Articulatory Controllable Speech Modification based on Gaussian Mixture Models with Direct Waveform Modification using Spectrum Differential

Patrick Lumban Tobing, Kazuhiro Kobayashi, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura

Graduate School of Information Science, Nara Institute of Technology (NAIST), Japan

{patrick.lumbantobing.pf3, kazuhiro-k, tomoki, neubig, ssakti, s-nakamura}@is.naist.jp

Abstract

In our previous work, we have developed a speech modification system capable of manipulating unobserved articulatory movements by sequentially performing speech-to-articulatory inversion mapping and articulatory-to-speech production mapping based on a Gaussian mixture model (GMM)-based statistical feature mapping technique. One of the biggest issues to be addressed in this system is quality degradation of the synthetic speech caused by modeling and conversion errors in a vocoder-based waveform generation framework. To address this issue, we propose several implementation methods of direct waveform modification. The proposed methods directly filter an input speech waveform with a time sequence of spectral differential parameters calculated between unmodified and modified spectral envelop parameters in order to avoid using vocoder-based excitation signal generation. The experimental results show that the proposed direct waveform modification methods yield significantly larger quality improvements in the synthetic speech while also keeping a capability of intuitively modifying phoneme sounds by manipulating the unobserved articulatory movements.

Index Terms: articulatory controllable speech modification, inversion and production mappings, direct waveform modification, spectrum differential, Gaussian mixture models

1. Introduction

Speech is one of the basic and universal ways for people to communicate with each other. During speech production, articulators are responsible in determining the resonance characteristics of the vocal tract by modulating the source excitation signal generated by vocal fold vibration. Therefore, speech can also be characterized by the articulatory parameters, which vary much more slowly than their acoustic parameters counterparts [1], such as the vocal tract spectrum. There have been several attempts at using these articulatory parameters in various speech applications, such as speech synthesis [2], speech recognition [3], and speech coding [4]. Moreover, articulatory parameters can also be used as intermediate features in a speech modification system where an input speech can be easily modified by manipulating the articulatory movements intuitively [5, 6].

The relationship between acoustic and articulatory parameters is basically defined into two mapping systems, namely acoustic-to-articulatory inversion mapping and articulatory-to-acoustic production mapping [7]. Earlier, approaches on developing these mapping systems are based on complex mathematical production models that need some approximations [4, 8]. Recently, statistical relationship between acoustic and articulatory parameters have been modeled in a data-driven manner thanks to a large amount of available parallel acoustic and articulatory database. This statistical approach has been developed for both inversion and production mappings utilizing various techniques, such as codebook [9, 10], hidden Markov models (HMMs) [11, 12], neural networks [13, 14], or Gaussian mixture models (GMMs) [7].

In the previous work, we have integrated both inversion and production mappings based on the GMMs in a unified articulatory controllable speech modification system that allows us to modify an input speech waveform by manipulating the unobserved articulatory parameters [5]. The controllability of the system has been indicated by the feasibility of modifying phonemic sounds through the manipulation of unobserved articulatory movements, which are estimated from the input speech waveform. Moreover, with its independency of the text specification input, this system can be easily implemented for any language.

However, in the conventional system, the quality of modified speech is significantly degraded compared to the original input speech. One of the main factors causing this quality degradation is the use of vocoder-based waveform generation process to generate the modified speech from the converted spectrum and the original source excitation parameters. This vocoder-based waveform generation framework is very sensitive to the errors from extraction and modelling of the spectral and source excitation parameters.

In this paper, in order to improve the quality of the modified speech, we propose a speech modification system with direct waveform modification based on spectrum differential [15] that replaces the vocoder-based waveform generation framework in the articulatory controllable speech modification system. In the direct waveform modification technique, the input natural speech waveform is directly filtered into the modified speech waveform based on a time sequence of spectral differential parameters. Therefore, the quality degradation found in the vocoder-based framework can be alleviated by the direct use of input natural speech. We propose three kinds of estimation process of the spectrum differential parameters through the inversion and production mappings. Experimental results show that proposed methods significantly improve the speech quality whilst also capable of modifying phonemic sounds through the articulatory manipulation.

2. Articulatory Controllable Speech Modification based on GMMs

The articulatory controllable speech modification system consists of two main mapping processes, acoustic-to-articulatory inversion and articulatory-to-acoustic production mapping processes, and a parameter manipulation process. Each of the inversion mapping and production mapping has its own training

and conversion process. The parameter manipulation process is performed to manipulate the articulatory parameters for modifying the articulatory movements.

In the training process of the inversion mapping, a joint probability density function of acoustic parameters (source features) and articulatory parameters (target features) is modeled by a Gaussian mixture model (GMM). Let $\boldsymbol{c}_t$ and $\boldsymbol{x}_t$ be mel-cepstrum parameters and articulatory parameters at frame t, respectively. As the source features, a mel-cepstrum segment feature vector denoted as $\boldsymbol{O}_t$ is used at frame t where it is extracted from mel-cepstrum parameters at multiple frames around the current frame as in [5]. As the target features, a joint static and dynamic feature vector of the articulatory parameters denoted as $\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top\right]^\top$ where $\Delta\boldsymbol{x}_t$ is the dynamic feature vector of the articulatory parameters is used at frame t. Their joint probability is then modeled with a GMM as follows:

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

where $\boldsymbol{\lambda}$ is a GMM parameter set consisting of the weight α_m, the mean vector $\boldsymbol{\mu}_m$, and the covariance matrix $\boldsymbol{\Sigma}_m$ of the m-th mixture component. The normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ is denoted as $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$. The mixture component index is m and the total number of mixture components is M. Then, in the conversion process of the inversion mapping, a time sequence of the articulatory parameters $\hat{\boldsymbol{x}}$ is converted from a given time sequence of the mel-cepstrum segment feature vectors $\boldsymbol{O}$ by maximizing the conditional probability density function analytically derived from the trained GMM for the inversion mapping given in Eq. (1) as follows:

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} P\left(\boldsymbol{X}|\boldsymbol{O}, \boldsymbol{\lambda}^{(O,X)}\right) \quad (2)$$

$$\text{subject to } \boldsymbol{X} = \boldsymbol{W}^{(x)}\boldsymbol{x}, \quad (3)$$

where $\boldsymbol{W}^{(x)}$ is a transformation matrix to expand the static articulatory parameter sequence $\boldsymbol{x}$ into the joint static and dynamic articulatory parameter sequence $\boldsymbol{X}$.

In the training process of the production mapping, another GMM is trained to model the joint probability density function of both articulatory and excitation parameters (source features) and acoustic parameters (target features). Let $\boldsymbol{s}_t$ be the excitation parameters at frame t. As the source features, a joint static and dynamic feature vector of both articulatory parameters and excitation parameters denoted as $\boldsymbol{Y}_t = \left[\boldsymbol{x}_t^\top, \boldsymbol{s}_t^\top, \Delta\boldsymbol{x}_t^\top, \Delta\boldsymbol{s}_t^\top\right]^\top$ is used at frame t. As the target features, a joint static and dynamic feature vector of the mel-cepstrum parameters denoted as $\boldsymbol{C}_t = \left[\boldsymbol{c}_t^\top, \Delta\boldsymbol{c}_t^\top\right]^\top$ is used at frame t. Their joint probability density function is then modeled with another GMM as follows:

$$P\left(\boldsymbol{Y}_t, \boldsymbol{C}_t|\boldsymbol{\lambda}^{(Y,C)}\right) = \sum_{m=1}^{M} \alpha_m^{(Y,C)} \mathcal{N}\left(\left[\boldsymbol{Y}_t^\top, \boldsymbol{C}_t^\top\right]^\top; \boldsymbol{\mu}_m^{(Y,C)}, \boldsymbol{\Sigma}_m^{(Y,C)}\right). \quad (4)$$

Then, in the conversion process of the production mapping, a time sequence of the mel-cepstrum parameters $\hat{\boldsymbol{c}}$ is converted from a given time sequence of articulatory and excitation features $\boldsymbol{Y}$ by maximizing the conditional probability density function analytically derived from the trained GMM for the production mapping given in Eq. (4) as follows:

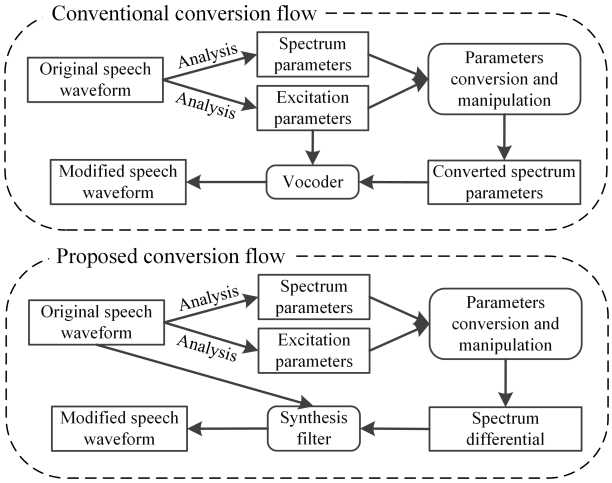

Figure 1: Diagrams of the conventional and proposed articulatory controllable speech modification systems.

$$\hat{\boldsymbol{c}} = \arg\max_{\boldsymbol{c}} P\left(\boldsymbol{C}|\boldsymbol{Y}, \boldsymbol{\lambda}^{(C,Y)}\right) \quad (5)$$

$$\text{subject to } \boldsymbol{C} = \boldsymbol{W}^{(c)}\boldsymbol{c}, \quad (6)$$

where $\boldsymbol{W}^{(c)}$ is a transformation matrix to expand the static mel-cepstrum parameter sequence $\boldsymbol{c}$ into the joint static and dynamic mel-cepstrum parameter sequence $\boldsymbol{C}$. The global variance (GV) [16] is also considered in the production mapping to improve the speech quality.

The complete flow of the conventional articulatory controllable speech modification system [5] illustrated by the upper diagram in Fig. 1 is described as follows. First, an input speech waveform is analyzed into its spectrum and excitation parameters. Then, through the inversion mapping, articulatory parameters are converted from the given spectrum parameters. A parameter manipulation method considering inter-dimensional correlation [5] is then performed in order to manipulate the converted articulatory parameters. Ater that, through the production mapping, corresponding spectrum parameters are converted from the manipulated articulatory parameters. Finally, modified speech waveform is generated from the converted spectrum parameters and the excitation parameters by using a vocoder.

3. Proposed articulatory controllable speech modification with direct waveform modification

The lower diagram in Fig. 1 shows the speech modification process of the proposed system utilizing the direct waveform modification technique [15]. In the proposed system, the use of vocoder-based waveform generation process is avoided by directly modifying the original speech waveform according to the differences between the modified and original spectrum parameters through a filtering process. In this paper, we propose three different methods to estimate the spectrum differential parameters.

The left-side diagram in Fig. 2 shows the modification process of the proposed system TYPE 1. Let $\hat{\boldsymbol{c}}_{mod}$ be a time sequence of the converted spectrum parameters of modified speech and $\boldsymbol{c}_{org}$ be a time sequence of the natural spectrum parameters of input speech. Note that even if using the GV in the production mapping, $\hat{\boldsymbol{c}}_{mod}$ is still oversmoothed compared to $\boldsymbol{c}_{org}$. A time sequence of spectrum differential parameters for the proposed system TYPE 1 is calculated as follows:

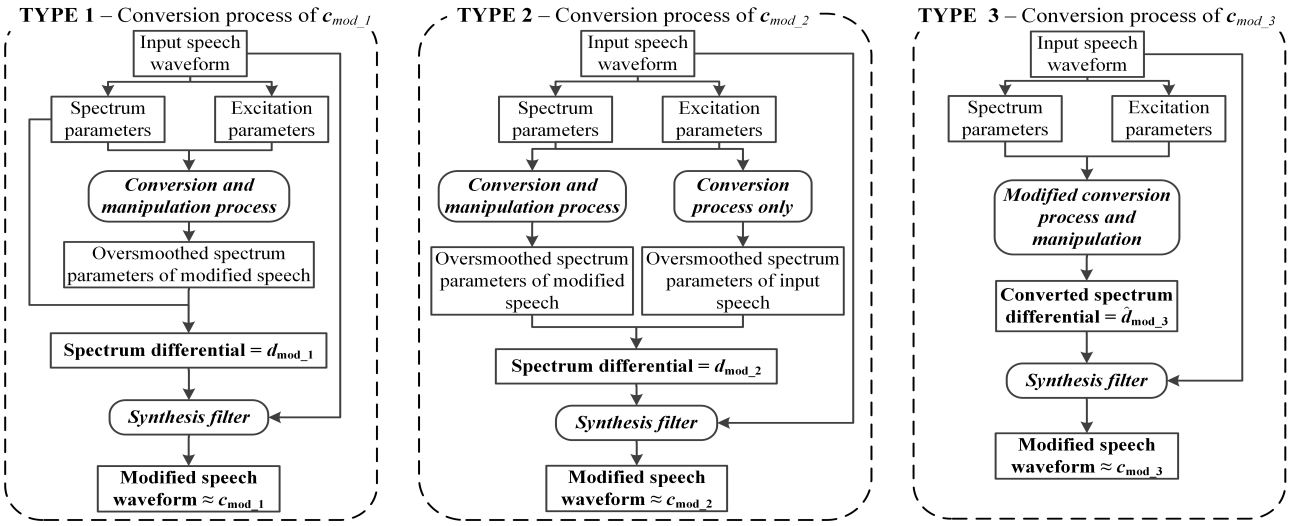

Figure 2: Three different proposed systems with different methods of the spectrum differential estimation.

$$\boldsymbol{d}_{mod\_1} = \hat{\boldsymbol{c}}_{mod} - \boldsymbol{c}_{org}. \tag{7}$$

The modified speech waveform of the proposed system TYPE 1 is then characterized by a time sequence of spectrum parameters $\boldsymbol{c}_{mod\_1}$, which is calculated by filtering the natural spectrum parameters of input speech $\boldsymbol{c}_{org}$ according to the spectrum differential $\boldsymbol{d}_{mod\_1}$ calculated in Eq. (7) as follows:

$$\boldsymbol{c}_{mod\_1} = \boldsymbol{c}_{org} + \left(\hat{\boldsymbol{c}}_{mod} - \boldsymbol{c}_{org}\right). \tag{8}$$

Therefore, the modified speech waveform of the proposed system TYPE 1 is still defined by the oversmoothed spectrum parameters $\hat{\boldsymbol{c}}_{mod}$ as in the conventional system. On the other hand, it is completely different from that of the conventional system in terms of the excitation signal because of directly filtering the input speech waveform without using the vocoder-based excitation generation.

The middle diagram in Fig. 2 shows the flow of the proposed system TYPE 2. Let $\hat{\boldsymbol{c}}_{org}$ be a time sequence of the oversmoothed spectrum parameters of input speech, which is determined by performing the production mapping without any modifications of the estimated articulatory parameters. A time sequence of spectrum differential parameters for the proposed system TYPE 2 is calculated as follows:

$$\boldsymbol{d}_{mod\_2} = \hat{\boldsymbol{c}}_{mod} - \hat{\boldsymbol{c}}_{org}. \tag{9}$$

The modified speech waveform of the proposed system TYPE 2 is then characterized by a time sequence of spectrum parameters $\boldsymbol{c}_{mod\_2}$, which is calculated by filtering the natural spectrum parameters of input speech $\boldsymbol{c}_{org}$ according to the spectrum differential $\boldsymbol{d}_{mod\_2}$ calculated in Eq. (9) as follows:

$$\boldsymbol{c}_{mod\_2} = \boldsymbol{c}_{org} + \left(\hat{\boldsymbol{c}}_{mod} - \hat{\boldsymbol{c}}_{org}\right). \tag{10}$$

Therefore, the modified speech waveform of the proposed system TYPE 2 is defined by not only the oversmoothed spectrum parameters $\hat{\boldsymbol{c}}_{mod}$ but also input residuals given by $\boldsymbol{c}_{org} - \hat{\boldsymbol{c}}_{org}$.

The right-side diagram in Fig. 2 shows the flow of the proposed method TYPE 3. In the proposed system TYPE 3, the spectrum differential is basically calculated in the same principal as the TYPE 2 but in a different manner. Referring to the production mapping process in Section 2, let $\boldsymbol{Y}'$ be a time sequence of the modified source features that is resulted from the manipulation of articulatory parameters. In TYPE 2, the production mapping is performed twice, i.e., the mapping from $\boldsymbol{Y}$ to $\hat{\boldsymbol{c}}_{org}$ and the mapping from $\boldsymbol{Y}'$ to $\hat{\boldsymbol{c}}_{mod}$, and then the spectral differential $\boldsymbol{d}_{mod\_2}$ is calculated. Instead, in TYPE 3, the spectral differential $\boldsymbol{d}_{mod\_3}$ is directly estimated using a differential GMM as follows:

$$\hat{\boldsymbol{d}}_{mod\_3} = \arg\max_{\boldsymbol{d}_{mod\_3}} P\left(\boldsymbol{D}_{mod\_3}|\boldsymbol{Y}', \boldsymbol{Y}, \boldsymbol{\lambda}^{(C,Y)}\right) \tag{11}$$

$$\text{subject to } \boldsymbol{D}_{mod\_3} = \boldsymbol{C}' - \boldsymbol{C} \quad \text{and } \boldsymbol{D}_{mod\_3} = \boldsymbol{W}^{(c)}\boldsymbol{d}_{mod\_3}, \tag{12}$$

where $\boldsymbol{C}'$ denotes a time sequence of the joint static and dynamic spectral parameters given the modified source feature $\boldsymbol{Y}'$. $\boldsymbol{C}$ denotes that given the unmodified source feature $\boldsymbol{Y}$, and the differential GMM is analytically derived from two GMMs for the production mapping used in TYPE 2 in the same manner as described in [15]. The modified speech waveform of the proposed system TYPE 3 is then characterized by a time sequence of spectrum parameters $\boldsymbol{c}_{mod\_3}$, which is calculated by filtering the natural spectrum parameters of input speech $\boldsymbol{c}_{org}$ according to the spectrum differential calculated in Eq. (11) as follows:

$$\boldsymbol{c}_{mod\_3} = \boldsymbol{c}_{org} + \hat{\boldsymbol{d}}_{mod\_3}. \tag{13}$$

In the TYPE 3 proposed system, the modified speech waveform is also defined by both the oversmoothed spectrum parameters and the input residuals as in the TYPE 2 proposed system. Moreover, it is straightforward to further apply some additional techniques such as the GV modelling [16] or the modulation spectrum modelling [17] to the TYPE 3 proposed system as the standard production mapping is performed only once.

4. Experimental evaluation

4.1. Experimental conditions

We used a set of simultaneously recorded speech and EMA data provided in MOCHA [18]. There are a total of 460 utterances spoken by one male British speaker. The sampling rate of the speech data was set to 16 kHz. The EMA data was used as the articulatory parameters.

STRAIGHT analysis [19] was used to extract spectral envelopes which were converted into the 1st-to-24th mel-cepstral coefficients as the spectrum parameters. The fixed-point analysis [20] in STRAIGHT was used to extract F_0 values. Both log-scaled F_0 values, including unvoiced/voiced binary decision feature, and log-scaled power values, extracted from 0-th

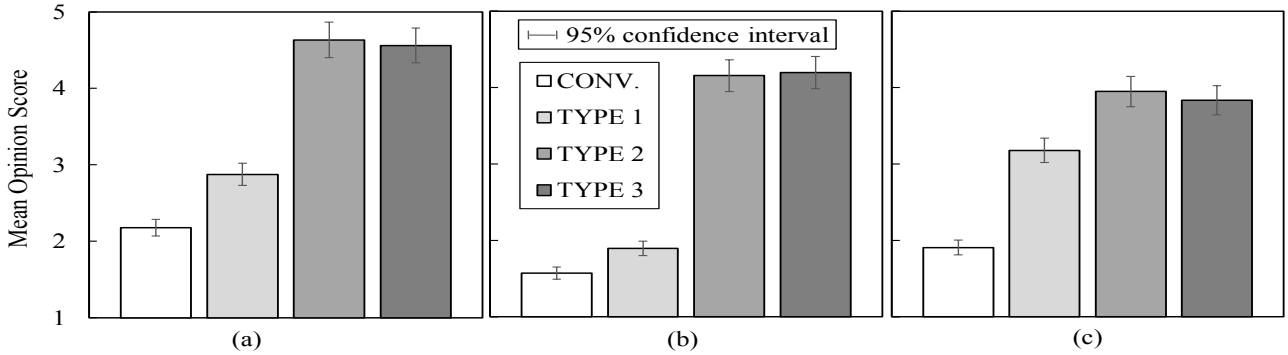

Figure 3: Mean opinion score (MOS) test result for quality evaluation of (a) normal articulated speech with 1.0-fold scaling value, (b) hypo-articulated speech with 0.5-fold scaling value, and (c) hyper-articulated speech with 2.0-fold scaling value.

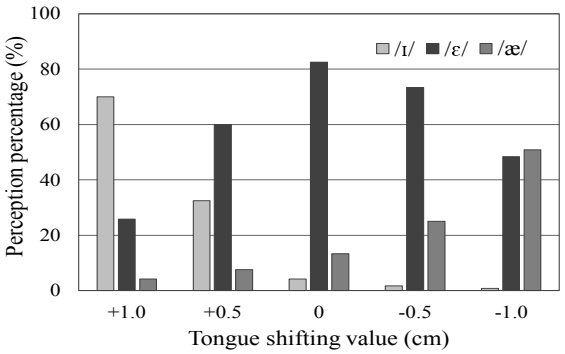

Figure 4: Perception percentage result for controllability evaluation of the proposed system by phoneme modification.

mel-cepstral coefficients, were used as the source excitation parameters. 14-dimensional EMA data in Z-score values, which represented the movements of 7 articulators: upper lip, lower lip, lower incisor, tongue tip, tongue body, tongue dorsum, and velum, in x- and y- coordinates on the midsagittal plane, was used as the articulatory parameters. Frame shift was set to 5 ms. The GV was considered only in the conventional system.

We used 350 utterances for training of two GMMs and 110 utterances for evaluation. The number of mixture components of GMMs for inversion and production mappings was set to 64.

Two subjective evaluations were conducted. In the first evaluation, the quality of generated speech waveforms was evaluated for all four different systems: the conventional system, the proposed systems TYPE 1, TYPE 2, and TYPE 3. Three scaling values were used to modify the articulatory movements. In the second evaluation, we assessed the controllability of the proposed system by modifying phoneme sound through manipulation of the corresponding articulatory movements.

4.2. Speech quality evaluation

We evaluated the quality of generated speech waveform by modifying all dimensions of the articulatory parameters using three different scaling values, 1.0-fold (normal-articulated), 0.5-fold (hypo-articulated), and 2.0-fold (hyper-articulated). The conventional and all three proposed systems were compared in this evaluation. Mean opinion score (MOS) test was used to evaluate by using 5-point scale of scoring from 1-to-5, with 5 as the best score. The number of listeners was 12. Each listener evaluated 96 speech samples including 8 different sentences for each system and scaling value.

Figures 3(a), 3(b), and 3(c) show the results for speech quality in settings of 1.0-fold, 0.5-fold, and 2.0-fold scaling values, respectively. In all three different articulation conditions, the proposed systems TYPE 2 and TYPE 3 significantly improve the quality of generated speech waveform compared to the conventional systems. These results indicate that the proposed direct waveform modification technique makes it possible to alleviate parameterization errors of spectrum and source excitation by avoiding the vocoder-based waveform generation, and to also alleviate the oversmoothing effect by using TYPE 2 and TYPE 3 methods.

4.3. Controllability evaluation by phoneme modification

We also evaluated the controllability of the proposed system in performing the speech modification by changing a specific vowel through articulatory manipulation. Three front vowels in English, /ɛ/, /ɪ/ and /æ/ were considered for the speech modification. Twelve words containing vowel /ɛ/ were chosen from the evaluation data and modified into vowels /ɪ/ and /æ/ by shifting the tongue's height position higher (+0.5 cm and +1.0 cm) and lower (-0.5 cm and -1.0 cm), respectively. Only the proposed system TYPE 3 was used in the evaluation. The number of listeners was 10. Text message corresponding to each speech sample was displayed during playback, where the modified vowel part to be guessed by the listeners as either /ɪ/, /ɛ/ or /æ/ was particularly written with a question mark. Note that the GMMs used in this evaluation were trained using the speech and EMA dataset excluding time frames corresponding to the target vowels /ɪ/ and /æ.

Figure 4 shows the result. A transition from vowel /ɛ/ to vowel /ɪ/ can be clearly seen as the tongue's height position gets higher. Similarly, a transition from vowel /ɛ/ to vowel /æ/ can also be seen as the tongue's height position gets lower. This tendency is similar to that observed in the conventional system as reported in [5]. This result demonstrates that the proposed system is able to perform speech sounds modification with reasonable accuracy through the manipulation of unobserved articulatory movements.

5. Conclusions

In this paper, we propose the articulatory controllable speech modification system based on direct waveform modification technique using spectrum differential to replace the conventional vocoder-based waveform generation framework. Experimental results show that the proposed method significantly improves the quality of generated speech waveform whilst also capable of modifying phoneme sounds through articulatory manipulation. In the future, we plan to implement the parameter algorithm considering the GV or the MS within the proposed framework.

6. Acknowledgements

This research was supported in part by JSPS KAKENHI Grant Number 22680060.

7. References

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

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

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

[4] J. Schroeter and M. M. Sondhi, "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.

[5] 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," *Proc. INTERSPEECH*, pp. 2298–2302, MAX Atria, Singapore, Sep. 2014.

[6] Z. Ling, K. Richmond, J. Yamagishi, and R. Wang, "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] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model," *Speech Communication*, Vol. 50, No. 3, pp. 215–227, Mar. 2008.

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

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

[10] T. Kaburagi, and M. Honda, "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.

[11] S. Hiroya, and M. Honda, "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.

[12] S. Hiroya, and M. Honda, "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.

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

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

[15] K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Statistical Singing Voice Conversion with Direct Waveform Modification based on the Spectrum Differential," *Proc. INTERSPEECH*, pp. 2514–2518, MAX Atria, Singapore, Sep. 2014.

[16] T. Toda, A. W. Black, and K. Tokuda, "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.

[17] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A postfilter to modify the modulation spectrum in HMM-based speech synthesis," *Proc. ICCASP*, pp. 290–294, May, 2014.

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

[19] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne "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.

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