# Articulatory Controllable Speech Modification Based on Statistical Feature Mapping with Gaussian Mixture Models

*Patrick Lumban Tobing*[1,2], *Tomoki Toda*[1], *Graham Neubig*[1], *Sakriani Sakti*[1], *Satoshi Nakamura*[1], *Ayu Purwarianti*[2]

[1]Graduate School of Information Science, Nara Institute of Science and Technology, Japan
[2]STEI, Institut Teknologi Bandung, Indonesia

13510013@std.stei.itb.ac.id, tomoki@is.naist.jp, neubig@is.naist.jp, ssakti@is.naist.jp, s-nakamura@is.naist.jp, ayu@stei.itb.ac.id

## Abstract

This paper presents a novel speech modification method capable of controlling unobservable articulatory parameters based on a statistical feature mapping technique with Gaussian Mixture Models (GMMs). In previous work [1], the GMM-based statistical feature mapping was successfully applied to acoustic-to-articulatory inversion mapping and articulatory-to-acoustic production mapping separately. In this paper, these two mapping frameworks are integrated to a unified framework to develop a novel speech modification system. The proposed system sequentially performs the inversion and the production mapping, making it possible to modify phonemic sounds of an input speech signal by intuitively manipulating articulatory parameters estimated from the input speech signal. We also propose a manipulation method to automatically compensate for unmodified articulatory movements considering inter-dimensional correlation of the articulatory parameters. The proposed system is implemented for a single English speaker and its effectiveness is evaluated experimentally. The experimental results demonstrate that the proposed system is capable of modifying phonemic sounds by manipulating the estimated articulatory movements and higher speech quality is achieved by considering the inter-dimensional correlation in the manipulation.

**Index Terms**: speech modification, acoustic-to-articulatory inversion mapping, articulatory-to-acoustic production mapping, Gaussian mixture model, inter-dimensional correlation

## 1. Introduction

Articulators are a set of human speech organs that are used in a unified way to control the resonance characteristics of the vocal tract. Therefore, speech can be characterized by articulatory parameters, such as movements of the articulators. Because the articulatory parameters vary much more slowly than the acoustic parameters of speech [2], they have a potential to yield better parameterization of speech in many applications such as speech coding [3], speech recognition [4], and speech synthesis [5]. Furthermore, speech is more easily modified in an understandable way by manipulating articulatory parameters rather than acoustic parameters [1, 6].

There have been many attempts at developing mapping systems between the speech acoustic parameters and the articulatory parameters [1, 3, 6, 7, 8, 9, 10, 11, 12, 13]. There are mainly two mapping systems: one is an acoustic-to-articulatory inversion mapping system to estimate the articulatory parameters from the given acoustic parameters and the other is an articulatory-to-acoustic production mapping system to estimate the acoustic parameters from the given articulatory parameters [1]. One of the typical approaches to these mapping systems is based on mathematical production models [3, 7]. However, the speech production mechanism is too complex to be mathematically modeled without some approximations.

Recently, some research has examined statistical approaches that do not mathematically model the speech production mechanism. These mapping systems between articulatory parameters and speech acoustics are developed in a data-driven manner using parallel acoustic-articulatory data.There have been proposed several statistical methods, e.g., the mapping system using codebooks [8, 9], hidden Markov models (HMMs) [10, 11], neural networks [12, 13], and Gaussian mixture models (GMMs) [1], and their effectiveness has been confirmed in both the inversion and production mapping. Moreover, it has been reported that phoneme sounds of synthetic speech are effectively modified by manipulating the articulatory parameters in articulatory controllable HMM-based text-to-speech synthesis, where the articulatory parameters are modeled as intermediate features [6].

Inspired by the conventional work [1, 6], we propose a novel articulatory controllable speech modification system. Specifically we do so by developing a new speech analysis/synthesis framework by combining the inversion and production mapping to make it possible to modify speech signals by manipulating the unobserved articulatory parameters. Such a framework has a great potential to develop various new speech applications, such as speech recovery for vocally disabled people, pronunciation enhancement in speaking foreign languages, and concealing messages by modifying phonemes/words.

In this paper, we focus on the GMM-based inversion/production mapping methods [1] as one of the promising methods capable of easily being applied to any language. Thanks to its independence of text/language specification input, which is needed by [6], as only speech signals are needed as the input of this system. In the proposed system, the articulatory parameters are first estimated from a given input speech signal using a GMM-based inversion mapping system. These articulatory parameters are manipulated, and then the acoustic parameters are estimated from the manipulated articulatory parameters using a GMM-based production mapping system. Finally, a modified speech signal is generated from the estimated acoustic parameters. We also propose an articulatory manipulation method for refining unmodified parts of the articulatory parameters according to the modified parts by considering their inter-dimensional correlation.

## 2. GMM-based Inversion and Production Mapping [1]

A simultaneously recorded speech and articulatory data set is used as training data to construct the GMMs for the inversion and production mapping. In this paper, we use speaker-dependent GMMs. For articulatory parameters, we use 14-

dimensional Electromagnetic articulograph (EMA) data, which are provided in MOCHA [14]. Locations of seven articulators (top lip, bottom lip, bottom incisor, tongue tip, tongue body, tongue dorsum, and velum) are measured in x- and y- coordinates on the midsagittal plane.

Let $\boldsymbol{c}_t$, $\boldsymbol{s}_t$, and $\boldsymbol{x}_t$ be spectral envelope parameters (*i.e.*, mel-cepstrum in this paper), source excitation parameters (*i.e.*, log-scaled $F_0$ and log-scaled waveform power in this paper), and the articulatory parameters. Time sequence vectors of these parameters over an utterance are $\boldsymbol{c} = \left[\boldsymbol{c}_1^\top, \cdots, \boldsymbol{c}_T^\top\right]^\top$, $\boldsymbol{s} = \left[\boldsymbol{s}_1^\top, \cdots, \boldsymbol{s}_T^\top\right]^\top$, and $\boldsymbol{x} = \left[\boldsymbol{x}_1^\top, \cdots, \boldsymbol{x}_T^\top\right]^\top$, respectively, where $T$ is the number of frames and $\top$ denotes the transposition of the vector.

### 2.1. Acoustic-to-articulatory inversion mapping

In the inversion mapping, spectral envelope parameters of an input speech signal (the source features) are converted to the corresponding articulatory parameters (the target features).

#### *2.1.1. Source and target features in inversion mapping*

The source features consist of a mel-cepstral segment feature vector extracted from mel-cepstrum parameters at multiple frames around the current frame. The mel-cepstral segment feature vector at frame $t$ is denoted as $\boldsymbol{O}_t$, which is given by

$$\boldsymbol{O}_t = \boldsymbol{A}\left[\boldsymbol{c}_{t-L}^\top, \cdots, \boldsymbol{c}_t^\top, \cdots, \boldsymbol{c}_{t+L}^\top\right]^\top + \boldsymbol{b}, \quad (1)$$

where linear transformation parameters $\boldsymbol{A}$ and $\boldsymbol{b}$ are determined with principal component analysis for the training data in advance. On the other hand, a joint static and dynamic feature vector of the articulatory parameters is used as the target feature, which is given by $\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top\right]^\top$, where $\Delta\boldsymbol{x}_t$ is the dynamic feature vector of the articulatory parameters at frame $t$.

#### *2.1.2. Training process in inversion mapping*

A joint source and target feature vector $[\boldsymbol{O}_t^\top, \boldsymbol{X}_t^\top]^\top$ is constructed at each frame in the training data. Then, the joint probability density function of the source and target features is modeled with the GMM for the inversion mapping as follows:

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

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

#### *2.1.3. Conversion process*

Given a time sequence of the mel-cepstral segment feature vectors $\boldsymbol{O}$, a time sequence of the articulatory parameters $\boldsymbol{x}$ is determined by maximizing the conditional probability density function $P\left(\boldsymbol{X}|\boldsymbol{O}, \boldsymbol{\lambda}^{(O,X)}\right)$, which is analytically derived from the GMM for the inversion mapping. In this paper, an approximation of the conditional probability density function using a single mixture component sequence $\boldsymbol{m} = \{m_1, \cdots, m_T\}$ [15] is employed, where $m_t$ shows the mixture component index at frame $t$. First, the suboptimum mixture component sequence $\hat{\boldsymbol{m}}^{(O)}$ is determined as follows:

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

Then, the converted articulatory parameter sequence vector $\hat{\boldsymbol{x}}$ is determined as follows:

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} P\left(\boldsymbol{X}|\boldsymbol{O}, \hat{\boldsymbol{m}}^{(O)}, \boldsymbol{\lambda}^{(O,X)}\right), \quad (4)$$

$$\text{subject to } \boldsymbol{X} = \boldsymbol{W}^{(x)}\boldsymbol{x}, \quad (5)$$

where $\boldsymbol{W}^{(x)}$ is a linear transform to expand the articulatory parameter sequence vector $\boldsymbol{x}$ into its joint static and dynamic feature sequence vector $\boldsymbol{X}$.

### 2.2. Articulatory-to-acoustic production mapping

In the production mapping, the spectral envelope parameters are determined from both the articulatory parameters and the excitation parameters.

#### *2.2.1. Source and target features in production mapping*

As the source features, a joint static and dynamic feature vector including not only the articulatory parameters but also the source excitation parameters is used, which is given by $\boldsymbol{Y}_t = \left[\boldsymbol{x}_t^\top, \boldsymbol{s}_t^\top, \Delta\boldsymbol{x}_t^\top, \Delta\boldsymbol{s}_t^\top\right]^\top$ at frame $t$. On the other hand, as the target features, a joint static and dynamic feature vector of the mel-cepstrum $\boldsymbol{C}_t = \left[\boldsymbol{c}_t^\top, \Delta\boldsymbol{c}_t^\top\right]^\top$ is used at frame $t$.

#### *2.2.2. Training process*

The training process is basically the same as described in **Section 2.1.2**. After constructing the joint source and target feature vectors in the training data, the joint probability density function of the source and target features is modeled with the GMM for the production mapping as follows:

$$P\left(\boldsymbol{Y}_t, \boldsymbol{C}_t|\boldsymbol{\lambda}^{(Y,C)}\right) = \sum_{m=1}^{M} \alpha_m^{(Y,C)} \mathcal{N}\left(\left[\boldsymbol{Y}_t^\top, \boldsymbol{C}_t^\top\right]^\top; \boldsymbol{\mu}_m^{(Y,C)}, \boldsymbol{\Sigma}_m^{(Y,C)}\right), \quad (6)$$

where $\boldsymbol{\lambda}^{(Y,C)}$ denotes a parameter set of the GMM for the production mapping, which consists of mixture-component weights $\alpha_m^{(Y,C)}$, mean vectors $\boldsymbol{\mu}_m^{(Y,C)}$, and full covariance matrices $\boldsymbol{\Sigma}_m^{(Y,C)}$ of individual mixture components.

#### *2.2.3. Conversion process*

The conversion process is also basically the same as described in **Section 2.1.3**. Given a time sequence of the source feature vectors $\boldsymbol{Y}$, that of the converted mel-cepstrum parameters $\hat{\boldsymbol{c}}$ is determined as follows:

$$\hat{\boldsymbol{m}}^{(Y)} = \arg\max_{\boldsymbol{m}} P\left(\boldsymbol{m}|\boldsymbol{Y}, \boldsymbol{\lambda}^{(Y,C)}\right), \quad (7)$$

$$\hat{\boldsymbol{c}} = \arg\max_{\boldsymbol{c}} P\left(\boldsymbol{C}|\boldsymbol{Y}, \hat{\boldsymbol{m}}^{(Y)}, \boldsymbol{\lambda}^{(Y,C)}\right) \quad (8)$$

$$\text{subject to } \boldsymbol{C} = \boldsymbol{W}^{(c)}\boldsymbol{c}, \quad (9)$$

where $\boldsymbol{W}^{(c)}$ is a linear transform to expand the static mel-cepstrum sequence vector $\boldsymbol{c}$ into its joint static and dynamic feature sequence vector $\boldsymbol{C}$. Note that the global variance (GV) [15] is also considered in the production mapping to improve the converted speech quality.

# 3. Articulatory Controllable Speech Modification

The proposed articulatory controllable speech modification process is shown in **Figure 1**. First a given input speech signal is analyzed to extract speech acoustic parameters, such as mel-cepstrum parameters $\boldsymbol{c}$ and the source excitation parameters $\boldsymbol{s}$ including waveform power and $F_0$. Then, the inversion mapping is performed to determine the estimated articulatory parameters $\hat{\boldsymbol{x}}$ corresponding to the given input speech signal from the mel-cepstral segment features $\boldsymbol{O}$ as described in **Section 2.1**. Next, the estimated articulatory parameters are modified manually, *e.g.*, scaling movements of some articulators or changing positions of some articulators to modify phoneme sounds. After that, the production mapping is performed to determine the estimated mel-cepstrum parameters $\hat{\boldsymbol{c}}$ corresponding to the modified articulatory parameters $\hat{\boldsymbol{x}}'$ and the extracted source excitation parameters $\boldsymbol{s}$ in the manner described in **Section 2.2**. Finally, the modified speech signal is generated from the estimated mel-cepstrum parameters $\hat{\boldsymbol{c}}$ and the extracted source excitation parameters $\boldsymbol{s}$ using a vocoder.

In the manipulation of articulatory parameters, it is convenient to manually control 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 paper, we implement two manipulation methods to do so.

### 3.1. Simple manipulation method

The articulatory parameters at frame $t$ estimated by the inversion mapping are denoted as the $D$-dimensional vector $\hat{\boldsymbol{x}}_t = [\hat{x}_t(1), \cdots, \hat{x}_t(D)]^\top$. Then, the manipulated articulatory parameters $\hat{\boldsymbol{x}}'_t$ are defined by changing only components corresponding to the movements of target articulators to be manipulated; *e.g.*, if only the first and second dimensional components are changed to $\hat{x}'_t(1)$ and $\hat{x}'_t(2)$, respectively, $\hat{\boldsymbol{x}}'_t$ is given by $[\hat{x}'_t(1), \hat{x}'_t(2), \hat{x}_t(3), \cdots, \hat{x}_t(D)]^\top$.

This method is capable of easily manipulating only the movements of the target articulators. However, because movements of some articulators are strongly correlated to each other [16], *e.g.*, the movements of the tongue tip affects those of the tongue body, this method possibly causes unnatural movements of the articulators.

### 3.2. Manipulation method considering inter-dimensional correlation of articulatory parameters

To consider the inter-dimensional correlation of the articulatory parameters, we propose a second manipulation method based on two stage inversion. After the inversion mapping and the simple manipulation of the articulatory parameters as mentioned above, the modified components of the articulatory parameters are appended to the source features. Then, the second stage inversion mapping is performed to refine the other components of the articulatory parameters using the conditional probability density function derived from the GMM for the inversion mapping.

The modified articulatory parameter vector consisting of only the manually modified components at frame $t$ is given by $\hat{\boldsymbol{x}}_t^{(m)'}$. A time sequence vector of the joint static and dynamic feature vectors is given by $\hat{\boldsymbol{X}}^{(m)'}$. On the other hand, the unmodified articulatory parameter vector consisting of the other components at frame $t$ is given by $\boldsymbol{x}_t^{(u)}$ and a time sequence vector of the joint static, and dynamic features is given by $\boldsymbol{X}^{(u)}$. The sum of the number of dimensions of $\hat{\boldsymbol{x}}_t^{(m)'}$ and that of $\boldsymbol{x}_t^{(u)}$ is equivalent to $D$. In the second stage inversion mapping, the unmodified articulatory parameter sequence vector is determined as follows:

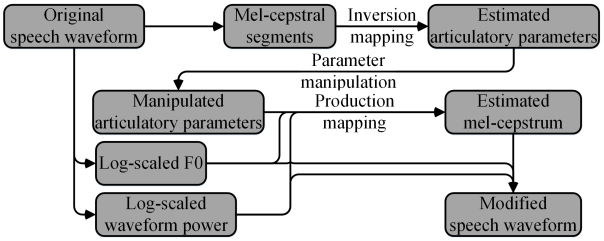


Figure 1: Proposed speech modification process

$$\hat{\boldsymbol{x}}^{(u)} = \arg\max_{\boldsymbol{x}^{(u)}} P\left(\boldsymbol{X}^{(u)} | \boldsymbol{O}, \hat{\boldsymbol{X}}^{(m)'}, \hat{\boldsymbol{m}}^{(O)}, \boldsymbol{\lambda}^{(O,X)}\right), \quad (10)$$

$$\text{subject to } \boldsymbol{X}^{(u)} = \boldsymbol{W}^{(x^{(u)})} \boldsymbol{x}^{(u)}, \quad (11)$$

where $\boldsymbol{W}^{(x^{(u)})}$ is a linear transform to expand the unmodified articulatory parameter sequence vector $\boldsymbol{x}^{(u)}$ into its joint static and dynamic feature vector sequence $\boldsymbol{X}^{(u)}$. The mixture component sequence $\hat{\boldsymbol{m}}^{(O)}$ is given by Eq. (3).

The conditional probability density function used in the inversion mapping effectively models inter-dimensional correlation of the articulatory parameters with the mixture-dependent full covariance matrices. Therefore, the unmodified articulatory parameters are automatically revised in Eq. (10) according to the modified articulatory parameters. Note that the inter-frame correlation of the articulatory parameters is also considered in this revision due to the trajectory-based conversion framework [1] using an explicit relationship between the static and dynamic features shown in Eq. (11). Consequently, it is expected that this manipulation method will yield more natural movements of the articulatory parameters compared to the simple manipulation method.

# 4. Experimental Evaluation

### 4.1. Experimental conditions

As a simultaneously recorded speech and articulatory data set, we used one British male speaker's data in MOCHA [14]. Speech data was sampled at 16 kHz. EMA data was used as the articulatory data.

In speech acoustic parameter extraction, we used the STRAIGHT analysis method [17] 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 envelop parameters. The current $\pm$ 10 frames were used to extract the mel-cepstral segments for the inversion mapping as described in **Section 2.1.1**. 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 [18] in STRAIGHT was employed to extract $F_0$ values. For the articulatory parameters, we used 14-dimensional EMA data as briefly explained in **Section 2**. These 14-dimensional articulatory feature vectors were converted to a Z-score (zero mean and unit variance). Frame shift was set to 5 ms.

We used 350 sentences for training and the remaining 110 sentences for evaluation. The silence frames were removed using phonetic segmentation information included in the MOCHA. We trained two GMMs separately for the inversion mapping and the production mapping, as described in **Sections 2.1** and **2.2**. The number of mixture components was optimized

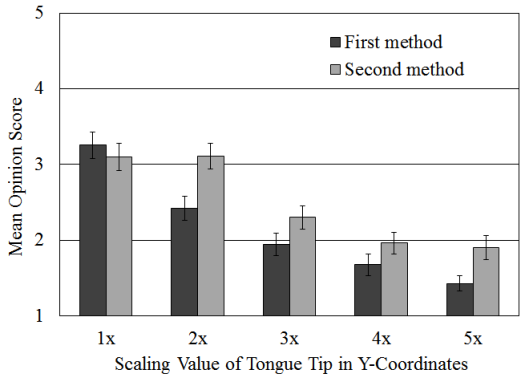


Figure 2: Mean Opinion Score (MOS) test result of the quality of modified synthetic speech from both manipulation methods

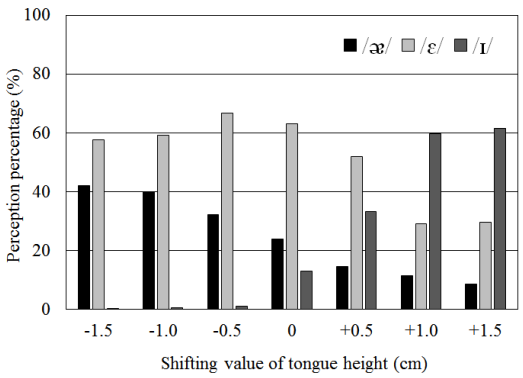


Figure 3: Perception percentage of vowel in modified words resulting from manipulation of tongue's height position

so that estimation accuracy in each mapping was maximized for the test data. The resulting numbers of mixture components were 64 for both the inversion mapping and the production mapping. In the inversion mapping, the correlation coefficient between the estimated and natural articulatory parameters was 0.79. In the production mapping, mel-cepstral distortion between the estimated and natural mel-cepstra was 4.70 dB. In the proposed sequential inversion and production mapping system, we also evaluated the estimation accuracy of mel-cepstrum without performing any modifications of the estimated articulatory parameters. The resulting mel-cepstral distortion between the estimated and natural mel-cepstral was 4.45 dB.

Two subjective evaluations were conducted to measure the performance of the proposed speech modification system. In the first evaluation, we compared two articulatory modification methods described in **Sections 3.1** and **3.2** in terms of speech quality by scaling the movements of an articulator. In the second evaluation, we investigate whether the proposed system are able to modify a specific phoneme by means of manipulating the articulatory positions. For both of the evaluations, there were ten listeners. Two of them were German native speakers, seven of them were Indonesian native speakers, and one was Vietnamese native speakers.

### 4.2. Speech quality evaluation for comparison of articulatory manipulation methods

We evaluated the quality of synthetic speech modified by the proposed system. The tongue tip's movement in y-coordinate was scaled to several values over an utterance from one-fold (*i.e.*, without modification) to 5-fold. An opinion test was conducted, setting the opinion score to a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). Every listener evaluated 15 distinct sentences, randomly selected from the test data, where each of them were modified with two manipulation methods using five scaling settings.

Figure 2 shows the result. Speech quality degrades more as the scaling value grows larger. The manipulation method considering the inter-dimensional correlation effectively reduces this quality degradation. We can see that it enables the system to keep speech quality of the modified speech equivalent to that of the unmodified speech, when the scaling value is set to 2. This result shows that the manipulation method considering the inter-dimensional correlation is more effective than the simple manipulation method.

### 4.3. Evaluation of phoneme sound modification

We investigated the feasibility of intentionally modifying certain phonemes with manual articulatory manipulations. We picked three front vowels in English, /æ/, /ɛ/ and /ɪ/, where the tongue tip's height is the obvious difference between them [6]; *i.e.*, /æ/ has the lowest position, /ɛ/ has the middle position, and ɪ/ has the highest position. We selected 10 distinct words from the test data, where each of them contained vowel /ɛ/. Then, the value of the tongue tip's height at the center frame of the vowel /ɛ/ was shifted from -1.5 cm to +1.5 cm in 0.5 cm intervals, relative to the originally observed position. A negative shifting value means the tongue tip is shifted down, whilst a positive value means it is shifted up. Spline interpolation was performed to generate continuous movements of the tongue tip. The manipulation method considering sthe inter-dimensional correlation was employed.

The result is shown in **Figure 3**. We can observe a clear transition between /ɛ/ and /ɪ/, as the tongue tip's height gets higher. On the other hand, the transition is not so clear from /ɛ/ to /æ/, as the tongue tip's height gets lower. However, we can still observe a reasonable tendency that the perception rate of /æ/ sound increases. Although further improvements will be necessary, the proposed system has a great potential to achieve manual modification of phoneme sounds of input speech by intuitively manipulating unobservable articulatory parameters.

## 5. Conclusions

In this paper, we proposed a novel speech modification method to make it possible to modify speech signals by manipulating unobserved articulatory movements in a sequential flow of statistical inversion and production mapping with Gaussian mixture models (GMMs). We also proposed a new articulatory manipulation method to automatically refine unmodified articulatory movements according to the modified ones by considering the inter-dimensional correlation between individual articulators. The experimental results showed that higher quality of the modified speech is yielded by considering the inter-dimensional correlation in the articulatory manipulation and that the proposed system is also capable of intentionally modifying phoneme sounds by manipulating the articulatory parameters. We plan to further improve quality of the modified speech and controllability of the articulatory parameters.

## 6. Acknowledgements

This research was supported in part by JSPS KAKENHI Grant Number 22680016 and JASSO (Japan Student Services Organization) scholarship.

# 7. References


[1] Toda, T., Black, A. W., and Tokuda, K., "Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model," Speech Communication, Vol. 50, No. 3, pp. 215–227, Mar. 2008.

[2] Parthasarathy, S., Schroeter, J., Coker, C. and Sondhi, M. M., "Articulatory analysis and synthesis of speech," Fourth IEEE region 10 international conference, pp. 760–764, Nov. 1989.

[3] Schroeter, J. and Sondhi, M. M., "Speech coding based on physiological models of speech production," Advances in Speech Signal Processing, S. Furui and M. M. Sondhi, Marcel Dekker New York, pp. 231–267, 1992.

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

[5] Bollepali, B., Black, A., and Prahallad, K., "Modeling a noisy-channel for voice conversion using articulatory features," Proc. INTERSPEECH, Portland, USA, Sep. 2012.

[6] Ling, Z., Richmond, K., Yamagishi, J., and Wang, R., "Articulatory control of HMM-based parametric speech synthesis driven by phonetic knowledge," IEEE Trans. Speech, Audio, and Lang. Process., Vol. 7, No. 6, pp. 697–708, 2008.

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

[8] Hogden, J., Lofqvist, A., Gracco, V., Zlokarnik, I., Rubin, P., and Saltzman, E., "Accurate recovery of articulator positions from acoustics: new conclusions based on human data," J. Acoust. Soc. Am., Vol. 100, pp. 1819–1834, 1996.

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

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

[11] Hiroya, S. and Honda, M., "Speaker adaptation method for acoustic-to-articulatory inversion using an HMM-based speech production model," IEICE Trans. Inf. and Syst., Vol. E87-D, No. 5, pp. 1071–1078, 2004.

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

[13] Richmond, K., King, S., and Taylor, P., "Modelling the uncertainty in recovering articulation from acoustics," Computer Speech and Language, Vol. 17, No. 2, pp. 153–172, 2003.

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

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

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

[17] Kawahara, H., Masuda-Katsuse, I., and de Cheveigne, A., "Restructuring speech representation using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based $F_0$ extraction: possible role of a repetitive structure in sounds", Speech Communication, Vol. 27, No. 3-4, pp. 187-207, 1999.

[18] Kawahara, H., Katayose, H., de Cheveigne, A., and Patterson, R. D., "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of $F_0$ and periodicity", Proc. EUROSPEECH, pp. 2781-2784, Budapest, Hungary, Sep. 1999.