}^{(Y)}|\boldsymbol{Y}',\boldsymbol{Y},\boldsymbol{\lambda}^{(O,X)}) P(\hat{\boldsymbol{G}}|\boldsymbol{Y}',\boldsymbol{Y},\hat{\boldsymbol{m}}^{(Y')},\hat{\boldsymbol{m}}^{(Y)},\boldsymbol{\lambda}^{(Y,C)}) \\ &= -\frac{1}{2}\hat{\boldsymbol{g}}^\top \boldsymbol{W}_c^\top \boldsymbol{D}^{(G)^{-1}}_{\hat{\boldsymbol{m}}^{(Y')},\hat{\boldsymbol{m}}^{(Y)}} \boldsymbol{W}_c \hat{\boldsymbol{g}} + \hat{\boldsymbol{g}}^\top \boldsymbol{W}_c^\top \boldsymbol{D}^{(G)^{-1}}_{\hat{\boldsymbol{m}}^{(Y')},\hat{\boldsymbol{m}}^{(Y)}} \boldsymbol{E}^{(G)}_{\hat{\boldsymbol{m}}^{(Y')},\hat{\boldsymbol{m}}^{(Y)}} + K', \end{aligned} \tag{3.54}$$

where

$$\boldsymbol{E}^{(G)}_{\hat{\boldsymbol{m}}^{(Y')},\hat{\boldsymbol{m}}^{(Y)}} = \big[\boldsymbol{E}^{(G)^\top}_{\hat{m}_1^{(Y')},\hat{m}_1^{(Y)},1}, \boldsymbol{E}^{(G)^\top}_{\hat{m}_2^{(Y')},\hat{m}_2^{(Y)},2}, \ldots, \boldsymbol{E}^{(G)^\top}_{\hat{m}_t^{(Y')},\hat{m}_t^{(Y)},t}, \ldots, \boldsymbol{E}^{(G)^\top}_{\hat{m}_T^{(Y')},\hat{m}_T^{(Y)},T}\big]^\top, \tag{3.55}$$

$$\boldsymbol{D}^{(G)^{-1}}_{\hat{\boldsymbol{m}}^{(Y')},\hat{\boldsymbol{m}}^{(Y)}} = \text{diag}\,\big[\boldsymbol{D}^{(G)^{-1}}_{\hat{m}_1^{(Y')},\hat{m}_1^{(Y)}}, \boldsymbol{D}^{(G)^{-1}}_{\hat{m}_2^{(Y')},\hat{m}_2^{(Y)}}, \ldots, \boldsymbol{D}^{(G)^{-1}}_{\hat{m}_t^{(Y')},\hat{m}_t^{(Y)}}, \ldots, \boldsymbol{D}^{(G)^{-1}}_{\hat{m}_T^{(Y')},\hat{m}_T^{(Y)}}\big]. \tag{3.56}$$

Finally, the time sequence vector of the spectrum differential parameters with this differential GMM method $\hat{\boldsymbol{g}}$ would be given by

$$\hat{\boldsymbol{g}} = (\boldsymbol{W}_c^\top \boldsymbol{D}^{(G)^{-1}}_{\hat{\boldsymbol{m}}^{(Y')},\hat{\boldsymbol{m}}^{(Y)}} \boldsymbol{W}_c)^{-1} \boldsymbol{W}_c^\top \boldsymbol{D}^{(G)^{-1}}_{\hat{\boldsymbol{m}}^{(Y')},\hat{\boldsymbol{m}}^{(Y)}} \boldsymbol{E}^{(G)}_{\hat{\boldsymbol{m}}^{(Y')},\hat{\boldsymbol{m}}^{(Y)}}. \tag{3.57}$$

The process flow of the refined direct waveform modification method with differential GMM is shown in Fig. 3.6 The modified speech waveform of the differential GMM method is characterized with a time sequence vector of spectrum parameters $\boldsymbol{c}_{\text{diffGMM}}$, which is written as

$$\boldsymbol{c}_{\text{diffGMM}} = \boldsymbol{c} + \hat{\boldsymbol{g}}. \tag{3.58}$$

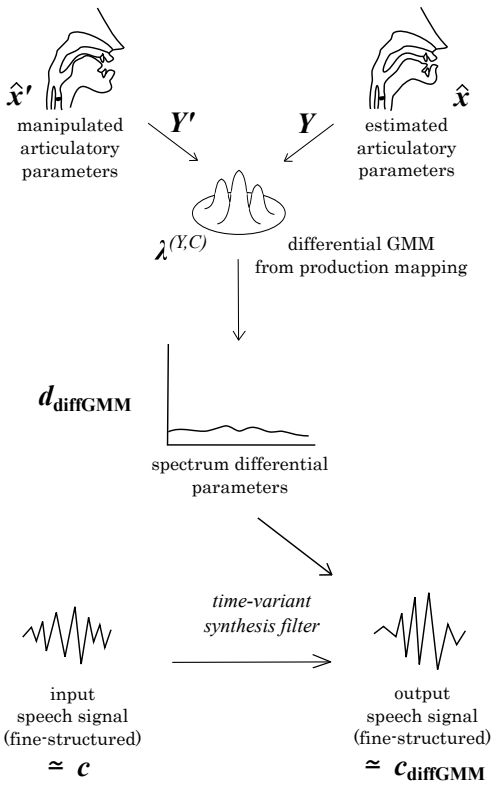



Figure 3.6. Process flow of the refined direct waveform modification method.

The characteristics of the spectrum differential parameters in $\boldsymbol{c}_{\mathrm{difGMM}}$ are the same as in the $\boldsymbol{c}_{\mathrm{diffRM}}$. Moreover, because of its convenient procedure by the utilization of the GMM parameters in the production mapping, it would be straightforward to further apply some additional techniques, such as the GV modelling [30, 46] or the modulation spectrum [47]. In brief, the differences between the three proposed direct waveform modification methods and the conventional vocoder-based method are shown in the Table 3.1.

## 3.5. Summary

In this chapter, the proposed articulatory controllable speech modification system using Gaussian mixture model (GMM) has been elaborated. The proposed system is capable of achieving a speech modification procedure by making it possible to manipulate the unobserved articulatory movements from the input speech

Table 3.1. Comparison of several traits between the proposed direct waveform modification methods and also the vocoder-based method

| trait \ method | vocoder | diffBM | diffRM | diffGMM |
|---|---|---|---|---|
| input #1 | spectrum | speech | speech | speech |
| input #2 | excitation | spectrum diff. | spectrum diff. | spectrum diff. |
| spectrum diff. | - | $\hat{\boldsymbol{c}}' - \boldsymbol{c}$ | $\hat{\boldsymbol{c}}' - \hat{\boldsymbol{c}}$ | $\hat{\boldsymbol{c}}' - \hat{\boldsymbol{c}}$ |
| structure | oversmoothed | less-oversmoothed | fine-structured | fine-structured |
| quality | very low | high | very high | very high |
| # of prod. map. | once | twice | twice | once |

signal. To do so, the GMM-based acoustic-to-articulatory inversion mapping and the GMM-based articulatory-to-acoustic production mapping are integrated in a sequential procedure, while allowing one to manually manipulate the articulatory parameters. Moreover, a manipulation method for controlling articulatory parameters by considering their inter-correlations to produce more natural movements is also described. Lastly, modified speech generation procedures that alleviate the quality degradation of vocoder-based process through directly filtering the speech waveform with spectrum differences have also been explained.

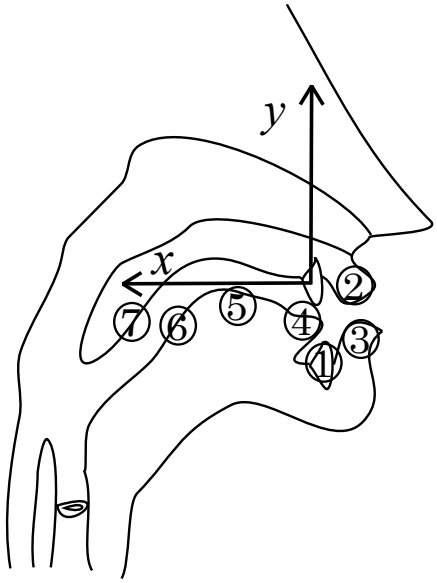
# Chapter 4

# Experimental Evaluation

## 4.1. Speech and Articulatory Data

In this thesis, we used the Multichannel Articulatory (MOCHA) [48] database as a set of simultaneously recorded speech and articulatory data. This database is accessible from the Centre for Speech and Technology, University of Edinburgh. The MOCHA database consists of one British male speaker and one British female speaker data. Each of the speaker uttered a set of 460 British TIMIT sentences and the sampling rate of this speech data is 16 kHz. The whole recording procedures were done in the Queen Margaret University, Edinburgh.

In the MOCHA database, the electromagnetic articulograph (EMA) device was used to record the movements of the articulators during speaking. To do that, seven coils were attached onto the articulators, i.e. lower incisor, upper lip, lower lip, tongue tip, tongue body, tongue dorsum, and velum, to record their movements. In addition, two other coils were attached onto the bridge of the nose and the upper incisor as reference points. The movements of the seven articulators were then recorded in on midsagittal plane at 500 Hz. Their positions over-time were represented on the $x$- and $y$-coordinates. These data were normalized to alleviate the noise effect from the measurements [49]. Figure 4.1 illustrates the cartesian coordinate used to represent the movements of the articulators.




1: LI = lower incisor
2: UL = upper lip
3: LL = lower lip
4: TT = tongue tip
5: TB = tongue body
6: TD = tongue dorsum
7: V = velum

Figure 4.1. Placement of coils for measuring EMA data and its cartesian coordinate with upper incisor as the origin.

## 4.2. Experimental Conditions

In speech acoustic parameter extraction, we used the STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) analysis method [50] to calculate the spectral envelope at each frame. It was then converted into the 1-st through 24−th mel-cepstral coefficients as the spectral envelope parameters. The current ± 10 frames were used to extract the mel-cepstral segments as described in section 2.2. For 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 [51] in STRAIGHT was employed to extract $F_0$ values. As for the articulatory parameters, we used the 14-dimensional EMA

data described in the section 4.1. These 14-dimensional articulatory data were converted into z-score (zero mean and unit variance). The frame shift was set to 5 ms.

In the experiments, we trained the Gaussian mixture models (GMMs) of the acoustic-to-articulatory inversion mapping and the articulatory-to-acoustic production mapping for each of the male and female speaker. The number of the GMMs, therefore, are four. From the 460 utterances of each speaker, we used 350 of them for training the GMMs and the remaining 110 for the evaluations. In the speech signal generation procedure, the mel log spectrum approximation (MLSA) filter [52] was used as the synthesis filter for direct waveform modification methods and the global variance (GV) [53] was considered in only the conventional vocoder-based process.

In order to evaluate the proposed articulatory controllable speech modification method, we conducted four different experiments. In the first evaluation, we investigated the accuracies of the GMM-based acoustic-to-articulatory inversion mapping, the articulatory-to-acoustic production mapping, and the proposed sequential mapping system using both of those mappings. In the second experiment, we performed subjective evaluation to compare the performance of the proposed methods for manipulating the articulatory movements, described in the section 3.3. In the third evaluation, we performed another subjective evaluations to compare the effectiveness of the proposed direct waveform modification methods with spectrum differential, described in the section 3.4, in generating modified speech sounds. In the final experiment, we performed a categorical perception evaluation to evaluate the controllability of the proposed system in modifying particular phonemic sounds through articulatory control.

## 4.3. Investigation on Mapping Accuracy between Acoustic and Articulatory Parameters

To objectively evaluate the performance of the proposed system, we investigated the accuracies of the acoustic-to-articulatory inversion mapping, the articulatory-to-acoustic production mapping, and the proposed sequential mapping system of them. In the evaluation on inversion mapping, we measured the errors between

Table 4.1. Average root-mean-square (RMS) error [mm] of estimated articulatory parameters for male and female speakers with varying number of mixture components from 1 to 128.

| # mix. comp. / speaker | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 |
|---|---|---|---|---|---|---|---|---|
| male | 1.98 | 1.87 | 1.77 | 1.68 | 1.54 | 1.47 | 1.43 | **1.42** |
| female | 1.90 | 1.80 | 1.75 | 1.65 | 1.56 | 1.48 | 1.43 | **1.41** |

measured articulatory movements and the estimated ones. On the other hand, in both the evaluation on production mapping and that of the proposed sequential mapping, the errors between extracted spectrum parameters and the estimated ones are measured. Both of the male and female speakers' data were used on all objective evaluations. The best number of mixture components were used for the subjective evaluations in the succeeding section.

### 4.3.1 Objective evaluation on inversion mapping

In the first objective evaluation, we evaluated the accuracy of the inversion mapping system in estimating the articulatory parameters. The number of mixture components in this evaluation was varied from 1 to 128. First, the accuracy of the inversion mapping was measured by calculating the root-mean-square (RMS) error between the measured articulatory movements and the estimated ones as follows:

$$RMSE(d) = \sqrt{\frac{\sum_{t=1}^{T}(a_t^{(o)}(d) - a_t^{(e)}(d))^2}{T}}, \tag{4.1}$$

where $RMSE(d)$ is the RMS error of $d$-th dimension of the articulatory parameters. $a_t^{(o)}(d)$ and $a_t^{(e)}(d)$ respectively denote the measured and the estimated $d$-th dimension articulatory parameter at frame $t$. Moreover, we calculated also the correlation coefficient of articulatory parameters between the measured and the estimated ones as follows:

$$r(d) = \frac{\sum_{t=1}^{T}(a_t^{(o)}(d) - \hat{a}^{(o)}(d))(a_t^{(e)}(d) - \hat{a}^{(e)}(d))}{\sqrt{\sum_{t=1}^{T}(a_t^{(o)}(d) - \hat{a}^{(o)}(d))^2}\sqrt{\sum_{t=1}^{T}(a_t^{(e)}(d) - \hat{a}^{(e)}(d))^2}}, \tag{4.2}$$

Table 4.2. Average correlation coefficient of estimated articulatory parameters for male and female speakers with varying number of mixture components from 1 to 128.

| # mix. comp. / speaker | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 |
|---|---|---|---|---|---|---|---|---|
| male | 0.59 | 0.65 | 0.69 | 0.72 | 0.76 | 0.78 | **0.79** | **0.79** |
| female | 0.62 | 0.67 | 0.69 | 0.73 | 0.76 | 0.78 | 0.79 | **0.80** |

where $r(d)$ is the correlation coefficient of $d$-th dimension of the articulatory parameters. $a_t^{(o)}(d)$ and $a_t^{(e)}(d)$ denote the $d$-th dimension articulatory parameter at frame $t$ for the measured and the estimated ones, respectively, while $\hat{a}^{(o)}(d)$ and $\hat{a}^{(e)}(d)$ denote their mean values. These RMS errors and the correlation coefficients were averaged over all dimensions for each setting of number of mixture components for each speaker.

The average RMS errors of all settings of number of mixture components for the male and female speakers are shown in Table 4.1. The best performance is achieved in both male and female speaker by using 128 mixture components. The lowest RMS value is 1.42 mm for the male speaker and 1.41 mm for the female speaker. Whereas the results of average correlation coefficients calculation for both of the speakers are shown in Table 4.2. The highest correlation coefficient for the male speaker is 0.79, achieved by using 64 and 128 mixture components, while for the female speaker, the same highest number of correlation coefficient is achieved using 128 mixture components. Overall, in this thesis, the performance of the inversion mapping is comparable as in [28].

### 4.3.2 Objective evaluation on production mapping

In the second objective evaluation, the accuracy of the production mapping in producing the spectrum parameters was investigated. The number of mixture components was varied from 1 to 128. The accuracy of the production mapping was measured by using the mel-cepstral distortion between the target mel-

Table 4.3. Average mel-cepstral distortion [dB] of mel-cepstrum parameters using conventional production mapping for male and female speakers with varying number of mixture components from 1 to 128.

| # mix. comp. / speaker | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 |
|---|---|---|---|---|---|---|---|---|
| male | 6.13 | 5.60 | 5.27 | 5.05 | 4.88 | 4.76 | **4.70** | 4.71 |
| female | 6.50 | 6.09 | 5.69 | 5.49 | 5.24 | 5.05 | **4.94** | 4.95 |

cepstrum parameters and the estimated ones as follows:

$$\text{Mel-CD[dB]} = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{24}(c_d^{(o)} - c_d^{(e)})^2}, \tag{4.3}$$

where the $d$th mel-cepstrum coefficient of the target and the estimated one are respectively denoted as $c_d^{(o)}$ and $c_d^{(e)}$. The mel-cepstral distortion values were then averaged over all evaluation data for each setting of speaker and number of mixture components.

The average mel-cepstral distortions of all number of mixture components for the male and female speakers are shown in Table 4.3. The highest accuracy is achieved by using 64 mixture components for both of the speakers having mel-cepstral distortion value as 4.70 dB and 4.94 dB for the male and female speaker, respectively. The performance of the production mapping in this thesis is also comparable to that of in [28].

### 4.3.3 Objective evaluation on sequential inversion and production mapping

In the third objective evaluation, we investigated the performance of the proposed sequential mapping system between acoustic and articulatory parameters, described in the section 3.2. Similar as in evaluating the production mapping, we calculated the mel-cepstral distortions between the extracted mel-cepstrum parameters and the estimated ones. The number of mixture components was varied from 1 to 128. We evaluated the performance of the system by training

Table 4.4. Average mel-cepstral distortion [dB] of mel-cepstrum parameters using proposed sequential mapping for male and female speakers with varying number of mixture components from 1 to 128.

| # mix. comp. / speaker | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 |
|---|---|---|---|---|---|---|---|---|
| male | 5.73 | 5.31 | 5.03 | 4.85 | 4.69 | 4.55 | 4.43 | **4.38** |
| female | 6.13 | 5.79 | 5.49 | 5.29 | 5.04 | 4.86 | 4.69 | **4.65** |

the GMMs for the production mapping using both the measured articulatory parameters and the estimated articulatory parameters, which are generated using the inversion mapping, from the training data.

The average mel-cepstral distortions over all evaluation data for each number setting of mixture components are shown in Table 4.4. The best performance is achieved by using 128 mixture components in both speakers. The distortion values are lower compared to that of the conventional production mapping using the measured articulatory parameters, having mel-cepstral distortion as 4.38 dB for the male speaker and 4.65 dB for the female speaker. This result shows that by performing the inversion and production mapping sequentially, reduction on the mel-cepstral distortion can be achieved. Moreover, by training the GMM of the production mapping with the converted articulatory data, the overall accuracy can be further improved, where the lowest mel-cepstral distortion is 3.99 dB for the male speaker and 4.20 dB for the female speaker using both 64 mixture components, as shown in Table 4.5.

Table 4.5. Average mel-cepstral distortion [dB] of mel-cepstrum parameters using proposed sequential mapping trained with converted EMA data.

| # mix. comp. / speaker | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 |
|---|---|---|---|---|---|---|---|---|
| male | 4.31 | 4.31 | 4.19 | 4.15 | 4.04 | 4.03 | **3.99** | – |
| female | 4.55 | 4.52 | 4.39 | 4.37 | 4.25 | 4.21 | **4.20** | – |

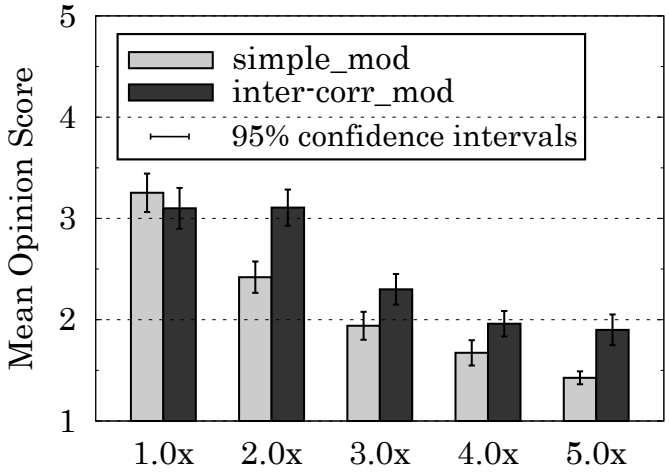



Figure 4.2. Mean Opinion Score (MOS) test result of the quality of modified synthetic speech from both manipulation methods

## 4.4. Comparison of Manipulation Methods for Controlling Articulatory Movements

We performed a subjective evaluation to confirm the performance of the proposed manipulation methods for controlling the articulatory movements, which are described in the section 3.3. In this experiment, we evaluated the quality of the synthetic speech modified by the proposed system. Specifically, the movement of the tongue tip in $y$-coordinate was scaled from 1-fold, i.e. without modification, to 5-fold. A mean opinion score (MOS) test was conducted, with the option of the scores was a 5-point scale, i.e. 5: excellent, 4: good, 3: fair, 2: poor, 1: bad. Ten listeners participated in the evaluation. Each listener evaluated 15 distinct utterances, randomly selected from the evaluation data, where each of the utterance was modified with both of the simple manipulation method and the manipulation method considering inter-correlations of articulatory parameters.

The subjective evaluation result on the male speaker is shown in Fig. 4.2. The result shows that the proposed manipulation method by considering inter-correlations of articulatory parameters capable of preserving higher quality of modified speech sounds. Whereas the quality of the modified speech sounds become significantly degraded if we did not consider those inter-correlations. This

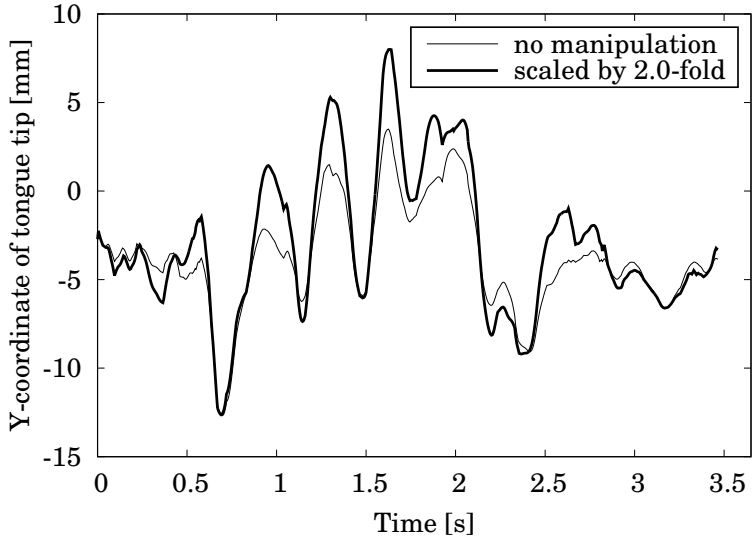



Figure 4.3. Trajectory of tongue-tip in y-coordinate with and without manipulation (2.0-fold scaled).

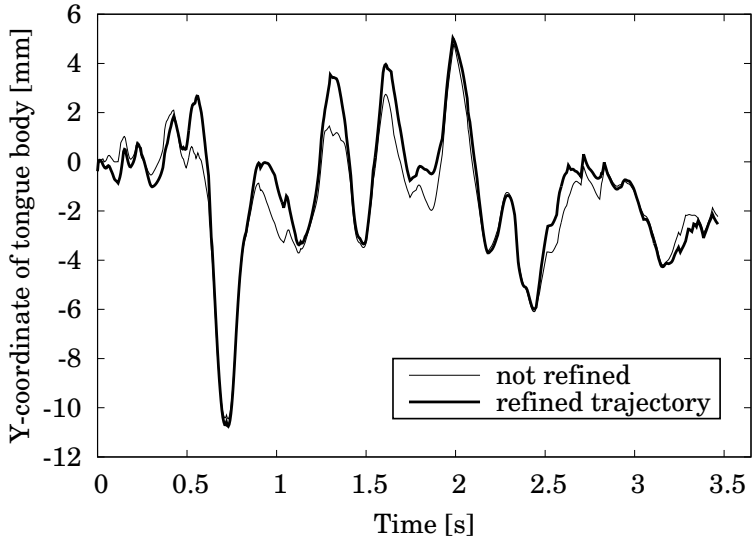



Figure 4.4. Trajectory of tongue-body in y-coordinate after manipulation of tongue tip.

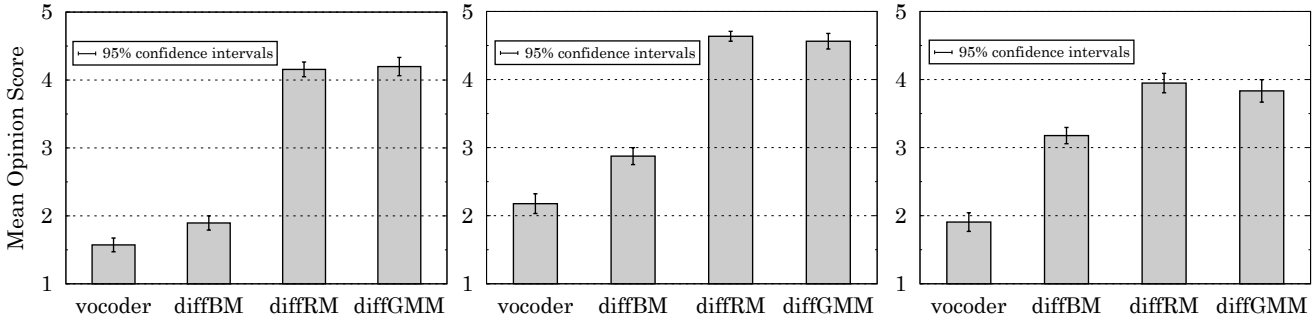



Figure 4.5. Mean Opinion Scores (MOS) on three different degree of articulations for male speaker, hypo-articulation (left), normal articulation (centre), and hyper-articulation (right)

result is consistent with the assumption that the movements of the articulators are correlated to one another. The manipulated trajectory of the tongue tip is shown in Fig. 4.3, where the overall structure is enlarged by a factor of 2. The refined-unmodified trajectory of the tongue body, generated by using the manipulation method by considering inter-correlations, is shown in Fig. 4.4. It can be observed that the refined trajectory of the tongue body follows a shape similar to that of the original one, but with several alterations according to the changes in the trajectory of the tongue tip. The experimental results conclude that the proposed manipulation method of articulatory movements is capable of preserving the quality of modified speech sounds by considering the inter-correlations of articulatory parameters in the manipulation process.

## 4.5. Comparison of Direct Waveform Modification Methods for Generating Modified Speech

In order to asses the proposed direct waveform modification methods, which are described in the section 3.4, we performed another subjective evaluation. In this experiment, we evaluated the quality of the modified speech sounds generated by using the proposed direct waveform modification methods. Specifically, we emulated three speaking conditions by scaling the articulatory movements as: normal-articulation (1-fold scale), hypo-articulated (0.5-fold scale), and hyper-

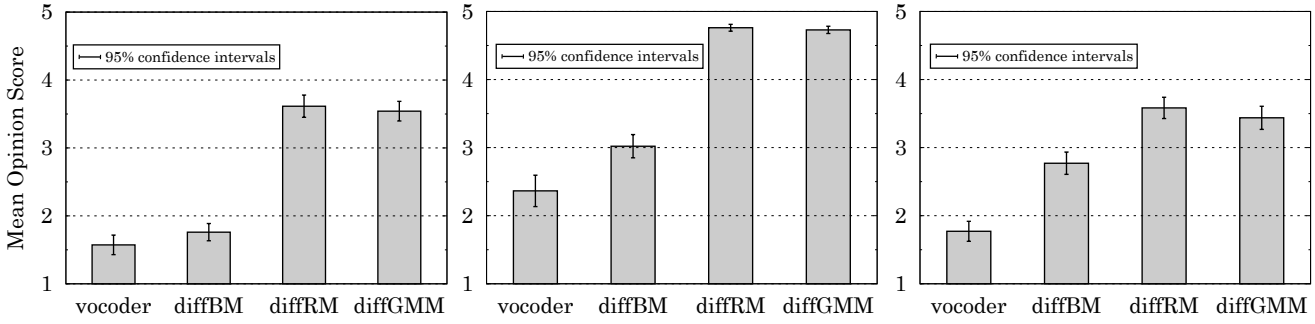



Figure 4.6. Mean Opinion Scores (MOS) on three different degree of articulations for female speaker, hypo-articulation (left), normal articulation (centre), and hyper-articulation (right)

articulated (2-fold scale). Modified speech sounds were generated by using these three settings of scaling manipulation and four different speech generation systems, i.e. using conventional vocoder-based method, basic direct waveform modification method (diffBM), refined direct waveform modification method (diffRM), and refined method with differential GMM (diffGMM). A mean opinion score (MOS) test was conducted using 5-point scale, as in the previous section. Twelve listeners participated in the evaluation. Each listener evaluated 96 speech samples including 8 different utterances for each speech generation system and each speaking condition setting.

The results of this subjective evaluation are shown in Fig. 4.5 and Fig. 4.6 for the male and female speaker, respectively. In both of the speakers, the proposed direct waveform modification methods are capable of improving the quality of the modified speech sounds. Moreover, by considering the refining factors (residuals), which alleviate the oversmoothing effect of the spectrum, described in the section 3.4.2, the diffBM method and the diffGMM method generate significantly higher-quality of modified speech sounds. These traits are observed within all speaking conditions for both the male and the female speakers. An example of comparison of spectrogram from the diffGMM method and the conventional-vocoder based method is shown in Fig. 4.7, where the sentence is "Dolphins are intelligent marine mammals.". From that figure, it can be observed that oversmoothed structures are possessed by the result from the vocoder. On the other hand, the proposed direct waveform modification methods are capable of preserving fine-

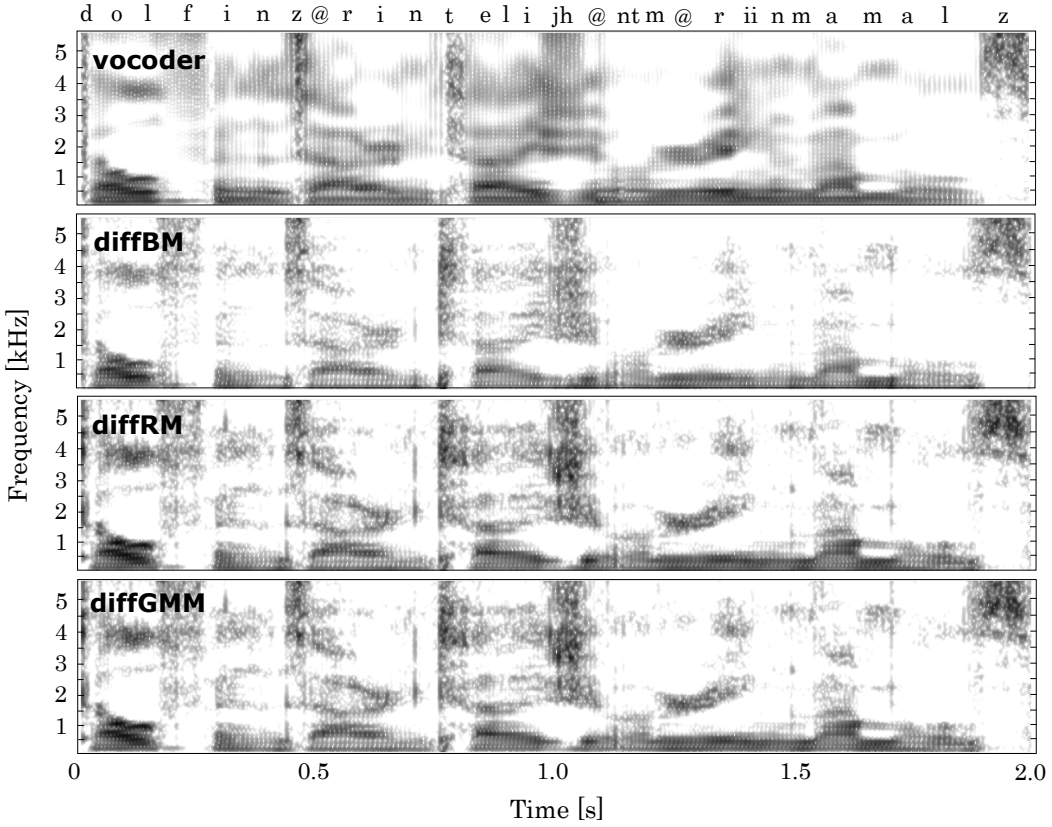



Figure 4.7. Comparison of spectrograms for sentence "Dolphines are intelligent marine mammals." from the male speaker in a hypo-articulated speaking condition (0.5-fold scaling) using vocoder process (top), basic direct waveform modification (second-top), refined direct waveform modification (second-bottom), and refined method with differential GMM (bottom).

structures of the spectrogram, where diffRM and diffGMM methods can generate finer structures compared to diffBM method. These results conclude that the proposed direct waveform modification method significantly improves the quality of the modified speech sounds by avoiding the use of vocoder-based excitation generation process and alleviating the oversmoothed structure of the spectrum with residual factors.

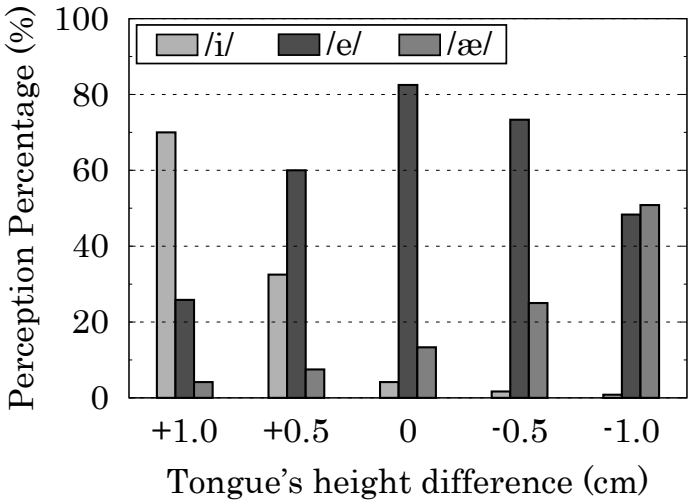



Figure 4.8. Perception percentage of the phoneme modification results for the male speaker

## 4.6. Evaluation on Controllability

In the last experiment, we evaluate the controllability of the proposed system in modifying phonemic sounds by means of controlling the articulators. In this evaluation, we perform modification of the sound of three front vowels in English, i.e. /ɪ/, /æ/, and /ɛ/. To do so, we utilize the dominant role of the tongue in producing these vowel sounds. Specifically, we manipulate the height configuration of the tongue as: in vowel /ɪ/, the height is the highest; in vowel /æ/, the height is the lowest; and in vowel /ɛ/, the height is between the former two. From the evaluation data, we extracted twelve different words containing the vowel /ɛ/. Then, in order to change its sound into the sound of the vowel /ɪ/, we modified the height of the tongue by +0.5 cm and +1.0 cm. On the contrary, in order to produce the sound of the vowel /æ/, we changed the height of the tongue by −0.5 cm and −1.0 cm. The manipulation of the height of the tongue was performed at the centre frame of the vowel /ɛ/ in each word. The altered position of the tongue was then interpolated to the centre frame of both the left phoneme and the right phoneme of the vowel /ɛ/, in the corresponding word, by using the cubic spline interpolation method [54]. A categorical perception evaluation was performed to assess the accuracy of the system in modifying these vowel sounds.

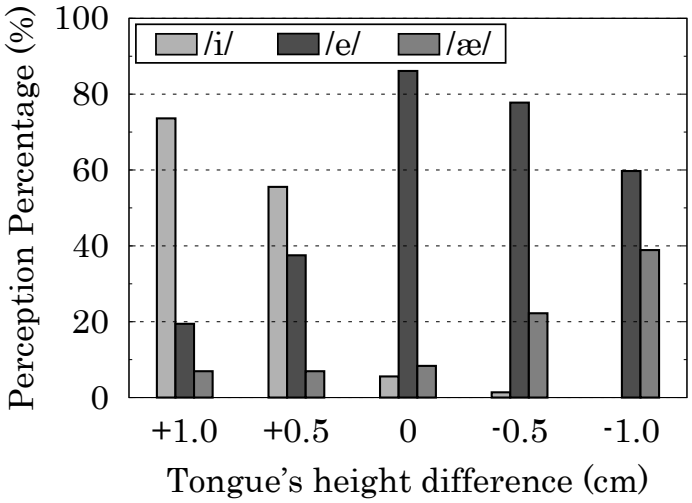



Figure 4.9. Perception percentage of the phoneme modification results for the female speaker

Twelve listeners participated in the evaluation. The modified vowel sound was marked with a question mark in each of the label of the uttered word. Each listener was requested to choose the marked vowel whether it was pronounced as /ɪ/, /ɛ/, or /æ/. Note that, to train the models used in this experiment, frames corresponding to the target vowels, /ɪ/ and /æ/, were removed from the training data. The third proposed direct waveform modification method, i.e. diffGMM, was used.

The evaluation results are shown in Fig. 4.8 for the male speaker and in Fig. 4.9 for the female speaker. Clear transitions from the vowel /ɛ/ to vowel /ɪ/ can be observed, in both male and female speakers, as the height of the tongue becomes higher. On the other hand, although such characteristic is not precisely shown for the vowel /æ/, the tendency of the vowels to be heard as /æ/ can be evidently observed as the height of the tongue becomes lower. Comparison of the spectrograms of modified word "stems" are shown in Fig. 4.10. In these spectrograms, the differences of formant characteristics can be clearly observed, where: the vowel /ɪ/ has the lowest formant $F_1$ and the highest formant $F_2$; the vowel /æ/ has the highest formant $F_1$ and the lowest formant $F_2$; and the vowel /ɛ/ posesses the middle values between those two former vowels. These results

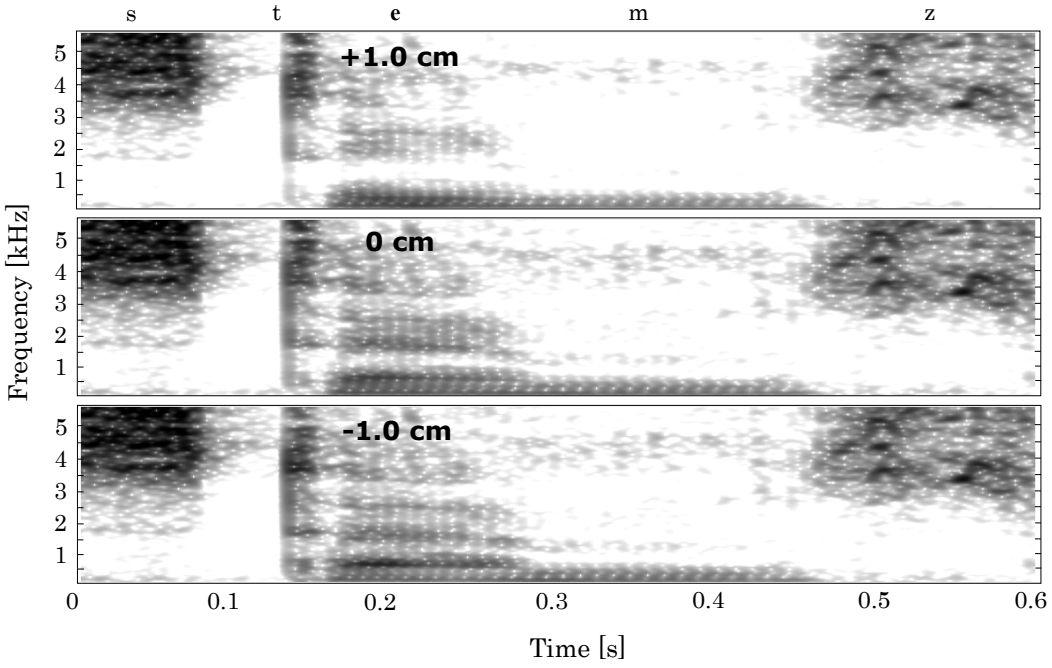



Figure 4.10. Comparison of spectrograms for word "stems" from male speaker, where the height of the tongue is lifted up 1.0cm (top) from the original position (middle) and also shifted down 1.0cm (bottom), showing the formant characteristics difference for vowel /ɪ/ (top), /ɛ/ (middle), and /æ/ (bottom).

indicate that the controllability of the proposed system in manipulating the articulators is very sufficient as shown by its capability of allowing modification of phonemic sounds through the manipulation of particular articulatory configurations. Furthermore, it also suggests that the proposed system could well produce intelligible sounds of foreign pronounciation because frames corresponding to the target vowels were removed in the training process.

## 4.7. Summary

In this chapter, we conducted several experiments to evaluate the proposed articulatory controllable speech modification system. In section 4.3, we demonstrated that the proposed sequential inversion-production mapping system yielded better accuracy compared to the conventional system. In section 4.4, we found that

the proposed method for controlling articulatory parameters by considering their inter-correlations was capable of producing higher quality of modified speech. Then, in section 4.5, we confirmed the effectiveness of the proposed direct waveform modification methods, which significantly improved the quality of the modified speech by avoiding the use of vocoder-based excitation generation process. Finally, in section 4.6, we approved the controllability of the proposed system in controlling the articulators, which was shown by its capability of changing the sound of several chosen vowels.

# Chapter 5

# Conclusion

## 5.1. Thesis Summary

Articulators, such as tongue or lips, have an essential role in the speech production process, i.e. to determine the resonance characteristics of the vocal tract. Therefore, speech can be parameterized not only by the spectrum of the vocal tract, but also by the smoothly time-varying parameters, such as articulatory parameters. Hence, exploiting the relationship between speech and articulators, it would be a beneficial contribution to the development of speech technologies if one could devise a system capable of utilizing the articulatory movements to produce altered speech sounds. This sort of system would become a great benefit, especially in the area of speech-assistive and language technologies. In this thesis, in order to lay a groundwork for such kind of applications, we proposed an articulatory controllable speech modification system using statistical feature mappings, specifically with Gaussian mixture model (GMM). The proposed system makes it possible for one to modify an input speech signal by performing manipulation of its unobserved articulatory movements.

In chapter 2, we reviewed the statistical feature mappings between acoustic and articulatory parameters using the GMM. First, we described the GMM-based acoustic-to-articulatory inversion mapping, along with its training and conversion process. Then, we explained the GMM-based articulatory-to-acoustic production mapping, along with also the training and conversion process. We showed that the GMM-based statistical feature mapping possesses not only sophisticated but also

convenient procedure. Thus, one would easily be able to accomplish in adjusting the parameters of mapping procedure for more diverge tasks.

In chapter 3, we proposed an articulatory controllable speech modificiation system by using the GMM-based statistical feature mappings. In the proposed system, we integrated the GMM-based inversion and production mappings in a sequential fashion. Thus, it allowed one to easily modify the acoustic spectrum of the input speech by manipulating the corresponding articulatory parameters. In order to perform the articulatory manipulation properly, we proposed also a manipulation method of articulatory parameters by considering their inter-correlations. Furthermore, to achieve generation of high-quality modified speech sounds, we proposed speech generation procedure by avoiding the use of vocoder-based excitation generation process through direct waveform modification methods using the spectrum differentials.

In chapter 4, we conducted several experimental evaluations to assess the performance and effectiveness of the proposed system. The experimental results demonstrated that: 1) the proposed sequential mapping system between acoustic and articulatory parameters generates higher accuracy in the estimation of the acoustic spectrum compared to the conventional production procedure using the measured articulatory parameters, 2) the proposed method for manipulation of articulatory movements by considering their inter-correlations is capable of bearing more natural quality of modified speech, 3) the proposed direct waveform modification methods using spectrum differential significantly improve the quality of the modified speech because of bypassing the use of vocoder-based excitation generation process, and 4) the controllability of the proposed system has been confirmed by its capability of modifying the sound of several chosen vowels through handling the configuration of particular articulators, such as the height of the tongue.

## 5.2. Future Work

Despite the success of developing an articulatory controllable speech modification system, some works need to be done for further development.

**Independency of speaker's characteristics and articulatory data:** In

order to develop a generally applicable system, it is important to consider an independency of both speaker's characteristics and availability limitation of articulatory data. First, to address the speaker-independency issue, one way to solve it is by employing the eigenvoice conversion technique which has been used in the voice conversion framework [55, 56]. Through this technique, one can easily adapt arbitrary source speaker into the trained model by using only a very few amount of adaptation utterances. Then, in order to address the limitation of articulatory data, one can deploy the speaker-normalization technique in [57], which unifies the GMM-based acoustic-to-articulatory inversion mapping and voice conversion, making it possible to train inversion mapping function without the need of articulatory data of the source speaker. Therefore, by integrating these two techniques, the GMM-based eigenvoice conversion and speaker-normalization, a speaker-independent articulatory controllable speech modification system that does not depend on also source articulatory data can be developed.

**Implementation of real-time computation and real-life applications:** To attain the essential goal in developing a system utilizing relationship between speech and articulators, of course, creation of real-life applications that can be deployed in daily activities need to be executed. In order to achieve that, first, a real-time computation process must be considered. One way to do that is by adapting the low-delay voice conversion framework which utilizes the time-recursive algorithm in achieving a high-quality real-time computation process [58]. Furthermore, a consideration of lower-quality of input speech parameter or diagonalization of covariance matrices can further reduces the computation cost of the real-time conversion procedure [59]. Finally, to embed the proposed system into a real-life application, one can reflect to the use of visualization-aid of articulatory movements for speech therapy and language learning programs in [15]. Incorporating the speaker-normalization procedure in [57], an articulatory recovery system can be achieved by assuming that healthy (in case of speech therapy) or native (in case of language) speaker would have the correct articulation movements. Then, integrating it with the eigenvoice conversion technique [56], one can develop a voice reconstruction system that utilize the voice and articulatory movements of healthy or native speaker as the reference one. Moreover, by incorporation of source excitation modification, i.e. by controlling the vocal folds

organ, and also duration control, one may develope a comprehensive and augmentative system capable of providing desirable change in input voice with full control of the speech organs that can be very effective for the use in education, medical, research and daily life area.

**Development of articulatory database:** We also plan to develop our own articulatory database. This database would be consisted of several non-native English speakers, e.g. Japanese. The articulatory movements would be recorded by the electromagnetic articulograph (EMA) device, where the speech sounds would be recorded simultaneously. Several pre-processing steps are needed to be performed, such as data normalization and noise removal. This database will be used for further developments of speech-articulatory mapping technique and its application in real life.

# Acknowledgements

I would like to convey my inmost gratitude to Professor Satoshi Nakamura of Nara Institute of Science and Technology for his treasured guidance and encouragement. He has given me a chance for studying in his Lab., Augmented Human Communication Laboratory, and it has widened my mind as well as provided me with a lot of opportunities.

I would also like to express my appreciation to Professor Kenji Sugimoto of Nara Institute of Science and Technology for his invaluable comments to this thesis.

I would like to declare my sincerest thanks in particular to Professor Tomoki Toda of Nagoya University for his precious teaching and support. One thing that keeps me going is because he always believes in me throughout my research study.

I would also like to express my thankfulness to Assistant Professor Sakriani Sakti for her encouragement and feedback during my study.

I would like to give my sincere gratitude to Assistant Professor Graham Neubig and Assistant Professor Koichiro Yoshino of Nara Institute of Science and Technology for their helpful advice and support in my research study.

I would also like to thank Associate Professor Hirokazu Kameoka and members of Communication Science Laboratory of NTT Atsugi, Kanagawa, for their invaluable comments, knowledge and experience.

I would like to convey my deepest appreciation to Ms. Manami Matsuda, secretary of Augmented Human Communication Laboratory of Nara Institute of Science and Technology, for her kindness and continuous support during my stay in Japan. I would also like to thank members of Augmented Human Communication Laboratory of Nara Institute of Science and Technology for their encouragement.

I would also like to assert my acknowledgement to the Ministry of Education, Culture, Sports, Science and Tecnology (MEXT) for the scholarship and tuition support during my study in Japan.

Finally, I would like to give my loving gratefulness to my family for their blessings and prayers. In the end, to the Lord Jesus Christ, my Savior and Provider, I am devoting this work for the greater glory of God.

# References


[1] S. Parthasarathy, J. Schroeter, C. Coker, and M. M. Sondhi, "Articulatory analysis and synthesis of speech," in *TENCON '89. Fourth IEEE Region 10 Int. Conf.*, Bombay, India, Nov. 1989, pp. 760—764.

[2] 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. 73–778.

[3] 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.

[4] A. A. Wrench and K. Richmond, "Continuous speech recognition using articulatory data," in *Proc. ICSLP*, Beijing, China, Oct. 2000, pp. 145–148.

[5] J. Frankel, K. Richmond, S. King, and P. Taylor, "An automatic speech recognition system using neural networks and linear dynamic models to recover and model articulatory traces," in *Proc. ICSLP*, Beijing, China, Oct. 2000, pp. 254–257.

[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] 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.

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

[9] D. W. Massaro, "A computer-animated tutor for spoken and written language learning," in *Proc. of the 5th International Conference on Multimodal Interfaces*, British Columbia, Canada, 2003, pp. 172—-175.

[10] ——, "The psychology and technology of talking heads: Applications in language learning," in *Advances in Natural Multimodal Dialogue Systems*. Heidelberg: Springer, 2005, vol. 30, pp. 183—-214.

[11] D. W. Massaro, Y. Liu, T. H. Chen, and C. Perfetti, "A multilingual embodied conversational agent for tutoring speech and language learning," in *Proc. INTERSPEECH*, Pittsburgh, USA, 2006, pp. 825—-828.

[12] P. Badin, G. Bailly, and L. Boë, "Towards the use of a virtual talking head and of speech mapping tools for pronunciation training," in *Proc. of the ESCA Tutorial and Research Workshop on Speech Technology in Language Learning (STiLL 1998)*, 1998, pp. 167—-170.

[13] O. Jokisch, U. Koloska, D. Hirschfeld, and R. Hoffmann, "Pronunciation learning and foreign accent reduction by an audiovisual feedback system," in *ACII 2005*. Heidelberg: Springer, 2005, vol. 3784, pp. 419—-425.

[14] O. Engwall and O. Bälter, "Pronunciation feedback from real and virtual language teachers," *Journal of Computer Assisted Language Learning*, vol. 20, pp. 235—-262, 2007.

[15] B. J. Kröger, P. Birkholz, R. Hoffman, 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.

[16] J. Schroeter and M. Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 133—-150, 1994.

[17] 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," *The Journal of the Acoust. Soc. of America*, vol. 100, no. 3, pp. 1819–1834, 1996.

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

[19] K. Richmond, S. King, and P. Taylor, "Modelling the uncertainty in recovering articulation from acoustics," *Computer Speech and Language*, vol. 17, no. 2, pp. 153—172, 2003.

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

[21] ——, "Speaker adaptation method for acoustic-to-articulatory inversion using an HMM-based speech production model," *IEICE Trans. on Inf. and Sys.*, vol. 87, no. 5, pp. 1071–1078, 2004.

[22] T. Toda, A. W. Black, and K. Tokuda, "Acoustic-to-articulatory inversion mapping with Gaussian mixture model," in *Proc. INTERSPEECH*, Jeju, Korea, 2004, pp. 1129—1 132.

[23] 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.

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

[25] Y. Shiga and S. King, "Accurate spectral envelope estimation for articulation-to-speech synthesis," in *Proc. 5th ISCA Speech Synthesis Workshop*, Pittsburgh, USA, 2004, pp. 19–24.

[26] 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, 2006, pp. 93–96.

[27] 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 Speech Synthesis Workshop*, Pittsburgh, USA, 2004, pp. 3067–3071.

[28] ——, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215—-227, 2008.

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

[30] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Trans. on Audio, Speech and Lang. Process*, vol. 15, no. 8, pp. 2222—2 235, 2007.

[31] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A hybrid approach to electrolaryngeal speech enhancement based on spectral subtraction and statistical voice conversion," in *Proc. INTERSPEECH*, Lyon, France, Aug. 2013, pp. 3067–3071.

[32] K. Kobayashi, T. Toda, G. Neubig, S. Sakriani, 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.

[33] Y. Tajiri, K. Tanaka, T. Toda, G. Neubig, S. Sakriani, and S. Nakamura, "Non-audible murmur enhancement based on statistical conversion using air- and body-conductive microphones in noisy environments," in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 2769–2773.

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

[35] G. Bouchard and B. Triggs, "The tradeoff between generative and discriminative classifiers," in *16th IASC International Symposium on Computational Statistics (COMPSTAT'04)*, Prague, Czech Republic, Aug. 2004, pp. 721–728.

[36] P. Liang and K. I. Jordan, "An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators," in *Proc. of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008, pp. 584–591.

[37] M. Paliwal and U. A. Kumar, "Neural networks and statistical techniques: A review of applications," *Expert Systems with Applications*, vol. 36, no. 1, pp. 2–17, 2009.

[38] I. Jolliffe, *Principal Component Analysis*. John Wiley & Sons, Ltd, 2002.

[39] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Soc. Series B*, vol. 39, no. 1, pp. 1–38, 1977.

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

[41] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi, and S. Imai, "An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features," in *Proc. EUROSPEECH*, Madrid, Spain, Sep 1995, pp. 757—-760.

[42] T. Toda and S. Young, "Trajectory training considering global variance for HMM-based speech synthesis," in *Proc. ICASSP*, Taipei, Taiwan, Aug 2009, pp. 4025–4028.

[43] S. Takamichi, T. Toda, A. W. Black, S. Sakriani, and S. Nakamura, "Parameter generation algorithm considering modulation spectrum for HMM-based speech synthesis," in *Proc. ICASSP*, Brisbane, Australia, April 2015, pp. 4210–4214.

[44] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura, "Modulation spectrum-constrained trajectory training algorithm for GMM-based voice conversion," in *Proc. ICASSP*, Brisbane, Australia, April 2015, pp. 4859–4863.

[45] A. Ben Youssef, P. 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.

[46] K. Kobayashi, T. Toda, G. Neubig, S. Sakriani, 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.

[47] S. Takamichi, T. Toda, G. Neubig, S. Sakriani, and S. Nakamura, "A post-filter to modify the modulation spectrum in HMM-based speech synthesis," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 290–294.

[48] A. Wrench, "The MOCHA-TIMIT articulatory database," http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html, 1999, Queen Margaret University College.

[49] K. Richmond, "Estimating articulatory parameters from the acoustic speech signal," Ph.D. dissertation, The Centre for Speech Technology Research, University of Edinburgh, 2001.

[50] 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 Communication*, vol. 27, no. 3–4, pp. 187–207, 1999.

[51] 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.

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

[53] T. Toda, A. W. Black, and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," in *Proc. ICASSP*, Philadelphia, USA, Mar. 2005, pp. 9–12.

[54] C. De Boor, *A Practical Guide to Splines*. New York: Springer-Verlag, 1978, vol. 27.

[55] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on Gaussian mixture model," in *Proc. INTERSPEECH*, Pittsburgh, PA, USA, Sep. 2006, pp. 2446–2449.

[56] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Many-to-many eigenvoice conversion with reference voice," in *Proc. INTERSPEECH*, Brighton, United Kingdom, Sep. 2009, pp. 1623–1626.

[57] H. Uchida, D. Saito, N. Minematsu, and K. Hirose, "Statistical acoustic-to-articulatory mapping unified with speaker normalization based on voice conversion," in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 588–592.

[58] T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory," in *Proc. INTERSPEECH*, Brisbane, Australia, Sep. 2008, pp. 1076–1079.

[59] T. Toda, T. Muramatsu, and H. Banno, "Implementation of computationally efficient real-time voice conversion," in *Proc. INTERSPEECH*, Portland, OR, USA, Sep. 2012, pp. 94–97.

# List of Publications

### International Conferences

1. **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 INTERSPEECH*, pp. 2298-2302, Singapore, Sep. 2014.

2. **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 INTERSPEECH*, pp. 3350-3354, Dresden, Germany, Sep. 2015.

3. **P. L. Tobing**, T. Toda, H. Kameoka, and S. Nakamura. Acoustic-to-articulatory inversion mapping based on latent trajectory Gaussian mixture model. *In INTERSPEECH*, Sep. 2016. (Acceptance)

### Technical Reports

1. **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. *IEICE Tech. Rep.*, Vol. 114, No. 365, SP2014-111, pp. 57-62, Dec. 2014

2. **P. L. Tobing**, T. Toda, H. Kameoka, and S. Nakamura. An evaluation of acoustic-to-articulatory inversion mapping with latent trajectory Gaussian mixture model. *IEICE Tech. Rep.*, Vol. 115, No. 523, SP2015-113, pp. 111-116, Mar. 2016

### Meetings

1. **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. *Spring Meeting Acoust. Soc. of Japan*, 2-2-8, pp. 267-268, 2015.

2. **P. L. Tobing**, T. Toda, G. Neubig, S. Sakti, and S. Nakamura. An evaluation of articulatory controllable speech modification based on Gaussian mixture models with direct waveform modification. *Autumn Meeting Acoust. Soc. of Japan*, 2-1-3, pp. 221-222, 2015.

3. **P. L. Tobing**, T. Toda, H. Kameoka, and S. Nakamura. An investigation of acoustic-to-articulatory inversion mapping with latent trajectory Gaussian mixture model. *Spring Meeting Acoust. Soc. of Japan*, 1-2-8, pp. 227-228, 2016.