HMM-based Text-to-Articulatory-Movement Prediction and Analysis of Critical Articulators

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

Abstract

In this paper we present a method to predict the movement of a speaker's mouth from text input using hidden Markov models (HMM). We have used a corpus of human articulatory movements, recorded by electromagnetic articulography (EMA), to train HMMs. To predict articulatory movements from text, a suitable model sequence is selected and the maximum-likelihood parameter generation (MLPG) algorithm is used to generate output articulatory trajectories. In our experiments, we find that fully context-dependent models outperform monophone and quinphone models, achieving an average root mean square (RMS) error of 1.945mm when state durations are predicted from text, and 0.872mm when natural state durations are used. Finally, we go on to analyze the prediction error for different EMA dimensions and phone types. We find a clear pattern emerges that the movements of so-called critical articulators can be predicted more accurately than the average performance.

Index Terms: Hidden Markov model, articulatory features, parameter generation, critical articulators

1. Introduction

When humans produce speech, it is the movement of articulators, such as the tongue, jaw, lips and velum, that generates the acoustic signal. Hence, articulatory features, which may be recorded by EMA [1], can provide an effective alternative description of speech. Similar to acoustic text-to-speech (TTS) synthesis, the generation of articulatory movements from text has many potential applications. For example, it could help users of a language tutoring system learn correct pronunciation; it could be exploited in an animated talking-head system; or it could form the heart of an articulation-based speech synthesis system.

This paper presents an approach to predicting articulatory movements from text which adopts a similar framework to HMM-based parametric speech synthesis [2]. HMMs are trained using the recorded articulatory features and context labelling of the EMA data set. To perform synthesis, the trained models are used in conjunction with a maximum-likelihood criterion with dynamic feature constraints [3] to generate optimal trajectories of articulatory movements.

This work was supported by the Marie Curie Early Stage Training (EST) Network, "Edinburgh Speech Science and Technology (EdSST)". Zhen-Hua Ling is funded by the National Nature Science Foundation of China (Grant No. 60905010). Korin Richmond is funded by the Engineering and Physical Sciences Research Council (EPSRC) grant EP/E01609x/1. Junichi Yamagishi is funded by EPSRC and EC FP7 collaborative project *EMIME*. More details about related work are introduced in a paper that has been submitted to the Speech Communication journal and which is currently under review.

Related research on predicting or estimating articulatory movements has been previously described in [4–7]. In [4], articulatory movements were predicted from phone strings using Gaussian distribution models at phone midpoints and an explicit coarticulation model. In contrast, we use an HMM here to provide temporal modelling of articulatory movements.

The work described in [5] and [6] was similarly based on the HMM. However, their focus was on the acoustic-to-articulatory (inversion) mapping, for which the aim is to estimate articulatory movements from a given acoustic speech signal. This limited them to using only simple context information to define their set of HMMs. In contrast, our aim here is to predict articulatory movements from *text*. Therefore, we can use much more "fine-grained" linguistic features to define our model set.

Finally, a similar HMM-based approach was also described in [7], where speaker adaptive training (SAT) was used to train a *speaker-independent* model to predict articulatory movements from text. The work presented here has two key differences. First, unlike [7], we evaluate using a broad set of linguistic context features for HMM training. Second, in [7], the state durations were not predicted, but derived from the recorded articulatory data by Viterbi alignment. In contrast, we use a statistical model to predict state durations from text. Furthermore, we investigate the influence of different forms of context information on state duration prediction in our experiments.

Finally, it is widely accepted that certain articulators may be more key to the production of a given phone than others. In [8], Papcun et al. presented evidence for what they termed *critical articulators*. They demonstrated, for example, that the variance of trajectories of a point at the tongue dorsum is significantly lower for phones for which this articulator is critical (i.e. for velar oral stops [k,g]) than for phones for which it is not (i.e. alveolar and bilabial stops [t,d,p,b]). The implication is that the movements of articulators that are critical to the production of a given phone are inherently more constrained, and may thus be estimated with lower error, than those which are non-critical. In this paper, we analyze the accuracy of movement prediction using the HMM-based method for the critical articulators of different phone types.

This paper is organized as follows. Section 2 describes the HMM-based articulatory-movement prediction method in detail. Section 3 presents our experiment results, and Section 4 gives the conclusions we draw on the basis of these.

2. Method

The framework of our HMM-based text-to-articulatory movement prediction method is similar to that of the HMM-based parametric speech synthesis. Articulatory movements of di-

mensionality D are recorded by human articulography to provide training data. A set of context-dependent HMMs λ are then trained to maximize the likelihood function $P(\boldsymbol{X}|\lambda)$. Here $\boldsymbol{X} = [\boldsymbol{x}_1^\mathsf{T}, \boldsymbol{x}_2^\mathsf{T}, ..., \boldsymbol{x}_N^\mathsf{T}]^\mathsf{T}$ is the observed articulatory feature sequence, $(\cdot)^\mathsf{T}$ denotes the matrix transpose and N is the length of the sequence. The observation feature vector $\boldsymbol{x}_t \in \mathcal{R}^{3D}$ for each frame consists of static articulatory parameters $\boldsymbol{x}_{S_t} \in \mathcal{R}^D$ and their velocities and accelerations such that

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

where

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

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

Following initial context-dependent HMM training, we use a decision tree to cluster the models [9] in order to address problems of data sparsity and to estimate the parameters for models whose context description is missing in the training set. Next, state alignment results are calculated and used to train context-dependent state duration probabilities [10].

To generate articulatory movements the result of front-end linguistic analysis on the input text is used to determine the sentence HMM by consulting the clustering decision tree built during training. The MLPG algorithm [3] is then applied to generate the optimal articulatory trajectories using dynamic features, such that

$$\boldsymbol{X}_S^* = \arg\max_{\boldsymbol{X}_S} P(\boldsymbol{X}|\lambda) = \arg\max_{\boldsymbol{X}_S} P(\boldsymbol{W}_{\boldsymbol{X}}\boldsymbol{X}_S|\lambda) \quad (4)$$

$$= \arg\max_{\boldsymbol{X}_S} \sum_{\forall \boldsymbol{q}} P(\boldsymbol{W}_{\boldsymbol{X}}\boldsymbol{X}_S, \boldsymbol{q}|\lambda). \quad (5)$$

where $\boldsymbol{X} = \boldsymbol{W}_{\boldsymbol{X}}\boldsymbol{X}_S$; $\boldsymbol{X}_S = [\boldsymbol{x}_{S_1}^\mathsf{T}, \boldsymbol{x}_{S_2}^\mathsf{T}, ..., \boldsymbol{x}_{S_N}^\mathsf{T}]^\mathsf{T}$ is the static articulatory feature sequence; $\boldsymbol{W}_{\boldsymbol{X}} \in \mathcal{R}^{3ND\times ND}$ is determined by the velocity and acceleration calculation functions in (1)-(3); and $\boldsymbol{q} = \{q_1, q_2, ..., q_N\}$ denotes the state sequence for the articulatory features. We solve (5) by keeping only the optimal state sequence in the accumulation and approximating it as a two-step optimization process

$$[\boldsymbol{X}_S^*, \boldsymbol{q}^*] \approx \arg\max_{\boldsymbol{X}_S, \boldsymbol{q}} P(\boldsymbol{W}_{\boldsymbol{X}}\boldsymbol{X}_S, \boldsymbol{q}|\lambda) \quad (6)$$

$$= \arg\max_{\boldsymbol{X}_S, \boldsymbol{q}} P(\boldsymbol{W}_{\boldsymbol{X}}\boldsymbol{X}_S|\lambda, \boldsymbol{q})P(\boldsymbol{q}|\lambda) \quad (7)$$

where the optimal state sequence

$$\boldsymbol{q}^* = \arg\max_{\boldsymbol{q}} P(\boldsymbol{q}|\lambda) \quad (8)$$

is determined from the trained state duration probabilities [10] and $\boldsymbol{X}_S^*$ is calculated by setting $\partial \log P(\boldsymbol{W}_{\boldsymbol{X}}\boldsymbol{X}_S|\lambda, \boldsymbol{q}^*)/\partial \boldsymbol{X}_S = \boldsymbol{0}$, as introduced in [3].

3. Experiments

3.1. Database and System Construction

We have used a multichannel articulatory database for our experiments (`mngu0`). The acoustic waveform was recorded concurrently with articulatory movements using a Carstens AG500 electromagnetic articulograph. A male British English speaker was recorded reading 1,263 phonetically balanced sentences. Of these, 63 sentences were selected evenly from throughout the corpus for a test set, while the remaining 1,200 sentences were used for training. Acoustic waveforms were in 16kHz PCM format with 16 bit precision. As depicted in Fig. 1, six EMA sensors were used, located at the *tongue dorsum* (T3), *tongue body* (T2), *tongue tip* (T1), *lower lip* (LL), *upper lip* (UL), and *lower incisor* (LI). Each sensor recorded spatial location in 3 dimensions at a 200Hz sample rate: coordinates on the x- (front to back), y- (bottom to top) and z- (left to right) axes (relative to viewing the speaker's face from the front). All six sensors were placed in the midsagittal plane, and their movements in the z-axis were very small. Therefore, only the x- and y-coordinates of the six receivers were used in our experiments, making a total of 12 static articulatory features at each sample instant.

Table 1: RMS error (mm) of EMA features predicted from text using monophone (*MONO*), quinphone (*QUIN*), and fully context-dependent (*FULL*) models.

	Predicted State Durations	Natural State Durations
MONO	2.178	1.147
QUIN	2.044	0.881
FULL	1.945	0.872

To create context-dependent HMMs, we first labelled the database using Unilex [11] and Festival [12] tools. Phone boundaries were determined automatically using HTK [13]. Our prediction implementation was based upon the HTS toolkits [14]. We evaluated 3 forms of HMMs:

- **Monophone models.** No context features used.
- **Quinphone models.** Context features comprised the identity of the current phone, together with those of the preceding and follow two neighbouring phones.
- **Fully context-dependent models.** A broad set of linguistic and prosodic features were adopted, similar to those used in HMM-based TTS systems [2], including neighbouring phones (as for quinphone models), lexical stress, part of speech, position in syllable, etc.

3.2. Evaluation of Prediction Accuracy

We have used RMS error calculated for the 63 test sentences (silence segments excluded) and averaged over all 12 EMA features as an objective measure to evaluate the accuracy of articulatory movement prediction. Results for the three model types are given in Table 1. The "predicted state durations" were calculated by solving (8) under the constraint that the total number of generated articulatory frames should be the same as that of the natural utterance [10], in order to facilitate the error calculation. Meanwhile, "natural state durations" were derived by Viterbi alignment with respect to the natural articulatory recordings, similar to [7]. A t-test informs us that the differences among the three systems in each column are significant ($p < 0.05$), with the exception of systems *QUIN* and *FULL* when natural state durations were used. From these results, we make the following observations:

- Context-dependent modelling, which is commonly used in HMM-based speech synthesis, is also effective for predicting articulatory movements from text. Compared with monophone models, using quinphone models improves the accuracy of articulatory feature prediction significantly, as the coarticulatory effects of nearby phones may be taken into account.

- Fully context-dependent models are significantly better than quinphone models when state durations must be predicted. This difference, though, is no longer significant when natural state durations are given. This implies the superiority of the fully context-dependent models lies in better duration prediction. This is reasonable, since they take into account context features related to prosody when training duration distributions.
- RMS error is greatly reduced for all 3 systems when natural state durations are used instead of ones predicted from text. Interestingly, although fully context-dependent models can provide better duration modelling than the monophone and quinphone models, they still fall well short of the performance achieved using natural durations.
- Even when fully context-dependent models and natural state durations are used, we still observe an RMS error of 0.872mm. At least part of this error may be attributed to inherent variability in articulatory movements themselves. Theory suggests the degree of this variability depends on whether a given articulator is "critical" at a given time or not. We explore this in more detail next.

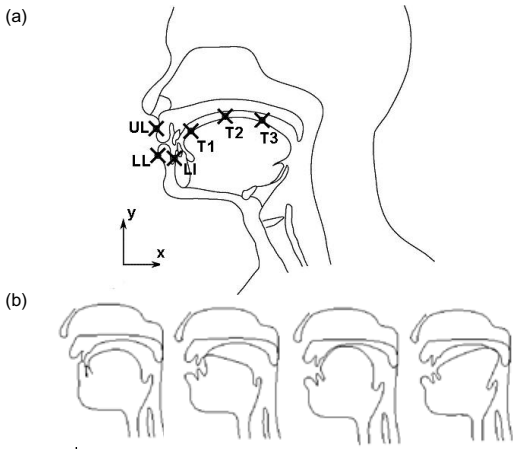
3.3. Movement Prediction for Critical Articulators

In order to investigate whether the accuracy of movement prediction might vary depending on how critical an articulator is to the production of a given phone, we have calculated RMS error for specific EMA sensor coordinates and phone types. Fig. 2 shows normalized RMS error for the y-coordinates of the LL, T1, T2 and T3 sensors according to phone type. Fully context-dependent HMMs and natural state durations have been used here. We observe that the movements of critical articulators can indeed be predicted more accurately than the average performance. Specifically, we note:

For vowels The position of the tongue body is important for defining the shape of the vocal tract. Fig. 2 shows that T2_y has the lowest prediction error (0.303) among the four EMA dimensions for type "Vowel", which is lower than the average T2_y error for all phones (0.317).

For consonants What constitutes a critical articulator depends upon a phone's place of articulation, e.g. the point where an obstruction occurs in the vocal tract. Fig. 1 illustrates the place of articulation for several consonant types, together with the placement of EMA sensors used in our experiments. It shows that the critical articulators for "Labiodental", "Alveolar", "Palatal" and "Velar" phone types correspond to the LL, T1, T2 and T3 sensors respectively. The clear pattern which emerges is that the critical articulator for each consonant type has the lowest prediction error among the four EMA dimensions. Furthermore, these EMA dimensions are predicted more accurately for the corresponding consonant types than for the others.

We propose that the lower error we observe for critical articulators may be due to there being less variability overall in their patterns of movement (i.e. across multiple instances) compared to those of non-critical articulators. Hence the distributions over EMA dimensions that correspond to critical articulators in the trained HMMs are likely to be "tighter", with lower variance. In addition, the movements of critical articulators in the test set are equally likely to be more constrained relative to those of non-critical articulators.



(RMSE)	Labiodental	Alveolar	Palatal	Velar
LL_y	<u>0.283</u>	0.330	0.283	0.387
T1_y	0.342	<u>0.289</u>	0.333	0.369
T2_y	0.321	0.338	<u>0.162</u>	0.260
T3_y	0.352	0.364	0.227	<u>0.128</u>

Figure 1: Illustrations for (a) the placement of the six EMA sensors used in our experiments and (b) the place of articulation and the normalized RMS error of four EMA dimensions for varying consonant types. We have underlined the EMA dimension which has the lowest prediction error among the four dimensions for each consonant type.

In order to test this explanation, we calculated the *normalized average standard deviation* of the trained distributions for each EMA dimension and phone type, using fully context-dependent models and natural state durations. When using fully context-dependent models together with decision-tree-based model clustering, it is not straightforward to select the distributions over EMA features for each phone directly. Instead, we used the test sentences to provide a state sequence using Viterbi alignment. First, we identified the states corresponding to a given phone type in the generated sequences for the test sentences. Next, we calculated the sum of the standard deviations in the corresponding Gaussian distributions over the static EMA features for all the frames in the states corresponding to a given phone type. We could then calculate their average by dividing by their total number of frames. Finally, this average standard deviation was normalized by the global standard deviation of each static EMA dimension separately.

The results for the y-coordinates of the LL, T1, T2 and T3 sensors are shown in Fig. 3. In this figure we see similar patterns to those found in Fig. 2. For example, the average standard deviation for the T3_y dimension is lower than the other three for the "Velar" phone type, and is also lower for this phone type than for any other type. This same pattern is also found in the y-coordinate of T2 for "Palatal", T1 for "Alveolar", and so on. This provides good support for our proposed explanation.

4. Conclusions

We have presented an HMM-based method to predict articulatory movements from text. Articulatory movements are generated from HMMs trained on articulatory features, using an MLPG algorithm. Our experiments have shown that using rich context features to define the model set reduces prediction error significantly. When fully context-dependent models are used,

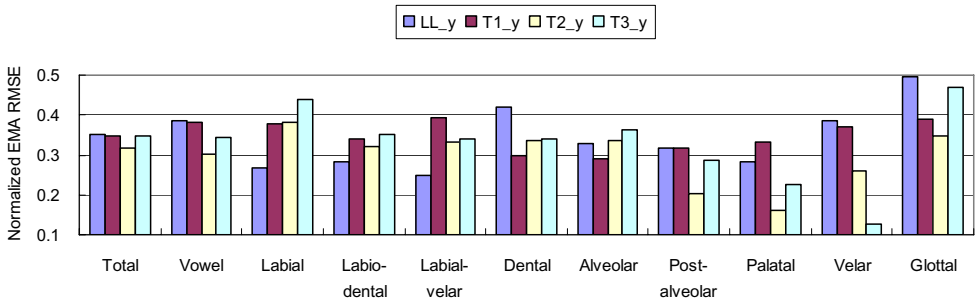


Figure 2: Normalized RMS error for the y-coordinates of the LL, T1, T2 and T3 sensors for different phone types using fully context-dependent models and natural state durations. RMS errors have been normalized by dividing by the global standard deviation for each EMA sensor coordinate separately.

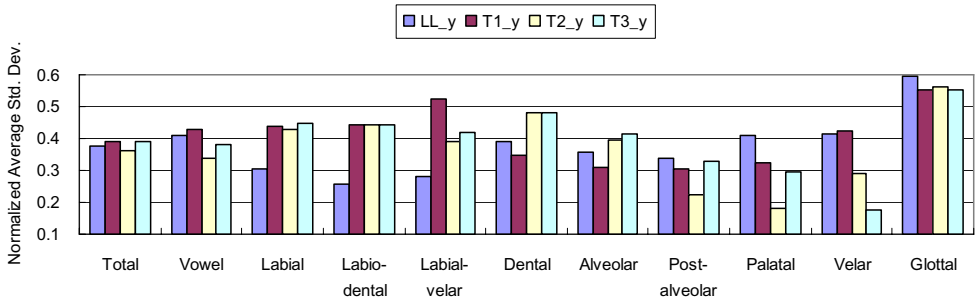


Figure 3: Normalized average standard deviation in the trained HMMs for the y-coordinates of the LL, T1, T2 and T3 sensors for different phone types. Fully context-dependent models and natural state durations have been used.

an RMS error of 1.945mm was achieved using predicted state durations, which decreased to 0.872mm when natural state durations were used. Furthermore, we have found that critical articulators have tighter distributions in the trained HMMs and consequently lower prediction error than non-critical ones for a range of phone types. Considering the different roles of articulators during pronunciation, to further improve the performance of movement prediction for critical articulators, will be the focus of our future work.

5. Acknowledgements

We thank Phil Hoole of Ludwig-Maximilian University, Munich for his great effort in helping record the EMA data.

6. References

[1] P. W. Schönle, K. Gräbe, P. Wenig, J. Höhne, J. Schrader, and B. Conrad, "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, 1987.

[2] K. Tokuda, H. Zen, and A. W. Black, "HMM-based approach to multilingual speech synthesis," in *Text to speech synthesis: New paradigms and advances*, S. Narayanan and A. Alwan, Eds. Prentice Hall, 2004.

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

[4] C. S. Blackburn and S. Young, "A self-learning predictive model of articulator movements during speech production," *Journal of the Acoustical Society of America*, vol. 107, no. 3, pp. 1659–1670, 2000.

[5] 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.

[6] L. Zhang and S. Renals, "Acoustic-articulatory modelling with the trajectory HMM," *IEEE Signal Processing Letters*, vol. 15, pp. 245–248, 2008.

[7] S. Hiroya and T. Mochida, "Multi-speaker articulatory trajectory formation based on speaker-independent articulatory HMMs," *Speech Communication*, vol. 48, no. 12, pp. 1677–1690, 2006.

[8] G. Papcun, J. Hochberg, T. R. Thomas, F. Laroche, J. Zachs, and S. Levy, "Inferring articulation and recognising gestures from acoustics with a neural network trained on X-ray microbeam data," *J. Acoust. Soc. Am.*, vol. 92, no. 2, pp. 688–700, August 1992.

[9] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Japan (E)*, vol. 21, no. 2, pp. 79–86, 2000.

[10] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration modeling in HMM-based speech synthesis system," in *ICSLP*, vol. 2, 1998, pp. 29–32.

[11] S. Fitt and S. Isard, "Synthesis of regional English using a keyword lexicon," in *Eurospeech*, vol. 2, 1999, pp. 823–826.

[12] P. Taylor, A. W. Black, and R. Caley, "The architecture of the Festival speech synthesis system," in *3rd ESCA Workshop in Speech Synthesis*, 1998, pp. 147–151.

[13] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK version 3.2)*. Cambridge University Engineering Department, 2002.

[14] 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 *6th ISCA Workshop on Speech Synthesis*, 2007, pp. 294–299.