Articulatory Controllable Speech Modification using Sequential Inversion and Production Mapping with Gaussian Mixture Models

Patrick LUMBAN TOBING[†], Tomoki TODA[†], Graham NEUBIG[†], Sakriani SAKTI[†], Satoshi NAKAMURA[†], and Ayu PURWARIANTI[††]

[†] Graduate School of Information Science, Nara Institute of Science and Technology
Takayama-cho 8916-5, Ikoma, Nara, 630–0192 Japan
[††] School of Electrical Engineering and Informatics, Institut Teknologi Bandung
Jl. Ganesha 10, Bandung 40132 Jawa Barat - Indonesia

Abstract In this report, we propose an articulatory controllable speech modification framework using statistical inversion and production mapping with Gaussian Mixture Models. The proposed framework enables us to modify speech waveforms by manipulating unobserved articulatory parameters estimated in the inversion mapping and generating the modified speech waveforms from the manipulated articulatory parameters in the production mapping. We also propose an articulatory manipulation method that considers inter-dimensional correlation between articulators. The experimental results show that the proposed framework is capable of successfully modifying phoneme sounds by manually controlling related articulators.

Key words speech modification, articulatory control, inversion mapping, production mapping, Gaussian Mixture Models

1. Introduction

During human speech production, articulators are moving together in a proper way so that a certain speech sound can be produced with particular resonance characteristics of the vocal tract. Hence, speech sound is able to be characterized not only by its acoustic properties, but also by articulators properties, such as articulatory movements during the speech production. In terms of modifying speech, it is of course a more intuitive way to manipulate the articulatory parameters rather than the acoustic parameters [1], [2]. In addition, with its slower varying trajectories compared to acoustic parameters [3], articulatory parameters are promising to become a better parameterization method in a variety of applications, such as speech recognition [4], speech synthesis [5], and speech coding [6].

In order to utilize the articulatory parameters in a speech modification system, it is inevitable to firstly develop mapping models between speech acoustic and articulatory parameters. Typically, there are two distinct mapping systems that cover this relationship [1]. One is acoustic-to-articulatory mapping process or so called inversion mapping. It is used to estimate articulatory parameters from acoustic parameters input. The other one is articulatory-to-acoustic mapping process or so called production mapping. It is used to estimate acoustic parameters from articulatory parameters input.

Earlier, those acoustic and articulatory mapping systems were approached by mathematical production models [6], [7]. However, speech production process is too complex to be mathematically modeled without some approximations. Recently, with the emergence of parallel acoustic-articulatory data, it is possible to estimate the speech production process with statistical approaches, instead of mathematical modelling. There have been proposed several works in the category of statistical methods, e.g., mapping systems using codebooks [8], [9], hidden Markov models (HMMs) [10], [11], neural networks [12], [13], and Gaussian Mixture Models (GMMs) [1], with the effectivity of both inversion and production mapping processes reported in each respective method. In addition, it is also reported that by manipulating certain articulatory movements, a particular phoneme sound can be effectively modified in an articulatory controllable HMM-based text-to-speech synthesis [2].

In this work, we propose a novel articulatory controllable speech modification system, inspired by the previous work for GMM-based inversion/ production mapping methods [1]. The proposed system is capable of sequentially estimating articulatory parameters from a given input speech signal using the GMM inversion mapping, manipulating the estimated articulatory parameters, estimating acoustic parameters from the manipulated articulatory parameters using the GMM production mapping, and finally synthesizing a modified speech signal from the estimated acoustic parameters. For the manipulation of the estimated articulatory parameters, we also develop a method capable of refining unmodified parts according to the modified parts of the estimated articulatory parameters by considering inter-dimensional correlation. The proposed system has a great potential to develop various speech applications, such as speech recovery for vocally disabled people, pronunciation enhancement in speaking foreign languages, and concealing messages by modifying particular phoneme/sounds. Furthermore, this system is capable of easily applied to any language, as it only needs a speech signal as input without the needs for text/language specification input like in [2].

2. Inversion and Production Mapping with GMM [1]

In this paper, a set of simultaneously recorded acoustic-articulatory data is used in the training process to develop speaker dependent GMMs as both inversion and production mapping models. The data are provided in MOCHA [14], where 14-dimensional electromagnetic articulograph (EMA) data is used as articulatory parameters. They represent movements of seven articulators (lower incisor, upper lip, bottom lip, tongue tip, tongue body, tongue dorsum, and velum) movements in x- and y- coordinates on the midsagittal plane.

Let assume $\boldsymbol{c}_t$, $\boldsymbol{s}_t$, and $\boldsymbol{x}_t$ as spectral envelope parameters (mel-cepstrum), source excitation parameters (log-scaled F0 and log-scaled power), and the articulatory parameters respectively. Time sequence vectors of individual parameters over an utterance are denoted as $\boldsymbol{c} = \left[\boldsymbol{c}_1^\top, \cdots, \boldsymbol{c}_T^\top\right]^\top$, $\boldsymbol{s} = \left[\boldsymbol{s}_1^\top, \cdots, \boldsymbol{s}_T^\top\right]^\top$, 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 Inversion mapping

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

2.1.1 Source and target features

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

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

where $\boldsymbol{A}$ and $\boldsymbol{b}$ are linear transformation parameters determined by performing principal component analysis for the training data. Whereas, the target feature consists of a joint static and dynamic feature of articulatory parameters, 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, which is calculated as $\Delta\boldsymbol{x}_t = \boldsymbol{x}_t - \boldsymbol{x}_{t-1}$.

2.1.2 Training process

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

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

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

2.1.3 Conversion process

In the conversion process, the target articulatory parameter sequence $\boldsymbol{x}$ corresponding to the given input speech is estimated by maximizing the conditional probability density function $P(\boldsymbol{X}|\boldsymbol{O},\boldsymbol{\lambda})$ given the mel-cepstral segment feature sequence $\boldsymbol{O}$. This conditional probability density function given by a GMM is effectively approximated with a single Gaussian distribution using a single mixture component sequence $\boldsymbol{m} = \{m_1, \cdots, m_T\}$. The suboptimum mixture component sequence $\hat{\boldsymbol{m}}$ is determined first by:

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

After that, the estimated articulatory parameter sequence is determined by:

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

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

where $\boldsymbol{W}^{(x)}$ is a transformation matrix for creating the joint static and dynamic feature sequence vector $\boldsymbol{X}$ from the static feature sequence vector $\boldsymbol{x}$.

2.2 Production mapping

In the production mapping, both the articulatory parameters and the excitation parameters are converted into the spectral envelope parameters.

2.2.1 Source and target features

The source feature consists of joint static and dynamic feature vectors of articulatory parameters and excitation parameters, which is given by $\boldsymbol{Y}_t = [\boldsymbol{x}_t, \Delta\boldsymbol{x}_t, \boldsymbol{s}_t, \Delta\boldsymbol{s}_t]$ at frame t. Whereas, the target feature consists of a joint static and dynamic feature vector of mel-cepstrum $\boldsymbol{C}_t = [\boldsymbol{c}_t, \Delta\boldsymbol{c}_t]$ at frame t.

2. 2. 2 Training process

The training process of the production mapping is similar to that of the inversion mapping described in **Section 2. 1. 2**. At each frame, a joint source and target feature vector $\left[\boldsymbol{Y}_t^\top, \boldsymbol{C}_t^\top\right]^\top$ is constructed. Then, the joint probability density function of the source and target features is modeled with a GMM as follows:

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

2. 2. 3 Conversion process

The conversion process of the production mapping is also similar to that of the inversion mapping described in **Section 2. 1. 3**. First, for a given time sequence of the source feature vectors $\boldsymbol{Y}$, the suboptimum mixture component sequence $\hat{\boldsymbol{m}}$ is determined by:

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

Then, the mel-cepstrum sequence vector $\hat{\boldsymbol{c}}$ is determined by:

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

$$\text{subject to } \boldsymbol{Y} = \boldsymbol{W}^{(y)}\boldsymbol{y}, \tag{9}$$

where $\boldsymbol{W}^{(y)}$ is a transformation matrix for creating the joint static and dynamic feature vector sequence $\boldsymbol{Y}$ from the static feature vector sequence $\boldsymbol{y}$. To improve quality of synthetic speech, we also consider the global variance [15] in the production mapping.

3. Articulatory Controllable Speech Modification

Figure 1 depicts process flow of the proposed articulatory controllable speech modification system. Firstly, an input speech signal is analyzed and parameterized into a time sequence of the mel-cepstrum parameters $\boldsymbol{c}$ and that of the source excitation parameters $\boldsymbol{s}$, i.e. log-scaled waveform power and log-scaled F_0. After extracting a time sequence of the mel-cepstral segment feature vectors $\boldsymbol{O}$, that of the articulatory parameters $\boldsymbol{x}$ is estimated from it by using the inversion mapping described in **Section 2. 1**. Then, the estimated articulatory parameters are manipulated as we want. Subsequently, by using the production mapping described in **Section 2. 2**, a time sequence of the corresponding mel-cepstrum parameters $\hat{\boldsymbol{c}}$ is estimated from that of the manipulated articulatory parameters and that of the source excitation parameters $\boldsymbol{s}$. In the end, a modified speech signal is generated from the converted mel-cepstrum parameter sequence $\hat{\boldsymbol{c}}$ and the source excitation parameter sequence $\boldsymbol{s}$ by using vocoder.

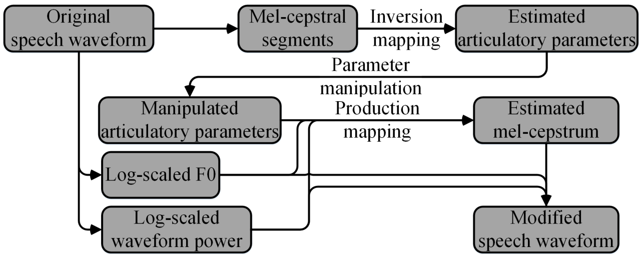

Figure 1 Proposed speech modification process [18]

In this framework, we often want to manipulate only the movements of some specific articulators, such as a tongue tip or velum, rather than to manipulate those of all articulators. In this report, we implement two manipulation methods for modifying the estimated articulatory parameters.

3. 1 Simple manipulation method

Consider the estimated D-dimensional articulatory parameters at frame t, $\hat{\boldsymbol{x}}_t = [\hat{x}_t(1), \cdots, \hat{x}_t(D)]^\top$ that is generated from the inversion mapping. The manipulated articulatory parameters are denoted as $\hat{\boldsymbol{x}}'_t$. In order to modify particular articulatory movements, only corresponding dimensions of articulatory parameters are modified. For example, if the first and second dimensions are modified, then the manipulated articulatory parameters become $\hat{\boldsymbol{x}}'_t = [\hat{x}'_t(1), \hat{x}'_t(2), \cdots, \hat{x}_t(D)]^\top$.

It is well known that some articulatory movements are strongly correlated to each other [16]. As this simple manipulation method ignores the correlation between them, it potentially produce unnatural articulatory movements.

3. 2 Manipulation method considering inter-dimensional correlation of articulatory parameters

In order to take into account the inter-dimensional correlation of articulatory parameters, we propose another method that performs two steps of the inversion mapping. In the first step, the articulatory parameters are estimated from the inversion mapping and then they are manipulated by using the first method as explained previously. In the second step, the modified parts of the articulatory parameters are attached to the source features, then the inversion mapping is performed to refine unmodified parts of the articulatory parameters.

At frame t, the modified parts of the articulatory parameters are denoted as $\hat{\boldsymbol{x}}_t^{(m)}$ and a time sequence of their joint static and dynamic feature vectors is denoted as $\hat{\boldsymbol{X}}^{(m)}$. Whereas, the unmodified parts of the articulatory parameters at frame t are denoted as $\boldsymbol{x}_t^{(u)}$ and a time sequence of their joint static and dynamic feature vectors is denoted as $\hat{\boldsymbol{X}}^{(u)}$. It has to be noted that the sum of dimensions of the unmodified parts $\hat{\boldsymbol{x}}_t^{(m)}$ and the modified parts $\boldsymbol{x}_t^{(u)}$ is equivalent to D. Considering the modified parts as the given features like the source features, the unmodified parts are determined as follows:

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

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

where $\boldsymbol{W}^{(x^{(u)})}$ is a linear transformation matrix for creating a time sequence vector of the joint static and dynamic feature vector of the unmodified articulatory parameters $\hat{\boldsymbol{X}}^{(u)}$ from that of their static feature vector $\boldsymbol{x}^{(u)}$.

The inter-dimensional correlation of the articulatory parameters is modeled by the mixture-dependent full covariance matrices in the inversion mapping. Consequently, in Eq. (10) the unmodified articulatory parameters are refined according to the modified parts. Note that not only inter-dimensional correlation but also inter-frame correlation is considered in this refinement process thanks to the trajectory-based conversion framework [1] using an explicit relationship between the static and dynamic features. Therefore, it is expected that this manipulation method produces more natural articulatory parameter trajectories than the simple manipulation method.

4. Experimental Evaluation

4. 1 Experimental conditions

We used STRAIGHT analysis method to calculate spectral envelope at each frame t. Then, it was converted to a 25-dimensional mel-cepstrum, where 1^{st} through 24^{th} coefficients were used as the spectral envelope parameters. To extract mel-cepstral segments for the inversion mapping, current ± 10 frames were used. As the source excitation parameters, we used log-scaled F_0 values extracted by fixed point analysis [19] in STRAIGHT, which also include unvoiced/voiced binary features, and log-scaled waveform power values extracted from STRAIGHT spectra. For the articulatory parameters, we used 14-dimensional EMA data provided in MOCHA as described in **Section 2.**, which are normalized to Z-scores (zero means and unit variance).

A set of simultaneously recorded acoustic-articulatory data of a British male speaker data set provided in MOCHA [14] was used. We used 350 sentences for training and 110 sentences not included in the training for evaluation. Silence frames were removed using phonetic segmentation included in the MOCHA. Two GMMs were trained separately for the inversion mapping and the production mapping, as described in **Sections 2. 1** and **2. 2**.

We conducted both objective and subjective evaluations. The objective evaluation was conducted to investigate the optimum number of mixture components in the inversion mapping and the production mapping and show accuracy of individual mapping processes. We conducted two subjective evaluations to evaluate the performance of the proposed system.The first subjective evaluation was conducted to compare two articulatory manipulation methods described in **Sections 3. 1** and **3. 2** in terms of speech quality. Second subjective evaluation was conducted to evaluate capability of the proposed system to modify particular vowel sounds as we want by manipulating articulatory movements based on our linguistic knowledge. The number of listeners was ten in each subjective evaluation.

4. 2 Objective evaluation of inversion and production mapping

In the inversion mapping, we calculated a correlation coefficient between the estimated articulatory parameters and the natural articulatory parameters. The number of mixture components was varied between 32, 64, and 128. As a result, 64 mixture components gave the highest correlation coefficient of 0.79 in the inversion mapping, which was comparable to the previous result shown in [1]

In the production mapping, mel-cepstral distortion between the estimated mel-cepstra and the natural mel-cepstra was calculated. The number of mixture components was also varied between 32, 64, and 128. As a result, 64 mixture components also gave the lowest mel-cepstral distortion of 4.70 dB in the production mapping.

We also evaluated the estimation accuracy of mel-cepstrum in the proposed sequential inversion and production mapping processes without performing any modifications of the estimated articulatory parameters. The number of mixture components was set to 64 in both the inversion mapping and the production mapping. The resulting mel-cepstral distortion between the estimated and natural mel-cepstra was 4.45 dB.

4. 3 Evaluation of speech quality for comparison of articulatory manipulation methods

The first subjective evaluation was conducted to evaluate the modified synthetic speech quality from the proposed system. The modification was applied to scale the tongue tip's movement in y-axis. The scale value ranged from 1× to 5× from the originally estimated articulatory parameters. Given 15 distinct sentences, each listeners then evaluated each sentence with an opinion score ranged from 1-to-5 point (bad-to-excellent). The modification was applied by using each of the two proposed articulatory manipulation method.

The result is shown in **Fig. 2**. As the scaling value gets larger, the speech quality also degrades. However, the manipulation method considering inter-dimensional correlation

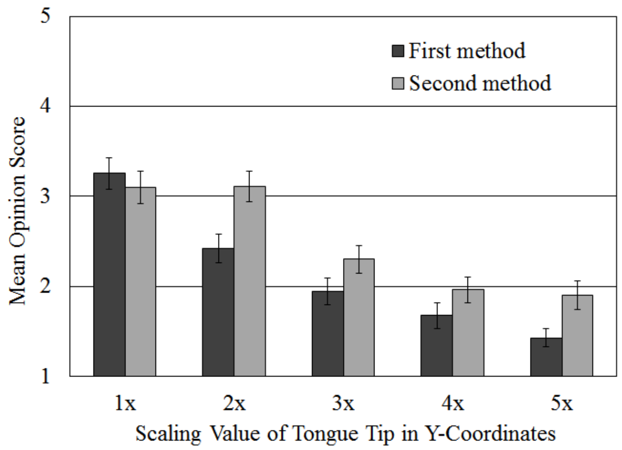

Figure 2 Mean Opinion Score (MOS) on the quality of synthetic speech modified by two proposed manipulation methods described in **Sections 3. 1** and **3. 2** (after [18])

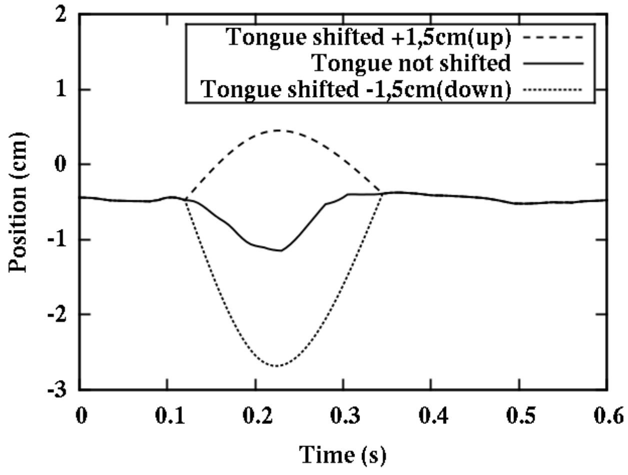

Figure 3 Three various tongue tip trajectories during word "stems"; non-modified, tongue shifted +1,5cm up, and tongue shifted -1,5cm down for the center frame of vowel /ɛ/

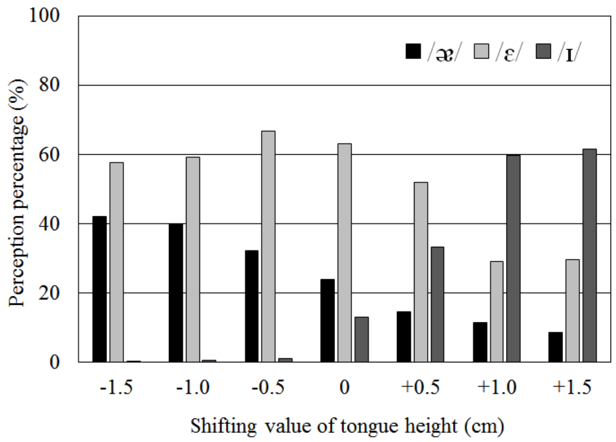

Figure 4 Perception percentage of vowel in modified words resulting from manipulation of tongue's height position (after [18])

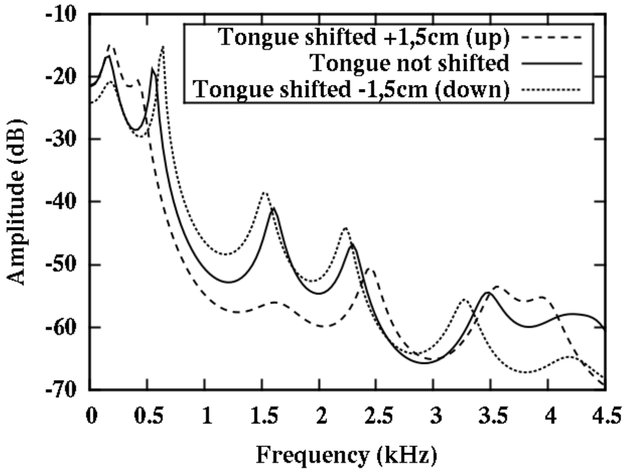

Figure 5 Plot of frequency as a function of amplitude of original vowel /ɛ/ in word "stems" and its modification results by shifting the tongue tip 1,5cm higher and lower

still retain the quality of speech when scaled to 2, as its quality is similar to that of the unmodified one. This method also alleviates the degradation of speech quality compared to the simple method. Overall, the manipulation method considering inter-dimensional correlation is more effective than the simple manipulation method.

4. 4 Evaluation of phoneme sound modification

In the second evaluation, we examined the possibility of modifying particular vowel sounds by manipulating corresponding articulatory movements. We performed similar experiments as in [2], where three front vowels in English were considered, /æ/, /ɛ/ and /ɪ/. Tongue tip's height configuration has the most influence in producing these vowels. In /ɪ/ vowel, the tongue tip is placed to the highest position. In /æ/ vowel, the tongue tip's height is the lowest. Whereas in /ɛ/ vowel, tongue tip's height is between previous two. We picked 10 words from test data, where each of them contained /ɛ/ vowel. The phoneme modification was done by shifting the tongue tip's height value at the center frame of /ɛ/ vowel ranged from -1.5 cm to +1.5 cm in 0.5 cm intervals. In order to produce smoothly varying articulatory trajectories after the manipulation, a cubic spline interpolation is performed [20] among a center frame of the left phoneme of the modified vowel, the modified frame (i.e. the center frame of the modified vowel), and a center frame of the right phoneme of the modified vowel. An example of the manipulated articulatory trajectories is shown in **Fig. 3**. The remaining unmodified articulatory trajectories were also refined by using the manipulation method considering inter-dimensional correlation.

The result is shown in **Fig. 4**. We can observe a clear transition between /ɛ/ and /ɪ/, as the tongue tip's height gets higher. On the other hand, as the tongue tip's height gets lower, the transition from /ɛ/ to /æ/ is not as clear as in that from /ɛ/ and /ɪ/. 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.

Figure 5 shows an example of all pole spectral envelope at the center frame of the modified vowel. Tongue tip height manipulation was performed for the vowel /ɛ/ of the sample word "stems". It is shown that a shift of the tongue tip position from the original position to 1,5 cm higher (+1,5 cm) makes the difference between the first formant frequency (F_1) and the second formant frequency (F_2) larger. It is consistent with the intention to modify vowel /ɛ/ to /ɪ/, as the tongue goes higher, vowel openness goes lower (related to F_1) and vowel frontness goes higher (related to F_2). Furthermore, it also shows the reasonable tendency that F_1 value goes higher and F_2 value goes lower as the tongue gets lower to modify vowel /ɛ/ to /æ/.

5. Conclusions

This report has presented a novel speech modification system based on a sequential inversion and production mapping process with Gaussian mixture models (GMMs). The modification system enables us to modify speech signals by manipulating unobserved articulatory movements. We have proposed a manipulation method considering inter-dimensional correlation of the articulatory movements to refine unmodified parts according to the modified parts on the articulatory parameter trajectories. Results of experimental evaluations have demonstrated that 1) higher speech quality is produced by considering the inter-dimensional correlation when the articulatory manipulation is performed and 2) vowel sounds are well modified by manipulating the corresponding articulatory parameters. We plan to improve quality of the synthetic speech and controllability of the articulatory parameters.

6. Acknowledgements

This research was supported in part by JSPS KAKENHI Grant Number 22680016.

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] 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.

[3] 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.

[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] 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.

[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] Tobing, P. L., Toda, T., Neubig, G., Sakti, S., Nakamura, S., and Purwarianti, A., "Articulatory Controllable Speech Modification Based on Statistical Feature Mapping with Gaussian Mixture Models", Proc. INTERSPEECH, pp. 2298-2302, MAX Atria, Singapore, Sep. 2014.

[19] 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.

[20] Wolberg, G., "Cubic Spline Interpolation: A Review", Department of Computer Science, Columbia University, New York, NY, Technical Report CUCS-389-88, 1988.