Vowel Creation by Articulatory Control in HMM-based Parametric Speech Synthesis

Zhen-Hua Ling[1], *Korin Richmond*[2], *Junichi Yamagishi*[2]

[1]iFLYTEK Speech Lab, University of Science and Technology of China, P.R.China
[2]CSTR, University of Edinburgh, United Kingdom

zhling@ustc.edu, korin@cstr.ed.ac.uk, jyamagis@inf.ed.ac.uk

Hidden Markov model (HMM)-based parametric speech synthesis has become a mainstream speech synthesis method in recent years. This method is able to synthesise highly intelligible and smooth speech sounds. In addition, it makes speech synthesis far more flexible compared to the conventional unit selection and waveform concatenation approach. Several adaptation and interpolation methods have been applied to control model parameters and so diversify the characteristics of the generated speech [1]. However, this flexibility relies upon data-driven machine learning algorithms and it is difficult to integrate phonetic knowledge into the system directly when corresponding training data is not available. In previous work, we have proposed a method to improve the flexibility of HMM-based parametric speech synthesis further by integrating articulatory features [2]. Here, we use "articulatory features" to refer to the continuous movements of a group of speech articulators, such as the tongue, jaw, lips and velum, recorded by human articulography techniques. In this method, a unified acoustic-articulatory HMM is trained. The dependency between acoustic and articulatory features is modelled by a group of linear transforms which are either trained and tied context-dependently [2] or switched in the articulatory feature space [3]. During synthesis, the characteristics of the synthetic speech can be controlled by modifying the generated articulatory features according to phonetic rules.

In this paper, we apply this method of articulatory control to the task of vowel creation in HMM-based parametric speech synthesis. In this task, the target vowel to be created does not occur in the training set, but its phonetic characteristics are known beforehand. We aim to produce this target vowel effectively at synthesis time once appropriate articulatory representations are provided. This is potentially useful for applications such as speech synthesis for limited resource languages, cross-language speaker adaptation, and so on. In our previous approach, articulatory features are treated as HMM observation vectors on which the acoustic features depend. In contrast, in this paper we treat the articulatory features as external explanatory variables for the mean vectors of Gaussians to make it simpler to control synthetic speech via articulation. This model is called a "multiple regression HMM" (MRHMM). Our feature-space transform tying strategy [3] is also applied here. Furthermore, we remove vowel identity from the set of context features used during context-dependent model training in order to ensure compatibility between the estimated model parameters and the articulatory features of a new target vowel at synthesis time.

This work is partially funded by the National Nature Science Foundation of China (Grant No. 60905010) and the National Natural Science Foundation of China - Royal Society of Edinburgh Joint Project (Grant No. 61111130120). The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 256230 (LISTA), and EPSRC grants EP/I027696/1 (Ultrax) and EP/J002526/1 (CAF).

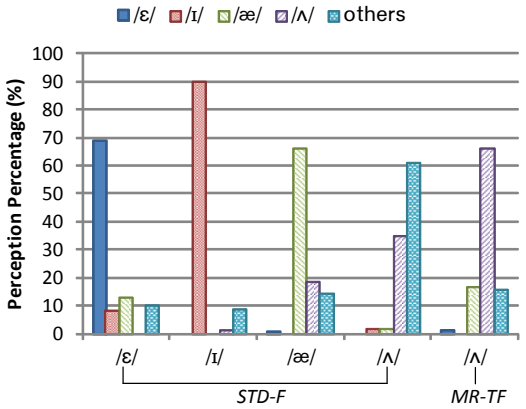

Figure 1: Vowel identity perception results for synthesising different vowels using the baseline system *STD-F* and creating vowel /ʌ/ by articulatory control using the proposed system *MR-TF*.

We have carried out a vowel identity perception test to evaluate the effectiveness of creating the target /ʌ/ vowel. Five monosyllabic words ("but", "hum", "puck", "tun", "dud") containing the /ʌ/ vowel were selected and embedded within a carrier sentence "Now we'll say ... again". For the purpose of comparison, we substituted the vowel /ʌ/ in the five monosyllabic words with /ɛ/, /ɪ/ and /æ/, and then synthesised the respective test sentences using the baseline system. Thirty-two native English listeners were asked to listen to these stimuli and to write down the key word in the carrier sentence they heard. These results are shown in Fig. 1. We see that only 35% of the synthesised vowels /ʌ/ were perceived correctly using the baseline system, due to the lack of acoustic training samples for this vowel. Using the proposed system and the generated articulatory features, this percentage increased to 66.25%, which is close to the perception accuracy of synthesising vowel /ɛ/ (68.75%) and /æ/ (66.25%) using the baseline system.

1. References

[1] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 9, pp. 1406–1413, 2007.

[2] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang, "Integrating articulatory features into HMM-based parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1171–1185, Aug. 2009.

[3] Z.-H. Ling, K. Richmond, and J Yamagishi, "Feature-space transform tying in unified acoustic-articulatory modelling for articulatory control of HMM-based speech synthesis," in *Interspeech*, 2011, pp. 117–120.