Articulatory Control of HMM-based Parametric Speech Synthesis Driven by Phonetic Knowledge

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

[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, rhw@ustc.edu.cn

Abstract

This paper presents a method to control the characteristics of synthetic speech flexibly by integrating articulatory features into a Hidden Markov Model (HMM)-based parametric speech synthesis system. In contrast to model adaptation and interpolation approaches for speaking style control, this method is driven by phonetic knowledge, and target speech samples are not required. The joint distribution of parallel acoustic and articulatory features considering cross-stream feature dependency is estimated. At synthesis time, acoustic and articulatory features are generated simultaneously based on the maximum-likelihood criterion. The synthetic speech can be controlled flexibly by modifying the generated articulatory features according to arbitrary phonetic rules in the parameter generation process. Our experiments show that the proposed method is effective in both changing the overall character of synthesized speech and in controlling the quality of a specific vowel.

Index Terms: speech synthesis, hidden Markov model, articulatory features, phonetic knowledge

1. Introduction

HMM-based parametric speech synthesis has made significant progress during the past decade [1]-[2]. In this method, the spectrum, F0 and segment duration are modeled simultaneously within a unified HMM framework [1]. To perform synthesis, these features are first directly predicted from the HMMs by means of a maximum-likelihood parameter generation algorithm which incorporates dynamic features [2]. Then, the predicted parameters are sent to a parametric synthesizer to generate the synthetic speech waveform. A significant advantage of this model-based parametric approach is that it is far more flexible compared with waveform concatenation (e.g. unit selection). Many model adaptation and interpolation methods can be adopted to modify the model parameters and thus diversify the characteristics of the generated speech [3]-[5]. Unfortunately, however, this flexibility is still constrained by the nature of the training data or adaptation data that is available. For example, to build a speech synthesis system with a child's voice, some training or adaptation data from a child must be available. In many cases, it would be desirable to realize control over synthesis on the basis of phonetic rules and knowledge, as an alternative to relying on purely data-driven approaches. However, it is very difficult to integrate phonetic knowledge concerning the properties of speech, such as the differences between an adult's speech and that of a child, into the system to control the generation of acoustic features directly.

In this paper, we investigate introducing articulatory features into HMM-based synthesis as a way to address this shortcoming. 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 articulographic techniques. Acoustic and articulatory features are of course related, as it is the movement of the articulators that generates the acoustic signal. However, the physical nature of the speech production mechanism gives rise to certain potential advantages in comparison with acoustic features. One significant advantage of articulatory features is that they have physiological meanings and can provide a straightforward and simple explanation for speech characteristics. Hence, it is much more convenient to modify them according to phonetic rules and linguistic knowledge than to modify acoustic features for the purpose of obtaining flexible speech synthesis.

In the method we propose here, a unified statistical model for acoustic and articulatory features is estimated. A piecewise linear transform is used to model the dependency between these two feature streams explicitly. At synthesis time, we manipulate the generated articulatory features and reproduce acoustic features respecting that modified articulatory representation to change the characteristics of the synthesized speech.

This paper is organized as follows. Section 2 first gives a brief overview of the conventional, acoustics-only HMM-based parametric speech synthesis system, and then goes on to describe how this is extended in our proposed method. Section 3 presents the results of our experiments and Section 4 presents the conclusions we draw from this work.

2. Method

2.1. HMM-based parametric speech synthesis

To train a conventional HMM-based speech synthesis system, where only acoustic features are used, the F0 and spectral parameters of D_X dimensions are first extracted from the waveforms contained in the training set. Then, a set of context-dependent HMMs λ are estimated to maximize the likelihood function $P(\boldsymbol{X}|\lambda)$ for the training acoustic features. Here $\boldsymbol{X}=[\boldsymbol{x}_1^T,\boldsymbol{x}_2^T,...,\boldsymbol{x}_N^T]^T$ is the observation feature sequence, $(\cdot)^T$ is the matrix transpose and N is the length of the sequence. The observation feature vector $\boldsymbol{x}_t \in \mathcal{R}^{3Dx}$ for each frame consists of static acoustic parameters $\boldsymbol{x}_{S_t} \in \mathcal{R}^{Dx}$ and their velocity and acceleration components as

$$\boldsymbol{x}_t=[\boldsymbol{x}_{S_t}^T,\Delta\boldsymbol{x}_{S_t}^T,\Delta^2\boldsymbol{x}_{S_t}^T]^T \quad (1)$$

This work is a part of the project "Integrating articulatory features into HMM-based parametric speech synthesis" supported by the Marie Curie Early Stage Training (EST) Network, "Edinburgh Speech Science and Technology (EdSST)". More details about this project are introduced in a paper that has been submitted to IEEE Trans. on Audio, Speech and Lang. Proc. and is currently under review.

$$\Delta \boldsymbol{x}_{S_t} = 0.5\boldsymbol{x}_{S_{t+1}} - 0.5\boldsymbol{x}_{S_{t-1}} \tag{2}$$

$$\Delta^2 \boldsymbol{x}_{S_t} = \boldsymbol{x}_{S_{t+1}} - 2\boldsymbol{x}_{S_t} + \boldsymbol{x}_{S_{t-1}} . \tag{3}$$

A multi-space probability distribution (MSD) [6] is used to model the F0 features. A decision-tree-based model clustering technique is applied after the training for context-dependent HMMs to deal with data-sparsity problems and to avoid over-fitting to the training data. Then, we take the state alignment results using trained HMMs and use them to train context-dependent state duration probabilities [1].

During synthesis, the result of text analysis is used to decide the sentence HMM according to the clustering decision tree. The maximum-likelihood parameter generation algorithm using dynamic features [2] is then applied to generate the optimal static acoustic parameters for each frame such that

$$\boldsymbol{X}_S^* = \arg\max_{\boldsymbol{X}_S} P\left(\boldsymbol{X} \mid \lambda\right) = \arg\max_{\boldsymbol{X}_S} P\left(\boldsymbol{W}_X \boldsymbol{X}_S \mid \lambda\right) \tag{4}$$

where $\boldsymbol{X} = \boldsymbol{W}_X \boldsymbol{X}_S$; $\boldsymbol{X}_S = [\boldsymbol{x}_{S_1}^T, \boldsymbol{x}_{S_2}^T, ..., \boldsymbol{x}_{S_N}^T]^T$ is the static feature sequence; $\boldsymbol{W}_X \in \mathcal{R}^{3ND_X \times ND_X}$ is the matrix used to calculate the complete feature sequence $\boldsymbol{X}$ based on static parameters $\boldsymbol{X}_S$, as introduced in [2].

Finally, these generated parameters are sent to a parametric synthesizer to generate the speech waveform.

2.2. Integrating articulatory features

Our method to integrate articulatory features follows the general framework of HMM-based speech synthesis described in Section 2.1. During training, using parallel acoustic and articulatory observation sequences of the same length N, a statistical model λ for the combined acoustic and articulatory features is estimated to maximize the likelihood function of their joint distribution $P(\boldsymbol{X}, \boldsymbol{Y} \mid \lambda)$, where $\boldsymbol{Y} = [\boldsymbol{y}_1^T, \boldsymbol{y}_2^T, ..., \boldsymbol{y}_N^T]^T$ denotes a given articulatory observation sequence. For each frame, the articulatory feature vector $\boldsymbol{y}_t \in \mathcal{R}^{3D_Y}$ is similarly composed of static component $\boldsymbol{y}_{S_t} \in \mathcal{R}^{D_Y}$ and their velocity and acceleration components as

$$\boldsymbol{y}_t = [\boldsymbol{y}_{S_t}^T, \Delta \boldsymbol{y}_{S_t}^T, \Delta^2 \boldsymbol{y}_{S_t}^T]^T \tag{5}$$

where D_Y is the dimension of the static articulatory features. In order to affect the generation of acoustic parameters by modifying articulatory features, the dependency between these two features is incorporated in λ explicitly by the feature production model as shown in Fig. 1, where the generation of acoustic features is decided not only by the context-dependent acoustic models but also by the parallel articulatory features. Here, these two feature sequences are assumed to be synchronous and generated from the same state sequence.

Similar to [7], a piecewise linear transform is adopted to represent the dependency between these two feature streams by adding a component to the mean vector of the state observation probability density function (PDF) for acoustic features which depends linearly on the articulatory features. Mathematically, we can write the joint distribution as

$$\begin{aligned} P(\boldsymbol{X}, \boldsymbol{Y} \mid \lambda) &= \sum_{\forall \boldsymbol{q}} P(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{q} \mid \lambda) \\ &= \sum_{\forall \boldsymbol{q}} \pi_{q_0} \prod_{t=1}^{N} a_{q_{t-1} q_t} b_{q_t}(\boldsymbol{x}_t, \boldsymbol{y}_t) \end{aligned} \tag{6}$$

$$b_j(\boldsymbol{x}_t, \boldsymbol{y}_t) = b_j(\boldsymbol{x}_t \mid \boldsymbol{y}_t) b_j(\boldsymbol{y}_t) \tag{7}$$

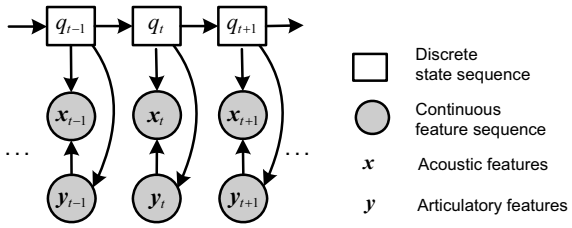

Figure 1: *Feature production model of proposed method.*

$$b_j(\boldsymbol{x}_t \mid \boldsymbol{y}_t) = \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{A}_j \boldsymbol{y}_t + \boldsymbol{\mu}_{\boldsymbol{X}_j}, \boldsymbol{\Sigma}_{\boldsymbol{X}_j}) \tag{8}$$

$$b_j(\boldsymbol{y}_t) = \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_{\boldsymbol{Y}_j}, \boldsymbol{\Sigma}_{\boldsymbol{Y}_j}) \tag{9}$$

where $\boldsymbol{q} = \{q_1, q_2, ..., q_N\}$ denotes the state sequence shared by two feature streams; π_j and a_{ij} represent initial state probability and state transit probability; $b_j(\cdot)$ means the state observation PDF for state j; $\mathcal{N}(;\boldsymbol{\mu}, \boldsymbol{\Sigma})$ represents a Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$; $\boldsymbol{A}_j \in \mathcal{R}^{3D_X \times 3D_Y}$ is the linear transform matrix for state j. The transform matrix is state-dependent, and so a globally piecewise linear transform can be achieved. An EM algorithm can be used to estimate the model parameters. The re-estimation formulas can be derived as

$$\boldsymbol{\mu}_{\boldsymbol{Y}_j}' = \frac{\sum_{t=1}^{T} \gamma_j(t) \cdot \boldsymbol{y}_t}{\sum_{t=1}^{T} \gamma_j(t)} \tag{10}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{Y}_j}' = \frac{\sum_{t=1}^{T} \gamma_j(t) \cdot (\boldsymbol{y}_t - \boldsymbol{\mu}_{\boldsymbol{Y}_j}') \cdot (\boldsymbol{y}_t - \boldsymbol{\mu}_{\boldsymbol{Y}_j}')^T}{\sum_{t=1}^{T} \gamma_j(t)} \tag{11}$$

$$\boldsymbol{A}_j' = \left(\sum_{t=1}^{T} \gamma_j(t) (\boldsymbol{x}_t - \boldsymbol{\mu}_{\boldsymbol{X}_j}) \boldsymbol{y}_t^T \right) \cdot \left(\sum_{t=1}^{T} \gamma_j(t) \boldsymbol{y}_t \boldsymbol{y}_t^T \right)^{-1} \tag{12}$$

$$\boldsymbol{\mu}_{\boldsymbol{X}_j}' = \frac{\sum_{t=1}^{T} \gamma_j(t) \cdot (\boldsymbol{x}_t - \boldsymbol{A}_j' \boldsymbol{y}_t)}{\sum_{t=1}^{T} \gamma_j(t)} \tag{13}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{X}_j}' = \frac{\sum_{t=1}^{T} \gamma_j(t) \cdot (\boldsymbol{x}_t - \boldsymbol{A}_j' \boldsymbol{y}_t - \boldsymbol{\mu}_{\boldsymbol{X}_j}') \cdot (\boldsymbol{x}_t - \boldsymbol{A}_j' \boldsymbol{y}_t - \boldsymbol{\mu}_{\boldsymbol{X}_j}')^T}{\sum_{t=1}^{T} \gamma_j(t)} \tag{14}$$

where $\gamma_j(t)$ is the state occupancy probability of frame t belonging to state j. In order to reduce the number of parameters that need to be estimated, all state-dependent transform matrices are tied to a given class using a decision tree.

For synthesis, the acoustic and articulatory features are simultaneously generated from the trained models based on a maximum-likelihood parameter generation method that considers explicit constraints of the dynamic features, so that

$$\begin{aligned} (\boldsymbol{X}_S^*, \boldsymbol{Y}_S^*) &= \arg\max_{\boldsymbol{X}_S, \boldsymbol{Y}_S} P(\boldsymbol{X}, \boldsymbol{Y} \mid \lambda) \\ &= \arg\max_{\boldsymbol{X}_S, \boldsymbol{Y}_S} \sum_{\forall \boldsymbol{q}} P(\boldsymbol{W}_X \boldsymbol{X}_S, \boldsymbol{W}_Y \boldsymbol{Y}_S, \boldsymbol{q} \mid \lambda) \end{aligned} \tag{15}$$

where

$$\boldsymbol{Y}_S = [\boldsymbol{y}_{S_1}^T, \boldsymbol{y}_{S_2}^T, ..., \boldsymbol{y}_{S_N}^T]^T \tag{16}$$

$$\boldsymbol{Y} = \boldsymbol{W}_{\boldsymbol{Y}} \boldsymbol{Y}_S. \tag{17}$$

$\boldsymbol{W}_{\boldsymbol{Y}} \in \mathcal{R}^{3ND_{\boldsymbol{Y}} \times ND_{\boldsymbol{Y}}}$ is the matrix used to calculate a complete articulatory feature sequence based on static parameters. If only the optimal state sequences is considered in the calculation, (15) can be simplified as

$$\begin{aligned}(\boldsymbol{X}_S^*, \boldsymbol{Y}_S^*) &\simeq \arg\max_{\boldsymbol{X}_S, \boldsymbol{Y}_S} \max_{\boldsymbol{q}} P(\boldsymbol{W}_X \boldsymbol{X}_S, \boldsymbol{W}_Y \boldsymbol{Y}_S, \boldsymbol{q} \mid \lambda) \\ &\simeq \arg\max_{\boldsymbol{X}_S, \boldsymbol{Y}_S} P(\boldsymbol{W}_X \boldsymbol{X}_S, \boldsymbol{W}_Y \boldsymbol{Y}_S \mid \lambda, \boldsymbol{q}^*) P(\boldsymbol{q}^* \mid \lambda)\end{aligned} \tag{18}$$

where

$$\boldsymbol{q}^* = \arg\max_{\boldsymbol{q}} P(\boldsymbol{q} \mid \lambda) \tag{19}$$

is the set of optimal state sequences determined from duration probabilities [1]. The joint distribution can be written as

$$\begin{aligned}\log P(\boldsymbol{W}_X \boldsymbol{X}_S, \boldsymbol{W}_Y \boldsymbol{Y}_S \mid \lambda, \boldsymbol{q}) &= \boldsymbol{X}_S^T \boldsymbol{W}_X^T \boldsymbol{U}_X^{-1} \boldsymbol{A} \boldsymbol{W}_Y \boldsymbol{Y}_S \\ &- \frac{1}{2} \boldsymbol{X}_S^T \boldsymbol{W}_X^T \boldsymbol{U}_X^{-1} \boldsymbol{W}_X \boldsymbol{X}_S + \boldsymbol{X}_S^T \boldsymbol{W}_X^T \boldsymbol{U}_X^{-1} \boldsymbol{M}_X \\ &- \frac{1}{2} \boldsymbol{Y}_S^T \boldsymbol{W}_Y^T (\boldsymbol{U}_Y^{-1} + \boldsymbol{A}^T \boldsymbol{U}_X^{-1} \boldsymbol{A}) \boldsymbol{W}_Y \boldsymbol{Y}_S \\ &+ \boldsymbol{Y}_S^T \boldsymbol{W}_Y^T (\boldsymbol{U}_Y^{-1} \boldsymbol{M}_Y - \boldsymbol{A}^T \boldsymbol{U}_X^{-1} \boldsymbol{M}_X) + K\end{aligned} \tag{20}$$

where

$$\boldsymbol{U}_X^{-1} = diag[\boldsymbol{\Sigma}_{X_{q1}}^{-1}, \boldsymbol{\Sigma}_{X_{q2}}^{-1}, ..., \boldsymbol{\Sigma}_{X_{qN}}^{-1}] \tag{21}$$

$$\boldsymbol{M}_X = [\boldsymbol{\mu}_{X_{q1}}^T, \boldsymbol{\mu}_{X_{q2}}^T, ..., \boldsymbol{\mu}_{X_{qN}}^T]^T \tag{22}$$

$$\boldsymbol{U}_Y^{-1} = diag[\boldsymbol{\Sigma}_{Y_{q1}}^{-1}, \boldsymbol{\Sigma}_{Y_{q2}}^{-1}, ..., \boldsymbol{\Sigma}_{Y_{qN}}^{-1}] \tag{23}$$

$$\boldsymbol{M}_Y = [\boldsymbol{\mu}_{Y_{q1}}^T, \boldsymbol{\mu}_{Y_{q2}}^T, ..., \boldsymbol{\mu}_{Y_{qN}}^T]^T \tag{24}$$

$$\boldsymbol{A} = diag[\boldsymbol{A}_{q1}, \boldsymbol{A}_{q2}, ..., \boldsymbol{A}_{qN}] \tag{25}$$

and K is a constant value. By setting

$$\frac{\partial \log P(\boldsymbol{W}_X \boldsymbol{X}_S, \boldsymbol{W}_Y \boldsymbol{Y}_S \mid \lambda, \boldsymbol{q}^*)}{\partial \boldsymbol{X}_S} = 0 \tag{26}$$

$$\frac{\partial \log P(\boldsymbol{W}_X \boldsymbol{X}_S, \boldsymbol{W}_Y \boldsymbol{Y}_S \mid \lambda, \boldsymbol{q}^*)}{\partial \boldsymbol{Y}_S} = 0, \tag{27}$$

we can obtain the optimal trajectories for acoustic features $\boldsymbol{X}_S^*$ and articulatory features $\boldsymbol{Y}_S^*$ as follows:

$$\boldsymbol{X}_S^* = \left(\boldsymbol{W}_X^T \boldsymbol{U}_X^{-1} \boldsymbol{W}_X\right)^{-1} \boldsymbol{W}_X^T \boldsymbol{U}_X^{-1} \left(\boldsymbol{M}_X + \boldsymbol{A} \boldsymbol{W}_Y \boldsymbol{Y}_S^*\right) \tag{28}$$

$$\begin{aligned}\boldsymbol{Y}_S^* = &\left(\boldsymbol{W}_Y^T \left(\boldsymbol{U}_Y^{-1} + \boldsymbol{A}^T \boldsymbol{U}_X^{-1} \boldsymbol{A} - \boldsymbol{A}^T \boldsymbol{Z}^{-1} \boldsymbol{A}\right) \boldsymbol{W}_Y\right)^{-1} \\ &\cdot \boldsymbol{W}_Y^T \left(\boldsymbol{U}_Y^{-1} \boldsymbol{M}_Y + \boldsymbol{A}^T \boldsymbol{Z}^{-1} \boldsymbol{M}_X - \boldsymbol{A}^T \boldsymbol{U}_X^{-1} \boldsymbol{M}_X\right)\end{aligned} \tag{29}$$

where

$$\boldsymbol{Z}^{-1} = \boldsymbol{U}_X^{-1} \boldsymbol{W}_X \left(\boldsymbol{W}_X^T \boldsymbol{U}_X^{-1} \boldsymbol{W}_X\right)^{-1} \boldsymbol{W}_X^T \boldsymbol{U}_X^{-1}. \tag{30}$$

We can control the characteristics of synthetic speech by modifying the generated articulatory features $\boldsymbol{Y}_S^*$ and reproducing acoustic parameters $\boldsymbol{X}_S^*$ as

$$\boldsymbol{X}_S^* = \left(\boldsymbol{W}_X^T \boldsymbol{U}_X^{-1} \boldsymbol{W}_X\right)^{-1} \boldsymbol{W}_X^T \boldsymbol{U}_X^{-1} \left(\boldsymbol{M}_X + \boldsymbol{A} \boldsymbol{W}_Y \cdot f\left(\boldsymbol{Y}_S^*\right)\right) \tag{31}$$

where $f(\cdot)$ is an arbitrary function to modify the articulatory features. This allows scope to perform modifications based on phonetic rules and linguistic knowledge of speech production. Because articulatory features have a straightforward physiological interpretation, it is much easier to control them than to control acoustic features directly, which as a result makes the speech synthesis more flexible. We should note that in (31) the articulatory features $\boldsymbol{Y}_S^*$ are also *generated*. This means no input of natural acoustic or articulatory features is required to carry out this modification, and so the modification can be performed for arbitrary novel synthetic utterances without supplementary articulatory data.

3. Experiments

3.1. System construction

A multi-channel articulatory database was used in our experiments, recorded using a Carstens AG500 electromagnetic articulograph. A male British English speaker was recorded reading 1,263 phonetically balanced sentences, using a 16kHz PCM wavefile format with 16 bit precision. We have used six EMA sensors, located at the *tongue dorsum*, *tongue body*, *tongue tip*, *lower lip*, *upper lip*, and *lower incisor*. Each receiver recorded spatial location in 3 dimensions at a 200Hz sample rate: coordinates on the *x*- (left to right), *y*- (front to back) and *z*- (bottom to top) axes (relative to viewing the speaker's face from the front). Because all six receivers were placed in the midsagittal plane of the speaker's head, their movements in the *x*-axis were very small. Consequently, only the *y*- and *z*-coordinates of the 6 receivers were used in our experiments, making a total of 12 static articulatory features.

A unified model for acoustic and articulatory features was trained following the proposed method. 1,200 sentences were selected for training and the remaining 63 sentences were used for a test set. 40-order frequency-warped LSFs and an extra gain dimension were derived from the spectral envelop provided by STRAIGHT [8] analysis, with a frame shift of 5ms. A 5-state, left-to-right HMM structure with no skips and diagonal covariance was adopted as the context-dependent phoneme models. Our implementation is based upon The HTS [9] toolkits. The transform matrix $\boldsymbol{A}_j$ was tied to 100 classes and was defined as a three-block matrix corresponding to static, velocity and acceleration components of the feature vector. Then, in the speech synthesis process, we can achieve flexible control of the characteristics of synthetic speech by generating acoustic features with different modification function $f(\cdot)$ in (31).

3.2. Experiments on speech characteristic control

We have tried to control the overall character of synthetic speech using the proposed method. The first experiment is to simulate lengthening a speaker's vocal tract by scaling the *y*-coordinate positions of EMA receivers in modification function $f(\cdot)$. Fig. 2 shows the spectrogram of speech synthesized with an articulatory scaling factor of 1.5 compared with that produced without modification. We find a decrease in formant frequencies and an increase in spectral tilt after modification. These are consistent with our prior knowledge about longer vocal tracts. Fig.3 shows another example, where we increase the *z*-coordinate positions of EMA receivers to simulate a speaking style with a more widely open mouth and more effort. As this figure shows, we find the formants become more pronounced and distinguishable. Listening to the synthesized speech, we perceive that the speech seems less muffled and more intelligible after modification.

In order to demonstrate the feasibility of controlling the quality of synthesized phones by manipulating articulatory features based on some phonetic motivation, a subjective experiment was carried out. We chose three front vowels / ɪ /, / ɛ / and / æ /[1] in English for this experiment. According to phonetic knowledge, the most significant difference in pronunciation between these three vowels is in tongue height. / ɪ / has the highest position, / ɛ / has the middle one and / æ / has the lowest position among these three vowels. In this experiment, $f(\cdot)$ was defined so as to modify the *z*-coordinate positions of EMA receivers corresponding to the *tongue dorsum*, *tongue body*, and *tongue tip* of the speaker. A positive modification means the tongue is raised and a negative value means the tongue is lowered. Five monosyllabic words ("*bet*", "*hem*", "*led*", "*peck*", and "*set*") with vowel / ɛ / were selected and embedded into a carrier sentence "Now we'll say … again". The modification distance was set from -1.5cm to +1.5cm in 0.5cm intervals, meaning we synthesized a total of 35 samples using (31). 20 listeners were asked to listen to these samples and write down the key word in the carrier sentence they heard. For each modification distance we calculated the overall percentage of instances the stimulus was perceived as each of the three vowels. The results are shown in Fig.4. This figure clearly shows the transition of vowel perception from / ɛ / to / ɪ / if we raise the tongue in modification function $f(\cdot)$ and from / ɛ / to / æ / when lowering the tongue.[2]

4. Conclusions

A method that improves the flexibility of conventional HMM-based parametric speech synthesis system through integrating articulatory features and using phonetic knowledge has been proposed. By modeling the dependency between acoustic and articulatory features, we can control articulatory features following phonetic rules and reproduce acoustic parameters with modified speech properties. Our results have proved the effectiveness of the proposed method in changing both the global characteristics of the synthetic speech as well as the quality of specific phones. Importantly, this method requires no additional natural acoustic and articulatory data for target speech, and thus the technique can be employed to synthesize arbitrary novel utterances.

5. Acknowledgements

Korin Richmond and Junichi Yamagishi are funded by the Engineering and Physical Sciences Research Council (EPSRC). The authors thank Prof. Phil Hoole of Ludwig-Maximilian University, Munich for his great effort in helping record the database.

6. References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in Proc. of Eurospeech, 1999, vol. 5, pp. 2347-2350.

[2] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in Proc. of ICASSP, 2000, vol. 3, pp. 1315-1318.

[1] All phonetic symbols in this paper are in International Phonetic Alphabet (IPA) format.

[2] Speech samples used in this experiment can be found at http://www.cstr.ed.ac.uk/research/projects/artsyn/art_hmm/.

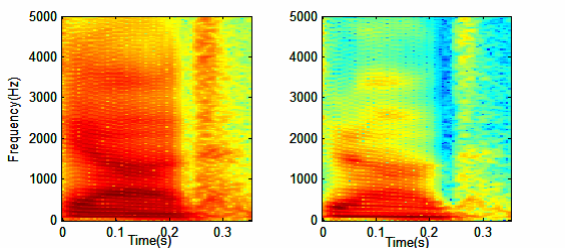

Figure 2: *The spectrograms of synthesized word "yard" without modification (left) and with a 1.5 scaling factor for the y-coordinate positions of all EMA receivers (right).*

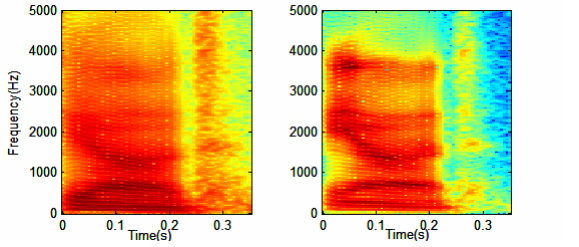

Figure 3: *The spectrograms of synthesized word "yard" without modification (left) and with a 1.5 scaling factor for the z-coordinate positions of all EMA receivers (right).*

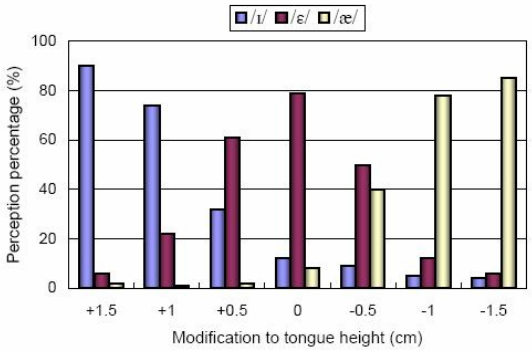

Figure 4: *The result of subjective evaluation on vowel quality perception after modifying the tongue height of EMA features to synthesize vowel /ɛ/.*

[3] J. Yamagishi, and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," IEICE Trans. on Inf. & Syst., vol. E90-D, no.2, pp. 533-543, Feb. 2007.

[4] K. Shichiri, A. Sawabe, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Eigenvoices for HMM-based speech synthesis," in Proc. of ICSLP, 2002, pp.1269-1272.

[5] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "Speech synthesis with various emotional expressions and speaking styles by style Interpolation and morphing," IEICE Trans. Inf. & Syst., vol. E88-D, no. 11, pp. 2484-2491, Nov. 2005.

[6] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in Proc. of ICASSP, 1999, pp. 229-232.

[7] S. Hiroya, and M. Honda, "Estimation of articulatory movements from speech acoustics using an HMM-based speech production model," IEEE Trans. Speech Audio Process., vol. 12, no. 2, pp. 175- 185, 2004.

[8] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using pitchadaptive timefrequency smoothing and an instanta-neousfrequency-based F0 extraction: possible role of a repetitive structure in sounds," Speech Communication, vol. 27, pp. 187-207, 1999.

[9] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0", in The 6th Speech Synthesis Workshop, 2007, pp. 294-299.