# An Evaluation of Articulatory Controllable Speech Modification based on Gaussian Mixture Models with Direct Waveform Modification *

◎ Patrick Lumban Tobing, Kazuhiro Kobayashi, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura (NAIST)

## 1 Introduction

Statistical inversion and production mapping systems based on Gaussian mixture models (GMMs) [1] have a great potential to be developed into various new applications of speech modification. As a first step, we have developed an articulatory controllable speech modification system based on GMMs [2] by sequentially performing the acoustic-to-articulatory inversion mapping and articulatory-to-acoustic production mapping. This system makes it possible to modify an input speech signal by manipulating unobserved articulatory parameters, with several implementations of direct waveform modification method [4]. In this paper, in order to continue our previous evaluation for a male speaker [4], we conduct additional experiments to evaluate both the speech quality and the controllability of articulatory parameters for a female speaker.

## 2 Articulatory controllable speech modification based on GMMs with direct waveform modification

In the GMM-based articulatory controllable speech modification system [2], given an input speech signal, spectrum parameters are extracted and articulatory parameters are estimated using the GMM-based inversion mapping. Then, the articulatory parameters are modified as we want while also revising unmodified parts of them by considering their inter-dimensional correlation [2]. Next, corresponding spectrum parameters are estimated from the modified articulatory parameters using the GMM-based production mapping. Finally, the modified speech waveform is synthesized with a vocoder-based waveform generation process in the conventional system (**Conv**). However, in the conventional system, the quality of the modified speech is degraded compared to the input natural speech due to parameterization errors of spectral and source excitation features in the vocoder-based framework.

To address this issue, we have developed an articulatory controllable speech modification system based on direct waveform modification method [3], capable of avoiding the use of vocoder-based excitation generation by directly filtering the input speech waveform according to spectrum differential parameters [4]. We have proposed three kinds of implementation to estimate the spectrum differential parameters composed by difference values between the spectrum parameters of the modified speech waveform and that of the input speech waveform. In the basic method (**DiffBM**), spectrum differences are calculated between estimated spectrum of the modified speech and natural spectrum of the input speech. Thus, in the filtering process, the natural spectrum is eliminated, giving only the oversmoothed spectrum of the modified speech. In the refined method (**DiffRM**), the differences are calculated between estimated spectrum of the modified speech and estimated spectrum of the input speech waveform, which is determined by performing the inversion and production mappings without any modification of the articulatory parameters. Because both the estimated spectra are oversmoothed, the resulting spectral differentials are also oversmoothed. Hence, in the filtering process, fine structure of spectrum of the input speech is still kept, which alleviates the over-smoothing problem. The third method (**DiffGMM**) is essentially similar to the DiffRM but it directly estimates the over-smoothed spectrum differentials using a differential GMM instead of performing the production mapping twice.

## 3 Investigation of effectiveness of proposed system in typical applications

In this paper, we investigate two typical applications of the proposed articulatory controllable speech modification system: control of articulation conditions and control of phonetic sounds. In the first application, we investigate the speech quality of modified speech signal in hypo and hyper articulation conditions. In the second application, we investigate the capability of modifying phonetic sounds through manipulation of articulatory parameters.

### 3.1 Control of hypo and hyper articulated speech

Hypo and hyper articulated speeches are produced with little effort of articulations (e.g. talk in a very close distance to someone) and with more effort (e.g. in noisy environment), respectively [5]. Therefore, it is expected that hypo articulated speech is generated by reducing the articulatory movements and hyper articulated speech is generated by increasing them. We have reported that the proposed system is capable of generating high speech quality of hypo and hyper articulated speech for a male speaker by scaling the degree of articulation. In this paper, we further evaluate it for a female speaker.

### 3.2 Control of phonetic sounds

Capability of the proposed system in modifying phonetic sounds through articulatory manipu-

*時間波形変形処理と混合正規分布モデルに基づく調音制御機能付き音声変換法の評価 by Patrick Lumban Tobing, 小林 和弘, 戸田 智基, Graham Neubig, Sakriani Sakti, 中村 哲（奈良先端大）

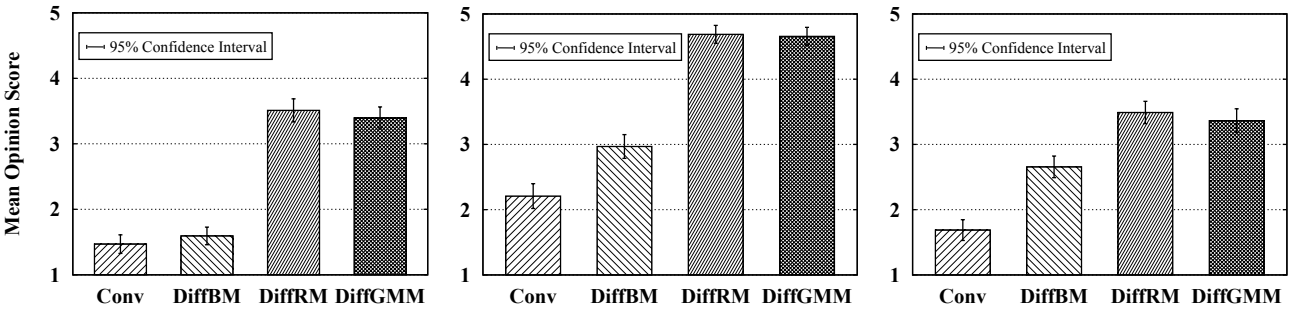


Fig. 1 Mean Opinion Scores (MOS) on three different degree of articulations for female speaker, hypo-articulation (left), normal articulation (centre), and hyper-articulation (right)

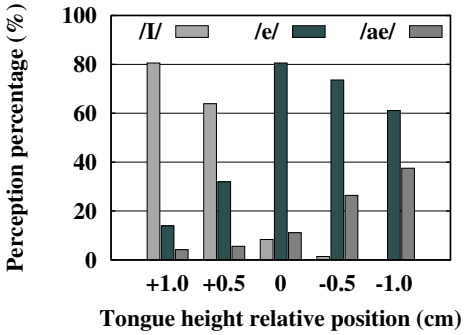


Fig. 2 Perception percentage of the phoneme modification results for the female speaker

lation makes it possible to be implemented in many applications, e.g. language learning, speech recovery, phonetic research, etc. We have reported that the proposed system is capable of modifying vowel sounds of a male speaker by manipulating the tongue movements [4]. In this paper, we also further evaluate it for a female speaker.

# 4 Experimental evaluation

## 4.1 Experimental conditions

We used a set of simultaneously recorded speech and articulatory data of one British female speaker, which is used also in [1]. STRAIGHT analysis [6] was used to extract spectral envelopes which were converted into the 1–24th mel-cepstral coefficients as the spectrum parameters. As the articulatory parameters, 14-dimensional EMA data in Z-scores were used. Two GMMs were trained separately for the inversion and production mappings.

We conducted a speech quality evaluation using the female speaker's data to compare the conventional vocoder-based system and three direct waveform modification-based systems described in **Section 2**. Three degree of articulations were employed by scaling the movements of all articulators: normal (1.0x), hypo-articulation (0.5x), and hyper-articulation (2.0x). 12 listeners were asked to evaluate quality of 96 speech samples each, consisted of 8 different utterances for all four systems and three speaking conditions, in an opinion test.

We also conducted a controllability evaluation by performing modification of three English front vowels, by shifting only the tongue height position. Original vowel /ɛ/ has the middle position, vowel /ɪ/ is in the highest position, and /æ/ has the lowest tongue position. 12 words containing vowel /ɛ/ were chosen from evaluation data. Five modifications of the tongue height position were employed, ranged from -1.0 cm to +1.0 cm. 10 listeners were asked to evaluate phonetic sounds of the modified vowel parts by selecting the most similar vowel among /ɪ/, /ɛ/, and /æ/. Note that time frames corresponding to the target vowels /ɪ/ and /æ/ were removed from the training data set of the GMMs. DiffGMM system was evaluated using the female speaker's data.

## 4.2 Experimental results

Figure 1 and Fig. 2 show the results of the speech quality evaluation and the phoneme modification, respectively. The speech quality evaluation result indicates that hypo and hyper articulated speeches are perceived with a reasonably high quality with the use of DiffRM and DiffGMM. Moreover, the phoneme modification result also demonstrates a tendency of vowel transition, as the tongue height position is changed, which validates the controllability of articulatory movements of the system. These results, which are quite important for this evaluation paper, suggest the capability of the system to perform suitable practices in speech applications.

# 5 Conclusion

We have conducted experimental evaluations to investigate the effectiveness of the proposed articulatory controllable speech modification system based on direct waveform modification. The experimental results have showed that the proposed system is capable of yielding high speech quality in various degree of articulations and also modifying vowel sounds by simply manipulating articulatory movements. We will further investigate a way of manipulating articulatory parameters to develop various speech modification applications.

**Acknowledgements** Part of this work was supported by JSPS KAKENHI Grant Number 26280060.

# 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] P. L. Tobing, *et al.*, Proc. INTERSPEECH, 2015.
[5] B. Picart, *et al.*, Neurocomputing, Vol. 132, pp. 142-147, 2014.
[6] H. Kawahara *et al.*, Speech Communication, Vol. 27, No. 3–4, pp. 187–207, 1999.