Minimum Generation Error Training for HMM-based Prediction of Articulatory Movements

Tian-Yi Zhao, Zhen-Hua Ling, Ming Lei, Li-Rong Dai, Qing-Feng Liu
iFLYTEK Speech Lab, EEIS, University of Science and Technology of China, HeFei
tyzhao@mail.ustc.edu.cn, zhling@ustc.edu, leiming@mail.ustc.edu.cn, lrdai@ustc.edu.cn, qfliu@iflytek.com

Abstract—**This paper presents a minimum generation error (MGE) training method for hidden Markov model (HMM) based prediction of articulatory movements when both text and audio inputs are given. In this method, MGE criterion is adopted to replace the maximum likelihood (ML) criterion to estimate model parameters for the unified acoustic-articulatory HMMs. Different from the MGE training for HMM-based acoustic speech synthesis, the generation error used here is defined as the distance between the generated and natural articulatory features. Experimental results show that our proposed method can improve the accuracy of articulatory movement prediction significantly. The average root mean square (RMS) error reduces from 1.002 mm to 0.913 mm on the test set.**

Keywords-hidden Markov model; articulatory features; minimum generation error training

I. INTRODUCTION

When humans speak, it is the movement of articulators, such as the tongue, jaw, lips and velum, that generates the acoustic signal. These movements of human articulators, i.e. articulatory features, can be recorded by human articulography, such as EMA [1], and they offer an effective description for speech production. Similar to traditional acoustic text-to-speech (TTS) synthesis, the prediction of articulatory features from text also has many potential applications. For example, it could be integrated in an animated talking-head system; or it could help users of a language tutoring system to correct their pronunciation.

Many methods have previously been proposed to predict or estimate articulatory movements, such as [2-4]. In **[**2**]**, combined with time-aligned phone strings, articulator movements were predicted using Gaussian distribution models at phone midpoints together with an explicit coarticulation model. In **[**3**]**, lip shapes (derived from video) were predicted alongside synchronous acoustic speech synthesis parameters from textual input using an HMM-based parameter generation method. The work described in [4] was based on a Gaussian mixture model for the joint distribution of acoustic and articulatory features to achieve the mapping from acoustic features to articulatory movements.

In our previous work [5], we adopted a framework similar to HMM-based parametric speech synthesis to predict the movement of articulators from text. When text was the only input, HMMs were trained using the recorded articulatory features and labeling information. When acoustic features were input with the text, unified acoustic-articulatory HMMs were trained to capture the relationship between the acoustic and articulatory features. Then the optimal trajectories of articulatory movements were generated from the trained models using a maximum-likelihood criterion with dynamic feature constraints. In [5], maximum likelihood criterion was adopted in model training, which may lead to two issues similar to the HMM-based parametric speech synthesis [6]. The first issue is the inconsistency between the model training criterion and the application of articulatory movement prediction. Another one is the ignorance of constraints between static and dynamic features during model training.

A minimum generation error (MGE) training method was proposed [6] to solve these two issues for HMM-based acoustic speech synthesis. In this paper, we introduce MGE criterion into the HMM-based prediction of articulatory movements when both text and audio inputs are available and the unified acoustic- articulatory HMMs are adopted. Here, we define the generation error as the distance between the predicted and natural articulatory parameters. By minimizing this generation error on the training database, model parameters of the unified HMMs are optimized.

The rest of this paper is organized as follows. Section 2 reviews our baseline system, i.e. the unified acoustic-articulatory HMM-based system trained under ML criterion. In Section 3, the proposed MGE training method is described in detail. Finally, the experimental results and conclusion are shown in Section 4 and 5.

II. BASELINE

The framework of HMM-based articulatory movement prediction method using both text and audio inputs is shown in Fig. 1, which consists of a training and a prediction stage.

A. Model Training

In training process of the unified acoustic-articulatory HMMs-based system [7], the parallel acoustic and articulatory observation sequences of the same length T are used to train a statistical model λ for the combined acoustic and articulatory features by maximizing the likelihood function of their joint distribution $P(\boldsymbol{X},\boldsymbol{Y}\mid\lambda)$, where $\boldsymbol{X}=[\boldsymbol{x}_1^{\mathsf{T}},\boldsymbol{x}_2^{\mathsf{T}},...,\boldsymbol{x}_T^{\mathsf{T}}]^{\mathsf{T}}$ and $\boldsymbol{Y}=[\boldsymbol{y}_1^{\mathsf{T}},\boldsymbol{y}_2^{\mathsf{T}},...,\boldsymbol{y}_T^{\mathsf{T}}]^{\mathsf{T}}$ denote the acoustic and articulatory observation sequence respectively, $(\cdot)^{\mathsf{T}}$ is the matrix transpose. For each frame, the acoustic and articulatory feature vector $\boldsymbol{x}_t\in\mathcal{R}^{3D_X}$ and $\boldsymbol{y}_t\in\mathcal{R}^{3D_Y}$ is similarly composed of static component $x_{S_t}\in\mathcal{R}^{D_X}$ and $\boldsymbol{y}_{S_t}\in\mathcal{R}^{D_Y}$, their velocity and acceleration components as

ISBN 978-1-4244-6245-2/10/$26.00 ©2010 IEEE

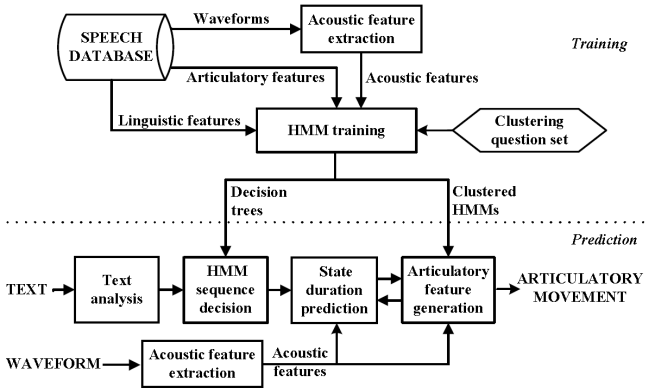

Figure 1. Flowchart of the HMM-based articulatory movement prediction method when both text and audio inputs are given

$$\boldsymbol{x}_t = [\boldsymbol{x}_{S_t}^{\mathsf{T}}, \Delta\boldsymbol{x}_{S_t}^{\mathsf{T}}, \Delta^2\boldsymbol{x}_{S_t}^{\mathsf{T}}]^{\mathsf{T}} \tag{1}$$

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

where D_X and D_Y are the dimensions of the static acoustic and articulatory features.

Synchronous-state model structure [7] is adopted here, which assumes that the acoustic features and the articulatory *features share the same state sequence for each sentence. The dependency* between the acoustic and articulatory features is modeled by a piecewise linear transform within the HMM states [7], which is shown as follows:

$$\begin{aligned} P(\boldsymbol{X},\boldsymbol{Y} \mid \lambda) &= \sum_{\forall \boldsymbol{q}} P(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{q} \mid \lambda) \\ &= \sum_{\forall \boldsymbol{q}} \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(\boldsymbol{x}_t,\boldsymbol{y}_t) \end{aligned} \tag{3}$$

$$b_j(\boldsymbol{x}_t,\boldsymbol{y}_t) = b_j(\boldsymbol{x}_t \mid \boldsymbol{y}_t) b_j(\boldsymbol{y}_t) \tag{4}$$

$$b_j(\boldsymbol{x}_t \mid \boldsymbol{y}_t) = \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{A}_j \boldsymbol{y}_t + \boldsymbol{\mu}_{\boldsymbol{X}_j}, \boldsymbol{\Sigma}_{\boldsymbol{X}_j}) \tag{5}$$

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

where $\boldsymbol{q} = \{q_1, q_2, ..., q_T\}$ 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 probability density function (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 is employed to estimate the model parameters and the details can be found in [7].

Similar to the standard procedure of model training for HMM-based acoustic speech synthesis, the fully context-dependent HMMs of the unified acoustic-articulatory features is built at first. Then a decision tree is trained using the minimum description length (MDL) criterion [8] to cluster the PDFs of all HMM states to handle the data sparsity issue and to estimate parameters for the models whose context description is missing in the training set.

B. *Articulatory Movement Prediction*

At prediction stage, the results of front-end text analysis on the input sentence are used to determine the sentence HMM by consulting the clustering decision tree. Then, the maximum-likelihood parameter generation algorithm [5] is followed to predict the articulatory movements. Here, we assume that the acoustic features are input with the text. Thus, the optimal static articulatory features $\boldsymbol{Y}_S^*$ are predicted as

$$\begin{aligned} \boldsymbol{Y}_S^* &= \arg\max_{\boldsymbol{Y}_S} P(\boldsymbol{Y} \mid \lambda, \boldsymbol{X}_S) \\ &= \arg\max_{\boldsymbol{Y}_S} \sum_{\forall \boldsymbol{q}} P(\boldsymbol{W}_Y \boldsymbol{Y}_S, \boldsymbol{q} \mid \lambda, \boldsymbol{X}_S) \end{aligned} \tag{6}$$

where $\boldsymbol{X}_S$ and $\boldsymbol{Y}_S$ denote the static acoustic and articulatory parameters respectively and $\boldsymbol{X}_S$ is given as input; $\boldsymbol{X} = \boldsymbol{W}_X \boldsymbol{X}_S$, $\boldsymbol{Y} = \boldsymbol{W}_Y \boldsymbol{Y}_S$; $\boldsymbol{W}_X$ and $\boldsymbol{W}_Y$ are determined by the velocity and acceleration calculation functions. Eq. (6) is further simplified by considering only the optimal state sequence $\boldsymbol{q}^*$ as

$$(\boldsymbol{Y}_S^*, \boldsymbol{q}^*) \approx \arg\max_{\boldsymbol{Y}_S, \boldsymbol{q}} P(\boldsymbol{W}_Y \boldsymbol{Y}_S, \boldsymbol{q} \mid \lambda, \boldsymbol{X}_S) \tag{7}$$

where the optimal articulatory features $\boldsymbol{Y}_S^*$ and the optimal state sequence $\boldsymbol{q}^*$ are estimated in an iterative way [5]. Each iteration consists two steps:

1) Optimize articulatory features $\boldsymbol{Y}_S$ given $\boldsymbol{X}_S$ and $\boldsymbol{q}$

$$\boldsymbol{Y}_{Si}^* = \arg\max_{\boldsymbol{Y}_S} P(\boldsymbol{W}_Y \boldsymbol{Y}_S \mid \lambda, q_{i-1}, \boldsymbol{X}_S) \tag{8}$$

where i denotes i-th iteration. For initialization, q_0 is calculated by Viterbi alignment using acoustic features $\boldsymbol{X}$ and an isolated acoustic model. Eq.(8) can be solved by setting $\partial P(\boldsymbol{W}_Y \boldsymbol{Y}_S \mid \lambda, q_{i-1}, \boldsymbol{X}_S) / \partial \boldsymbol{Y}_S = 0$, and we have

$$\begin{aligned} \boldsymbol{Y}_{Si}^* = &\left(\boldsymbol{W}_Y^{\mathsf{T}} \left(\boldsymbol{U}_Y^{-1} + \boldsymbol{A}^{\mathsf{T}} \boldsymbol{U}_X^{-1} \boldsymbol{A}\right) \boldsymbol{W}_Y\right)^{-1} \\ &\cdot \left(\boldsymbol{W}_Y^{\mathsf{T}} \boldsymbol{U}_Y^{-1} \boldsymbol{M}_Y + \boldsymbol{W}_Y^{\mathsf{T}} \boldsymbol{A}^{\mathsf{T}} \boldsymbol{U}_X^{-1} (\boldsymbol{W}_X \boldsymbol{X}_S - \boldsymbol{M}_X)\right) \end{aligned} \tag{9}$$

where $\boldsymbol{M}_X$, $\boldsymbol{M}_Y$, $\boldsymbol{U}_X$, $\boldsymbol{U}_Y$ and $\boldsymbol{A}$ are the model parameters of sentence HMM which is decided by the state sequence. $\boldsymbol{M}_X$ and $\boldsymbol{M}_Y$ are mean vectors; $\boldsymbol{U}_X$ and $\boldsymbol{U}_Y$ are covariance matrices; $\boldsymbol{A}$ is the transform matrix. The detailed definition of these parameters can be found in [5].

2) Optimize state sequence $\boldsymbol{q}$ given $\boldsymbol{Y}_S^*$ and $\boldsymbol{X}_S$

$$\boldsymbol{q}_i^* = \arg\max_{\boldsymbol{q}} P(\boldsymbol{q} \mid \lambda, \boldsymbol{X}_S, \boldsymbol{Y}_{Si}^*) \tag{10}$$

This can be solved by Viterbi algorithm using trained unified models on feature sequence pair $(\boldsymbol{W}_Y \boldsymbol{Y}_{Si}^*, \boldsymbol{X})$. The updated $\boldsymbol{q}_i^*$ is used in the first step of next iteration.

III. MGE TRAINING FOR ARTICULATORY MOVEMENT PREDICTION

Although the previous method can predict articulatory features with good accuracy [5], there are still issues existed in current framework, which would take negative impact on prediction of articulatory features. With the purpose of eliminating the inconsistency between training and generation, and considering the constraints between static and dynamic

features, we introduce MGE training into this framework to optimize parameters of acoustic-articulatory model.

MGE criterion is to optimize model λ by minimizing defined generation error between generated and natural parameters on the training set. Here, the generation error is defined as distance between generated articulatory parameters $\boldsymbol{Y}_S^*$ and natural articulatory parameters $\overline{\boldsymbol{Y}}\boldsymbol{s}$, and we have

$$\bar{\lambda} = \arg\min D(\overline{\boldsymbol{Y}}\boldsymbol{s}, \boldsymbol{Y}_S^*) \tag{11}$$

$$D(\overline{\boldsymbol{Y}}\boldsymbol{s}, \boldsymbol{Y}_S^*) = \sum_{t=1}^{T}\sum_{j=1}^{D_Y}(\bar{y}_{t,j} - y_{t,j}^*)^2 \tag{12}$$

where T is the length of the articulatory features sequence; $\bar{y}_{t,j}$ and $y_{t,j}$ are the j-th dimension of generated and natural static articulatory features at frame t.

Similar to the original MGE training method [6], to minimize this generation error of articulatory features, a probabilistic descent (PD) algorithm is applied to update the model parameters. In the iterative model updating using probabilistic descent, we have

$$\begin{aligned}\lambda(\tau+1) &= \lambda(\tau) - \varepsilon_\tau \frac{\partial D(\overline{\boldsymbol{Y}}\boldsymbol{s}, \boldsymbol{Y}_S^*)}{\partial \lambda}\bigg|_{\lambda=\lambda(\tau)} \\ &= \lambda(\tau) - \varepsilon_\tau \frac{\partial D(\overline{\boldsymbol{Y}}\boldsymbol{s}, \boldsymbol{Y}_S^*)}{\partial \boldsymbol{Y}_S^*} \cdot \frac{\partial \boldsymbol{Y}_S^*}{\partial \lambda}\bigg|_{\lambda=\lambda(\tau)}\end{aligned} \tag{13}$$

where τ is the number of iteration and ε_τ is the updating step size for the τ-th iteration. When diagonal covariance matrices are used, the derivative of generation error with respect to model parameters can be derived from Eq.(9) as

$$\frac{\partial \boldsymbol{Y}_S^*}{\partial \mu_{X\mathrm{ij}}} = -\boldsymbol{R}^{-1}\boldsymbol{W}_Y^\top \boldsymbol{A}^\top \boldsymbol{U}_X^{-1}\boldsymbol{J}_{X\mathrm{ij}} \tag{14}$$

$$\frac{\partial \boldsymbol{Y}_S^*}{\partial v_{X\mathrm{ij}}} = \boldsymbol{R}^{-1}\boldsymbol{W}_Y^\top \boldsymbol{A}^\top \boldsymbol{P}_{Xij}(\boldsymbol{X} - \boldsymbol{M}_X - \boldsymbol{A}\boldsymbol{W}_Y\boldsymbol{Y}_S^*) \tag{15}$$

$$\frac{\partial \boldsymbol{Y}_S^*}{\partial \mu_{\mathrm{Yij}}} = \boldsymbol{R}^{-1}\boldsymbol{W}_Y^\top \boldsymbol{U}_Y^{-1}\boldsymbol{J}_{Y\mathrm{ij}} \tag{16}$$

$$\frac{\partial \boldsymbol{Y}_S^*}{\partial v_{\mathrm{Yij}}} = \boldsymbol{R}^{-1}\boldsymbol{W}_Y^\top \boldsymbol{P}_{Y\mathrm{ij}}(\boldsymbol{M}_Y - \boldsymbol{W}_Y\boldsymbol{Y}_S^*) \tag{17}$$

$$\begin{aligned}\frac{\partial \boldsymbol{Y}_S^*}{\partial a_{\mathrm{ijk}}} = \boldsymbol{R}^{-1}[&\boldsymbol{W}_Y^\top \boldsymbol{L}_{ijk}^\top \boldsymbol{U}_X^{-1}(\boldsymbol{X} - \boldsymbol{M}_X - \boldsymbol{A}\boldsymbol{W}_Y\boldsymbol{Y}_S^*) \\ &- \boldsymbol{W}_Y^\top \boldsymbol{A}^\top \boldsymbol{U}_X^{-1}\boldsymbol{L}_{ijk}\boldsymbol{W}_Y\boldsymbol{Y}_S^*]\end{aligned} \tag{18}$$

where

$$\boldsymbol{R} = \boldsymbol{W}_Y^\top(\boldsymbol{U}_Y^{-1} + \boldsymbol{A}^\top\boldsymbol{U}_X^{-1}\boldsymbol{A})\boldsymbol{W}_Y \tag{19}$$

$\mu_{X\mathrm{ij}}$ and μ_{Yij} are the j-th dimension of mean vectors of state i for acoustic and articulatory features respectively, $v_{X\mathrm{ij}}$ and v_{Yij} are the corresponding variance parameters. a_{ijk} is the linear transform coefficient for state i to model the dependency for j-th dimension of acoustic features on k-th dimension of articulatory features. In the vector $\boldsymbol{J}_{X\mathrm{ij}} \in \mathcal{R}^{3D_X T}$, elements for the j-th dimension of the frames belonging to state i are set to be 1, others are set to be 0, which is the same as $\boldsymbol{J}_{Y\mathrm{ij}}$. In the diagonal matrix $\boldsymbol{P}_{Xij} \in \mathcal{R}^{3D_X T \times 3D_X T}$, diagonal elements for the j-th dimension of the frames belonging to state i are set to be 1, others are set to be 0, which is the same as $\boldsymbol{P}_{Yij}$. In $\boldsymbol{L}_{ijk} \in \mathcal{R}^{3D_X T \times 3D_Y T}$, for the frames belonging to state i, the (j,k) element is set to be 1, and others are set to be 0.

The model parameters given by ML training are used as the initial parameters for the iterative updating. Because the MGE training is conducted on the training set, the state sequences used in parameter generation for model updating are obtained by Viterbi alignment using initial models on both acoustic and articulatory observations and they are fixed in the iterative updating.

IV. EXPERIMENTS

A. Experimental conditions

A multi-channel articulatory database was used in our experiments. It was recorded using a Carstens AG500 electro-magnetic articulograph and contained 1,263 phonetically balanced sentences read by a male British English speaker. 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 sample rate of 200Hz: 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. Therefore, only the y- and z-coordinates of the 6 receivers were used in our experiments, making a total of 12 static articulatory features.

Firstly, a unified model for acoustic and articulatory features was trained under ML criterion as a baseline system. The shared-clustering, state-synchronous and dependent-feature system [7] of acoustic and articulatory HMMs was adopted in our experiment. 1,200 sentences were selected for training and the remaining 63 sentences were used for testing. 40-order frequency-warped LSFs and an extra gain dimension were derived from the spectral envelop provided by STRAIGHT 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 15 classes and was defined as a three-block matrix corresponding to static, velocity and acceleration components of the feature vector.

In MGE training, 1000 sentences of original training set were used for training and other 200 sentences (randomly selected) were used as development set to control the number of iterative updating. Due to large computation complexity, the transform matrix $\boldsymbol{A}_j$ were not updated in this experiment. After each iteration of MGE training, we calculated average RMSE over all 12 articulatory dimensions between natural

and generated parameters on the development set, and select the final optimal iteration number if the reduction of the average RMSE was smaller than a threshold.

When evaluating the performance of proposed method on the training set, development set and test set, we simplified the iterative method of articulatory movement prediction introduced in Section II.B and the optimal state sequences are obtained by Viterbi alignment using the ML-trained model on the combined natural acoustic-articulatory observations to avoid the impact of different state sequences for the baseline and proposed methods.

B. Experimental results

Fig. 2 shows the reduction of the average RMSE (corresponding to total generation error) during MGE training on training set and development set. We can see from this figure that, the generation error on training set was reduced steadily during MGE training. And from that on development set, we can conclude that the optimal iteration number is 4. Thus the optimized model after 4-th iteration of MGE training was used as final optimization result.

Using the final optimized model, we evaluate performance of proposed method on test set by clustered decision tree and tied optimized model (*proposed*). The RMSE of *proposed* method on test set is shown in Fig.3, compared with result from the model before MGE training (*Baseline*) with same state sequence, and the RMSE on training set and development set was also shown as reference. We can see from this figure that *proposed* method is significant better than *baseline* with much smaller RMSE. The relative reduction of RMSE on test set is 8.8%.

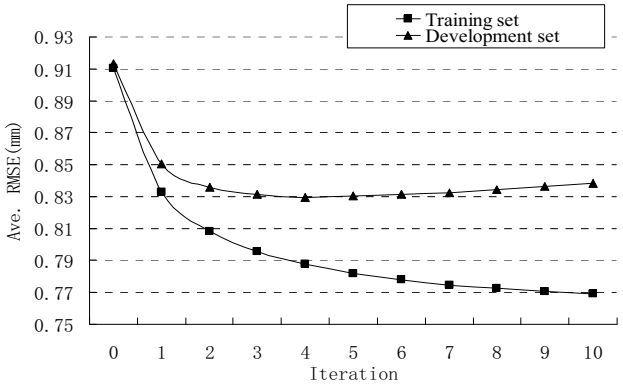

Figure 2. Generation error of articulatory features during MGE training on training set and development set

V. CONCLUSION

In this paper, by considering existed issues in the framework of predicting articulatory features, which are the inconsistency between model training criterion and application of the prediction of articulatory movement, and the ignorance of constraints between static and dynamic parameters during model training, we defined generation error as distance of natural and generated articulatory features, then introduced MGE training to solve the issues and thus to optimize parameters of the unified acoustic-articulatory model. By experiments, we can conclude that MGE training can bring significant improvement for prediction of articulatory features with relative reduction of 8.8% on RMSE of test set. To update the transform matrices in the MGE training and to evaluate our proposed method for different model structures will be the task of our future work.

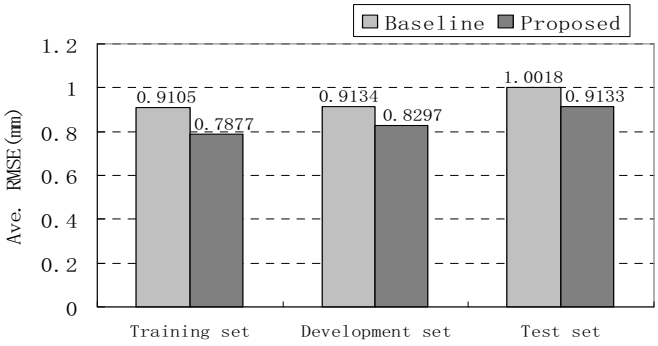

Figure 3. Final RMSE of articulatory movement prediction on training set, development set, and test set.

ACKNOWLDEGEMENT

This work was partially supported by the China Postdoctoral Science Foundation and the National Nature Science Foundation of China (Grant No. 60905010). We thank Phil Hoole of Ludwig-Maximilian University, Munich for his great effort in helping record the EMA data.

REFERENCES

[1] P. W. Schönle, K. Gräbe, P. Wenig, J. Höhne, J. Schrader, and B. Conrad, 1987. Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. Brain Lang., vol. 31, pp. 26–35.

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

[3] M. Tamura, S. Kondo, T. Masuko, T. Kobayashi, 1999. Text-to-audiovisual speech synthesis based on parameter generation from HMM. In: Eurospeech. pp. 959–962.

[4] T. Toda, A. W. Black, K. Tokuda, 2008. Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. Speech Communication 50, pp. 215–227.

[5] Z.-H. Ling, K. Richmond and J. Yamagishi, 2010. An Analysis of HMM-based Prediction of Articulatory Movements. Speech Communication, in press.

[6] Y.-J.Wu, R.-H.Wang, 2006. Minimum Generation Error Training for HMM-Based Speech Synthesis, In: Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP), vol. 1, pp. 889–892.

[7] Z.-H. Ling, K.Richmond, J.Yamagishi and R.-H.Wang,Aug. 2009. Integrating articulatory features into HMM-based parametric speech synthesis. Audio, Speech, and Language Processing, IEEE Transactions on 17 (6), pp.1171–1185.

[8] K. Shinoda, T. Watanabe, 2000. MDL-based context-dependent subword modeling for speech recognition. J. Acoust. Soc. Japan (E) 21 (2), pp.79–86.

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