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 (NAIST)

1 Introduction

The mapping systems to model the relationship between articulatory parameters and acoustic parameters with Gaussian mixture models (GMMs) are capable of achieving acoustic-to-articulatory inversion mapping and articulatory-to-acoustic production mapping without any text information [1]. We have recently developed an articulatory controllable speech modification system by combining these two mapping systems [2]. Our system enables to modify input speech waveform by manipulating unobserved articulatory parameters. However, significant quality degradation in the generated speech waveform is caused by using the vocoder-based waveform generation framework, which easily suffers from errors in the spectral and source excitation parameter extraction and modeling.

In this work, we propose the articulatory controllable speech modification system based on a direct waveform modification technique [3] to avoid using the vocoder-based waveform generation. We conduct subjective evaluations, demonstrating that the proposed system can significantly improve quality of the generated speech waveform.

2 GMM-based articulatory controllable speech modification

Our previously proposed articulatory controllable speech modification system is shown in top of **Fig. 1**. First, an input speech waveform is analyzed into its mel-cepstrum parameters and source excitation parameters. The mel-cepstral segments are then extracted from the mel-cepstrum parameters at multiple frames. Then, the articulatory parameters are estimated from the mel-cepstral segments by the GMM for the inversion mapping. These articulatory parameters are then manipulated as we want while also revising unmanipulated articulatory parameters by considering inter-dimensional correlation among the articulatory parameters [2]. The manipulated articulatory parameters together with the source excitation parameters are then converted into the corresponding mel-cepstrum parameters by the GMM for the production mapping. Finally, speech waveform is generated from the converted mel-cepstrum and the natural source excitation parameters by using vocoder.

3 Implementation of waveform modification with spectrum differential

The conventional framework using the vocoder-based waveform generation is sensitive to extraction and modeling errors of the excitation and spectral parameters. To address this issue, the proposed framework directly modifies the input speech waveform according to the spectrum differential between the target and input speech waveforms through a filtering process. A sequence of spectral differentials then needs to be estimated. In this paper, we propose three different estimation methods.

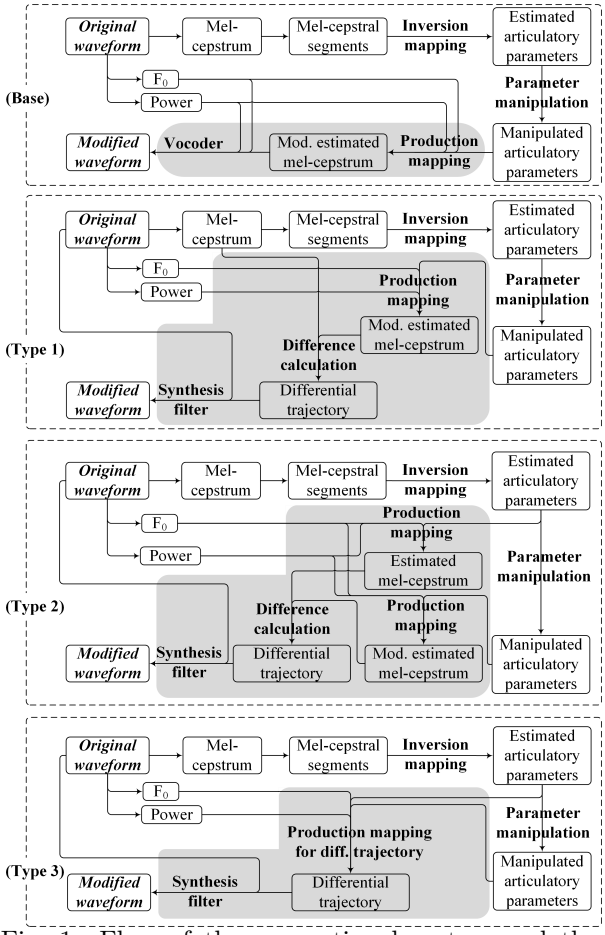

Fig. 1 Flow of the conventional system and the proposed systems

Fig. 1 also illustrates the processing of the three different proposed systems. The first proposed system (**Type 1**) calculates the spectral differential from the modified mel-cepstrum and the original mel-ceptrum extracted from the input speech waveform. This method may suffer from mismatches between the statistically estimated mel-cepstrum and the extracted mel-cepstrum. To avoid them, the second proposed system (**Type 2**) calculates the spectral differential from the modified mel-cepstrum and the converted mel-cepstrum estimated from the unmodified articulatory parameters. On the other

* 混合正規分布モデルと波形変形処理に基づく調音制御を可能とする音声変換 by Patrick Lumban Tobing, 小林 和弘, 戸田 智基, Graham Neubig, Sakriani Sakti, 中村 哲（奈良先端大）

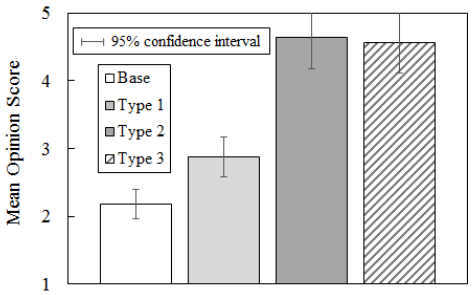

Fig. 2 Mean opinion scores (MOS) result for one-fold scaling value (normal articulation)

hand, the third proposed method (**Type 3**) analytically determines a differential GMM on a joint static and dynamic feature space from the GMM for estimating the modified mel-cepstrum and that for estimating the unmodified mel-cepstrum, and then the spectral differential is generated from the modified GMM. Both the second method and the third method calculate the spectral differential from the statistically estimated mel-cepstra without using the extracted mel-cepstrum, but the second method determines the differential GMM on a static feature sequence space rather than the joint static and dynamic feature space.

4 Experimental evaluation

4.1 Experimental conditions

We used a set of simultaneously recorded speech and articulatory data set with one British male speaker provided in MOCHA [4]. The speech data was sampled at 16 kHz. STRAIGHT [5] was used to extract spectral envelope at each frame. The 1st through 24th mel-cepstral coefficients were used as the spectral parameters. A current $\pm$ 10 frames were used to extract mel-cepstral segments as the source feature for the inversion mapping. As the articulatory parameters, we used the 14-dimensional EMA data in Z-scores. The frame shift was 5 ms. The MLSA filter [6] was used as the synthesis filter in both the vocoder-based waveform generation and the direct waveform modification. We used 350 sentences for training and the remaining 110 sentences for evaluation. Two GMMs were trained separately for the inversion and production mappings. The number of mixture components was set to 64 for both mappings. Note that the global variance (GV) [7] is only considered in the baseline system.

A subjective evaluation was conducted to compare the speech quality of the conventional and all three proposed methods described in **Section 3** by using mean opinion score (MOS) test. All 14-dimensions of the articulatory data were scaled with three different values, one-fold (normal articulation), half-fold (hypo-articulated) and two-fold (hyper-articulated) to evaluate the performance of individual methods in various modification settings. The number of listeners was 12 and each listener evaluated 96 synthetic speech samples consisting of 8 different sentences, which were generated by each system and each scaling setting.

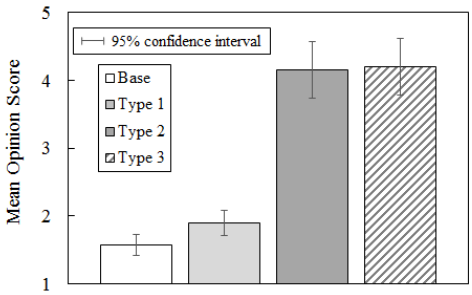

Fig. 3 Mean opinion scores (MOS) result for half-fold scaling value (hypo-articulated)

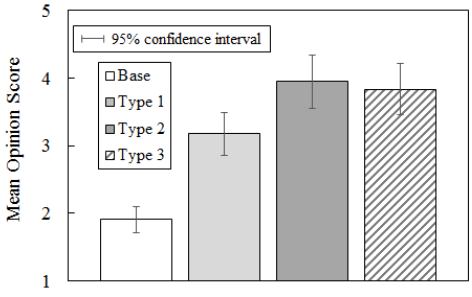

Fig. 4 Mean opinion scores (MOS) result for two-fold scaling value (hyper-articulated)

4.2 Experimental results

Figs. 2, **3**, and **4** show the MOS results for one-fold scaling value, half-fold scaling value, and two-fold scaling value, respectively. These results show that the Type 2 and Type 3 methods can significantly improve quality of the generated speech compared to the conventional method. These two methods can produce similar highest average scores across scaling values.

5 Conclusion

We have proposed the GMM-based articulatory controllable speech modification systems with direct waveform modification using spectrum differential. The results of the subjective evaluations have demonstrated that the proposed systems yield significant quality improvements in the generated speech waveform by avoiding using the vocoder-based waveform generation. We plan to conduct further evaluations in terms of controllability of phoneme sounds.

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

References

[1] T. Toda *et al.*, Speech Communication, vol. 50, No. 3, pp. 215-227, 2008.
[2] P. L. Tobing *et al.*, Proc. INTERSPEECH, pp. 2298-2302, 2014.
[3] K. Kobayashi *et al.*, Proc. INTERSPEECH, pp. 2514-2518, 2014.
[4] A. Wrench, http://www.cstr.ed.ac.uk/artic /mocha.html, 1999.
[5] H. Kawahara *et al.*, Speech Communication, Vol. 27, No. 3–4, pp. 187–207, 1999.
[6] S. Imai *et al.*, Electron. Comm. Jpn., Vol. 66, No. 2, pp. 10–18, 1983.
[7] T. Toda *et al.*, IEEE Trans. ASLP, vol. 15, No. 8, pp. 2222-2235, 2007