Target-Filtering Model based Articulatory Movement Prediction for Articulatory Control of HMM-based Speech Synthesis

Ming-Qi Cai, Zhen-Hua Ling, Li-Rong Dai
iFLYTEK Speech Lab, University of Science and Technology of China, Hefei, China
Email: mqcai@mail.ustc.edu.cn, zhling@ustc.edu.cn, lrdai@ustc.edu.cn

Abstract—**This paper presents a target-filtering model to predict the movements of articulators for articulatory control of hidden Markov model (HMM) based speech synthesis. This model is a bidirectional filtering process on the time-aligned articulation target sequence. The bidirectional filtering could achieve both anticipatory coarticulation and regressive coarticulation. As all the parameters of the model have definite physical meaning, we can control the generation of the articulatory features flexibly with the guidance of articulatory phonetics. And the articulatory features produced by the target-filtering model can be adopted for a multiple regression HMM (MRHMM)-based parametric speech synthesis system. So we can control the pronunciation of vowels by articulatory features instead of the set of context features. Experimental results show that we can control the pronunciation among /ɪ/, /ɛ/, /æ/ effectively just by modifying the articulation targets.**

Keywords-articulatory features; target-filtering model; speech synthesis; articulatory features

I. INTRODUCTION

During the production of human speech, the movements of articulators (e.g. tongue, jaw, lips) generate and shape the acoustic signal. A method to manipulate the pronunciation of vowels by articulatory control in HMM based parametric speech synthesis has been presented in [1]. In this method, articulatory features were adopted as auxiliary features to decide the distribution of acoustic features at each HMM state. Here "articulatory features" refers to the continuous movements of a group of articulators, recorded by electromagnetic articulography (EMA). In our previous work, we focus on modeling the dependency between acoustic and articulatory features. However, a simple but effective model for the prediction of articulatory movements is also very important for the articulatory control of HMM-based speech synthesis. Such a model should be able to generate articulatory features accurately as well as integrate articulatory phonetics easily, i.e. we can control the generation of articulatory features with the guidance of articulatory phonetics.

There have been some related researches on predicting articulatory movements from text. In [2], articulatory movements were predicted from time-aligned phone strings by explicit coarticulation model and Gaussian distribution models at phone midpoints. A kinematic triphone model and a minimum-acceleration model were used to predict the trajectories of articulatory for continuous speech in [3]. However, these two methods above use only simple statistics values as model parameters, and the accuracy of prediction is not satisfactory. Different from the idea that the movement of an articulator is predicted by the interpolation of a sequence of spatial target positions, Ling and Richmond proposed a method to predict the movements of articulators from text using HMM, which could readily take use of acoustic features and fine-grained linguistic features simultaneously [4]. The HMM-based method could achieve an average root mean square (RMS) error of 1.034mm, but the HMM model is too complex to integrate the knowledge of articulatory phonetics and to control the prediction of articulatory features.

This paper presents a modeling approach to predicting articulatory movements from text. This approach was first used by Deng and Yu for predicting formant trajectories. In [5], a quantitative model of coarticulation was presented that could generate formant dynamics in fluent speech using resonance targets in time-aligned phone strings. This model could predict actual formant trajectories for natural speech utterances. More important is that all the parameters have definite physical meaning and can be readily controlled by the guidance of articulatory phonetics. We make a similar assumption as Deng used for the generation of formants, each articulator has a specific articulation target during the generation of a phone together with a stiffness parameter specifying the degree of articulation. The articulation target here means a vector of articulatory features. In the target-filtering model, the stiffness parameter is used to control the temporal filtering of the time-aligned articulation targets.

In the remainder of the paper, Section 2 describes the target-filtering model in detail. Section 3 presents the performance of our method and a series of pronunciation control experiments. At last we make our conclusions in Section 4.

II. METHOD

A. Target-filtering model

The target-filtering model is a bidirectional filtering process on the time-aligned articulation target sequence. The model could achieve both anticipatory coarticulation and regressive coarticulation because when it generates the articulatory movements at each time taking not only the current phone's target but also the adjacent phones' targets into account. The model is consisted of a finite impulse response (FIR) filter characterized by the following non-causal impulse function [5]:

978-1-4673-2197-6/12/$31.00 ©2012 IEEE

$$h_s(k) = \begin{cases} C\gamma_{s(k)}^{-k} & -D \le k \le 0 \\ C & k = 0 \\ C\gamma_{s(k)}^{k} & 0 \le k \le D \end{cases}, \quad (1)$$

where k represents time frame, the frame length in our system is 5 msec. $\gamma_{s(k)}$ is the stiffness parameter specifying the degree of articulation, which is positive and real-valued, ranging from zero to one. The subscript $s(k)$ in $\gamma_{s(k)}$ indicates that the stiffness parameter is dependent on the segments state $s(k)$ which varies over time. D is the unidirectional length of the impulse response, which represents the temporal extent of coarticulation. In our implementation, we set the length for forward and backward direction to be equal for simplicity.

B. *Training articulation targets*

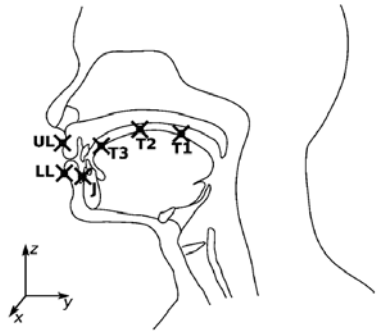
Given the target-filtering model above, the articulation targets are the key component of the model. In this part, we present a method to train the articulation target vectors T_s, which is phone dependent. Given the articulatory features recorded concurrently with the acoustic waveform, this paper presents a maximum likelihood training method [6]. We assume that articulatory features obey a Gaussian distribution, the mean vector is the articulatory trajectory generated from the target-filtering model and the covariance matrix is denoted by Q_s. Then we can write the objective function as:

$$logP = -0.5Q_s^{-1}\sum_{k=1}^{K}[\bar{z}(k) - \hat{g}_s(k)]^2, \quad (2)$$

where $\bar{z}(k)$ is the natural articulatory recordings and $\hat{g}_s(k)$ is the articulatory trajectory produced by the target-filtering model. The detailed algorithms to optimize the target values of each phone by maximizing (2) can be found in [6]. This is an iterative training method, so we need an initiation of the target vectors. First, we choose the instances of each phone with duration over 100msec. Then we calculate the means of the articulatory features of these instances' middle states to get the initial target vectors.

C. *MRHMM-based parametric speech synthesis*

A method of controlling the characteristics of synthetic speech using articulatory features and MRHMM was proposed in [1]. In MRHMM, an auxiliary articulatory feature sequence is used to supplement the state sequence for determining the distribution of acoustic features. To train the MRHMM-based speech synthesis system, context-dependent HMMs are first trained and a decision-tree-based model clustering technique is used to solve the data-sparsity problem. Then, the estimated parameters are used as the initial values in the MRHMM and a zero matrix is used as the initiation for the regression matrix which is used to model the relationship between the articulatory and acoustic features. And these parameters are iteratively updated to maximum $P(\boldsymbol{X}|\lambda, \boldsymbol{Y})$ using EM algorithm. Here, $\boldsymbol{X} = [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \ldots, \boldsymbol{x}_T^\top]^\top$ and $\boldsymbol{Y} = [\boldsymbol{y}_1^\top, \boldsymbol{y}_2^\top, \ldots, \boldsymbol{y}_T^\top]^\top$ are the parallel acoustic and articulatory feature sequence of the same length T, $(\cdot)^\top$ denotes the matrix transpose. Next, a context-dependent state duration model is trained by state-aligned acoustic features. At synthesis time, only the optimal HMM state sequence is considered. This optimal state sequence $\boldsymbol{q}^* = \{q_1^*, q_2^*, \ldots, q_T^*\}$ is determined using the trained duration distribution model. At synthesis time, the articulatory feature sequence $\boldsymbol{Y}$ is firstly predicted from text using the articulatory movement prediction method. Then, the optimal acoustic feature sequence $\boldsymbol{X}^*$ is generated by maximizing $P(\boldsymbol{X}|\lambda, \boldsymbol{Y}, \boldsymbol{q}^*)$ [1].



Label	Location	Label	Location
T1	Tongue dorsum	J	Jaw
T2	Tongue body	LL	Lower lip
T3	Tongue tip	UL	Upper lip

Figure 1. Placement of the six EMA sensors used in the database

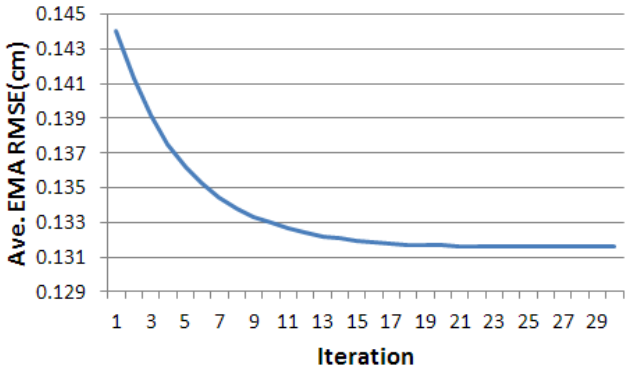


Figure 2. RMS error of articularoty features predicted from text. The x-axis refers to the number of iterations in the articulation targets training.

III. EXPERIMENTS

A. *Database*

In our experiments, an articulatory database which contains 1,263 phonetically balanced sentences read by a male British English speaker was used. A Carstens AG500 electromagnetic articulography was used to record the articulatory movements and acoustic waveforms simultaneously [7]. The waveforms were in 16kHz PCM format with 16-bit precision. Six EMA sensors were used in our experiments, e.g. tongue dorsum, tongue body, jaw, as shown in Fig. 1. Each sensor recorded spatial location in 3 dimensions at a 200Hz sample rate. All these six sensors were placed in the midsagittal plan of the speaker's head, so the values in the x-axis (left to right) were very small. Therefore, only the y-coordinate (front to back) and z-coordinate (bottom to top) of the six sensors were used in our experiments, making a total of 12 static articulatory features. 1,200 sentences were selected for training, and the rest 63 sentences were used as a test set.

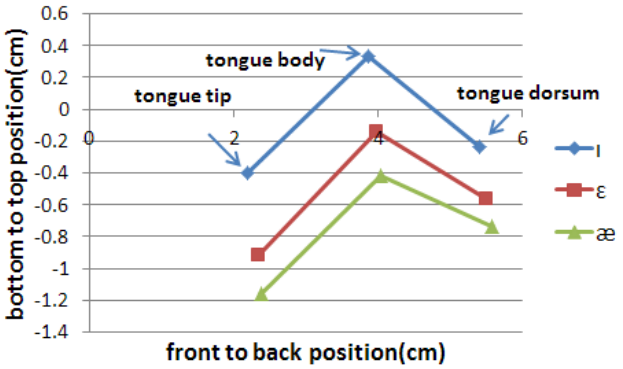


Figure 3. Position of the estimated articulation targets for the tongue for the vowels /ɪ/, /ɛ/ and /æ/.

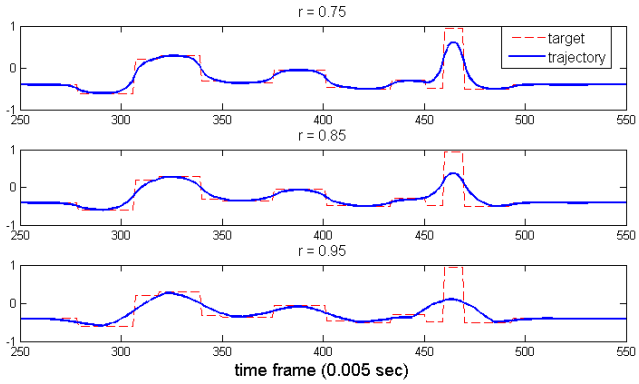


Figure 4. The tongue dorsum's z-axis trajectories for the sentence "Now we will say bet again". The three trajectories from top to bottom correspond to the use of the stiffness parameter values of $\gamma = 0.75, 0.85$ and 0.95 respectively.

B. *Effects of the gradient descent estimation*

The results of the gradient descent estimate are shown in Fig. 2, where we set $\gamma = 0.95, D = 15$ for (1), so the filter has a reaction time of 75 msec, which corresponds to the measured delays between the onsets of muscle activity and articulatory motion [8]. Although the gradient descent method do not guarantee to find the global optimum, but it could get a good result with a fine trained initialization. And the tongue positions of the estimated articulation targets of /ɪ/, /ɛ/ and /æ/ were shown in Fig. 3. We can see that the tongue position of /ɛ/ is between the ones of /ɪ/ and /æ/, which matches the knowledge of articulation phonetics well. Such knowledge will be used in the experiments of vowel quality modification in Section III. E.

C. *Effects of different stiffness parameter*

As we presented in Section II, the stiffness parameter can reflect the degree of articulation. So we illustrate the effects of the target-filtering model's stiffness parameter in Fig. 4. The articulation targets (trained with $\gamma = 0.85$) for the three sentences are the same, but the trajectories produced from the model are obviously different with different stiffness parameter values. The smaller the stiffness parameter value is, the sharper the articulatory trajectory is. So this model could be used to simulate the two degrees of articulation: hyperarticulated speech, which tends to maximize the speech clarity, and hypoarticulated speech is produced with minimal efforts [9].

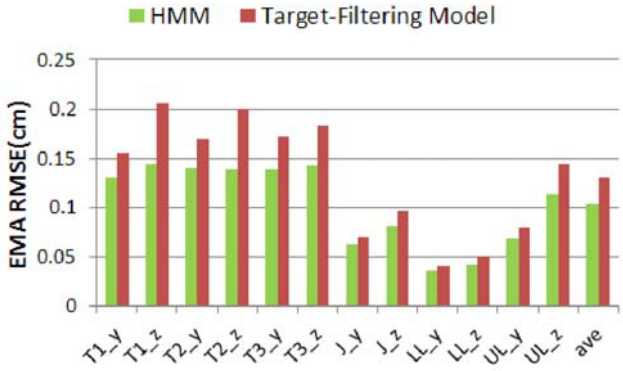


Figure 5. RMS error of articulatory features predicted from text and phone boundaries by HMM-based system and target-filtering model.

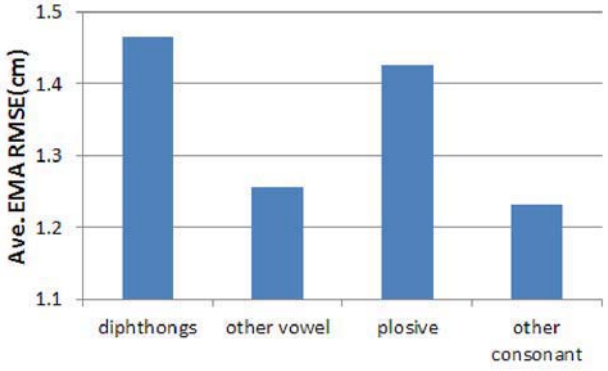


Figure 6. Average RMS error for diphthongs, other vowel, plosive and other consonant.

D. *Prediction of articulatory trajectoris*

In this experiment, the articulatory movements produced by target-filtering model are compared with the articulatory movements predicted by a HMM-based system which used a quinphone model [4]. Only text and phone boundaries are input to the models for the prediction.

The performance of the two systems is compared in Fig. 5, the HMM-based system achieves an average RMS error of 0.1034 cm and an average correlation coefficient of 0.8324 while the target-filtering system achieves an average RMS error of 0.1310 cm and an average correlation coefficient of 0.7384. We note that the target-filtering system underperforms the HMM-based system, but the target-filtering system uses less context features and the computation cost is lower. And the precision of the target-filtering system is acceptable for the aim of controlling speech synthesis.

As presented in Section II, we trained phone dependent articulation targets with the assumption that the articulation targets keep constant during a phone. But there are some compound phones (diphthongs and plosive) in our phone list. We calculated the average RMS error for diphthongs, other vowel, plosive and other consonant separately in Fig. 6. We can see that the compound phones have lager average RMS error than other phones. If we break up these compound phones into their constituents, the target-filtering system could achieve a better performance. It will be a task of our future work.

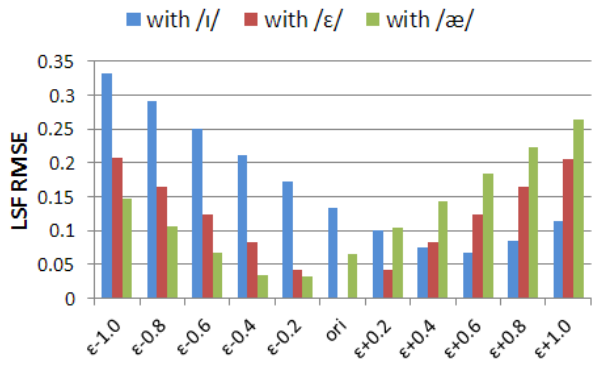


Figure 7. Objective evaluation of LSF RMSE on /ɪ/, /ɛ/ and /æ/. The x-axis indicates how to modify the articulation target.

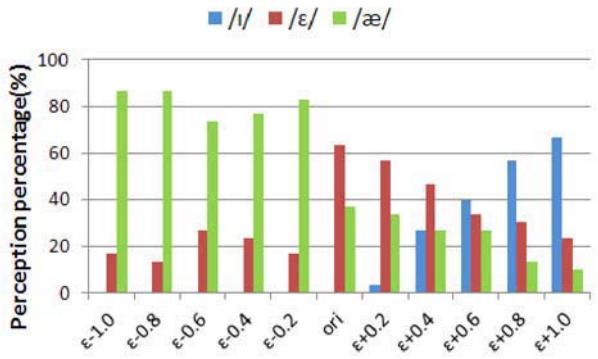


Figure 8. Vowel identity perception results for modifying the articularoty target of /ɛ/.

E. Pronunciation control by modifying articulatory targets

An MRHMM-based parametric speech synthesis system with articulatory control was constructed following the method introduced in [1]. In this system, 100 context-dependent transform matrices were estimated to model the dependency between acoustic and articulatory features. Vowel-related questions were removed from the question set for decision-tree-based model clustering to make the model appropriate for the vowel modification task of this experiment. We carried out a pronunciation controllable experiment on the English vowel /ɛ/. Five monosyllabic words ("bet", "hem", "pek", "ten", "ded") were embedded within a carrier sentence "Now we'll say ... again". As none natural articulatory recording of these sentences exists, we use the target-filtering model and the articulation target of /ɛ/ to produce articulatory features as standard ones. And we modified the articulation target of /ɛ/ by step size of 0.2cm only in 3 dimensions (the z-axis of tongue dorsum, tongue body and tongue tip) to generate articulatory features of these five sentences. These produced articulatory features were adopted for the MRHMM-based speech synthesis system. So 55 sentences were generated in total. We use the RMSE of the generated LSF feature sequence compared with the standard ones for these sentences [10]. The result of objective evaluation is shown in Fig. 7. From this figure, we see that the RMSE between /ɛ/ and /æ/ first decreases then increases when we reduce the target of /ɛ/. As the target of /ɛ/ is 0.6 cm lower than the one of /ɪ/, the RMSE between modified /ɛ/ and /ɪ/ gets a minimum at the iteration of "ɛ+0.6". And the same trend can be observed in the RMSE between modified /ɛ/ and /æ/. And we also carried out a vowel identity perception test to evaluate the effectiveness of the articulatory controllable speech synthesis by modifying articulation target. Six Chinese listeners were asked to listen to these sentences and to write down the key word in the carrier sentence they heard. And the results of the percentages for how the vowels were perceived are shown in Fig. 8, we see that the modification in articulation targets has an obvious effect on the synthesized speech although the listeners are not native English. The perception percentage of /ɪ/ increases with the modification of the articulation target value of /ɛ/ toward the one of /ɪ/. With the modification distance over 0.6cm, the LSF RMSE between modified /ɛ/ and /ɪ/ increases meanwhile the LSF RMSE between modified /ɛ/ and original /ɛ/ increases more. So the perception percentage keeps increasing when the modification distance is over 0.6cm.

IV. CONCLUSION

A target-filtering model has been presented for predicting articulatory moments. Only the phone sequence and phone boundaries are input to the model for prediction. We tend to use this model to control the generation of the articulatory trajectories as the input of a MRHMM-based speech synthesis system, i.e. we can control the pronunciation by just modifying the articulation targets with the knowledge of articulatory phonetics. Both the objective evaluation and subjective evaluation experiments on modifying the English vowel /ɛ/ have shown the effectiveness of this method.

ACKNOWLEDGMENT

This work was partially funded by the National Nature Science Foundation of China (Grant No. 60905010) and the Fundamental Research Funds for the Central Universities (Grant No. WK2100060005).

REFERENCES

[1] Zhen-Hua Ling, Korin Richmond and Junichi Yamagishi, "Vowel Creation by Articulatory Control in HMM-based Parametric Speech Synthesis," Interspeech, 2012.

[2] C. Simon Blackburn and Steve Young, "A self-learning predictive model of articulator movements during speech production," Journal of the Acoustic Society of America 107(3), pp. 1659-1670, 2000.

[3] Takeshi Okadome and Masaaki Honda, "Generation of articulatory movements by using a kinematic triphone model," Journal of the Acoustic Society of America 110(1), pp. 453-462, 2001.

[4] Zhen-Hua Ling, Korin Richmond and Junichi Yamagishi, "An analysis of HMM-based prediction of articulatory movements," Speech Communication 52, pp. 834-846, 2010.

[5] Li Deng, Dong Yu and Alex Acero, "A Quantitative Model for Formant Dynamics and Contextually Assimilated Reduction in Fluent Speech," ICSLP, 2004.

[6] Dong Yu, Li Deng and Alex Acero, "Speaker-adaptive learning of resonance targets in a hidden trajectory model of speech coarticulation," Computer Speech and Language 21, pp. 72-87, 2007.

[7] K. Richmond, P. Hoole and S. King, "Announcing the electro-magnetic articulography (day 1) subset of the mngu0 articulatory corpus," in Interspeech, 2011, pp. 1505-1508.

[8] R. Netsell and B. Daniel, "Neural and mechanical response time for speech production," Journal of Speech and Hearing Research 17, pp. 608-618, 1974.

[9] Benjamin Picart, Thomas Drugman and Thierry Dutoit, "Continuous Control of the Degree of Articulation in HMM-based Speech Synthesis," Interspeech, 2011.

[10] Zhen-Hua Ling, Korin Richmond, Junichi Yamagishi and Ren-Hua Wang, "Integrating Articulatory Features Into HMM-based Parametric Speech Synthesis," IEEE Transactions on Audio, Speech and Language Processing 17(6), PP. 1171-1185, 2009.