An Analysis of HMM-based Prediction of Articulatory Movements

Citation for published version:
Ling, Z-H, Richmond, K & Yamagishi, J 2010, 'An Analysis of HMM-based Prediction of Articulatory Movements' Speech Communication, vol 52, no. 10, pp. 834-846., 10.1016/j.specom.2010.06.006

Digital Object Identifier (DOI):
10.1016/j.specom.2010.06.006

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Preprint (usually an early version)

Published In:
Speech Communication

An Analysis of HMM-based Prediction of Articulatory Movements

Zhen-Hua Ling[a,*], Korin Richmond[b], Junichi Yamagishi[b]

[a]*iFLYTEK Speech Lab, University of Science and Technology of China, Hefei, Anhui, 230027, P.R.China*
[b]*The Centre for Speech Technology Research (CSTR), University of Edinburgh, Edinburgh, EH8 9LW, United Kingdom*

Abstract

This paper presents an investigation into predicting the movement of a speaker's mouth from text input using hidden Markov models (HMM). A corpus of human articulatory movements, recorded by electromagnetic articulography (EMA), is used to train HMMs. To predict articulatory movements for input text, a suitable model sequence is selected and a maximum-likelihood parameter generation (MLPG) algorithm is used to generate output articulatory trajectories. Unified acoustic-articulatory HMMs are introduced to integrate acoustic features when an acoustic signal is also provided with the input text. Several aspects of this method are analyzed in this paper, including the effectiveness of context-dependent modeling, the role of supplementary acoustic input, and the appropriateness of certain model structures for the unified acoustic-articulatory models. When text is the sole input, we find that fully context-dependent models significantly outperform monophone and quinphone models, achieving an average root mean square (RMS) error of 1.945mm and an average correlation coefficient of 0.600. When both text and acoustic features are given as input to the system, the difference between the performance of quinphone models and fully-context dependent models is no longer significant. The best performance overall is achieved using unified acoustic-articulatory quinphone HMMs with separate clustering of acoustic and articulatory model parameters, a synchronous-state sequence, and a dependent-feature model structure, with an RMS error of 0.900mm and a correlation coefficient of 0.855 on average. Finally, we also apply the same quinphone HMMs to the acoustic-articulatory, or inversion, mapping problem, where only acoustic input is available. An average root mean square (RMS) error of 1.076mm and an average correlation coefficient of 0.812 are achieved. Taken together, our results demonstrate how text and acoustic inputs both contribute to the prediction of articulatory movements in the method used.

Keywords:
Hidden Markov model, articulatory features, parameter generation

1. Introduction

In human speech production it is the movements of articulators, such as the tongue, jaw, lips and velum, that generate and shape the acoustic signal. Hence, articulatory features which may be recorded by human articulography (Schönle et al., 1987; Kiritani, 1986; Baer et al., 1987), provide an effective and important description of speech as an alternative to an acoustic representation. Similar to the generation of an acoustic representation of speech in standard 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 to learn correct pronunciation, or for the analysis of pronunciation defects; it could be employed in an animated talking-head system; or it could feature in an articulation-based speech synthesis system.

This paper presents an approach to predicting articulatory movements from text that adopts a similar framework to hidden Markov model (HMM) based parametric speech synthesis (Tokuda et al., 2004). When text is the only input from which to predict articulatory movements, HMMs are trained using the recorded articulatory features and linguistic context labeling of a speech corpus recorded with a human articulography technique, here electromagnetic articulography (EMA). When acoustic features are provided to supplement the text, it is necessary to train unified acoustic-articulatory HMMs to capture the relationship between the acoustic and articulatory features. To perform synthesis, optimal trajectories of articulatory movements are generated from the trained models using a maximum-likelihood criterion with dynamic feature constraints (Tokuda et al., 2000).

Related research on predicting or estimating articulatory movements has previously been presented in the literature, and we consider here a few of the most relevant examples. In Blackburn and Young (2000), articulator movements were predicted from time-aligned phone strings using Gaussian distribution models at phone midpoints together with an explicit coarticulation model. In contrast, we use an HMM here to achieve temporal modeling of articulatory movements. In Tamura et al. (1999), lip shapes (derived from video) were predicted alongside synchronous acoustic speech synthesis parameters from textual input using an HMM-based parameter generation method. Here, we predict not only lip movements, but also movements of articulators inside the mouth, with

[*]Corresponding author. Tel: +86-551-5331851 ext. 8050, fax: +86-551-5331801.

Email addresses: zhling@ustc.edu (Zhen-Hua Ling), korin@cstr.ed.ac.uk (Korin Richmond), jyamagis@inf.ed.ac.uk (Junichi Yamagishi)

EMA providing the articulatory training and testing data. In addition, we investigate optionally using an acoustic speech signal to supplement the input text in order to guide prediction of articulatory movements.

The focus of Toda et al. (2008); Richmond (2007, 2009); Hiroya and Honda (2004) and Zhang and Renals (2008) was the *inversion mapping* (also known as the acoustic-articulatory mapping), where the aim is to estimate the articulatory movements underlying a given acoustic speech signal. In Toda et al. (2008), a Gaussian mixture model for the joint distribution of acoustic and articulatory features was adopted to achieve the mapping from acoustic features to articulatory movements. In Richmond (2007, 2009), an artificial neural network (ANN) and MLPG algorithm were combined to form a statistical trajectory model to estimate articulation from an acoustic speech signal. The work described in Hiroya and Honda (2004) and Zhang and Renals (2008) was based on the HMM, which is similar to the approach presented in this paper. However, since their focus was on the inversion mapping, they were limited to using only very simple context information to define the set of HMMs. Our aim here, in contrast, is primarily to predict articulatory movements from *text*. Therefore, we can readily use much more fine-grained linguistic features to define our model set, as is common in acoustic speech synthesis, since we do not face the problem of a decoding search with a huge model set.

Finally, a similar HMM-based approach was also used in Hiroya and Mochida (2006). Their aim was to use speaker adaptive training (SAT) to train a *speaker-independent* model to predict articulatory movements from text. The work presented here has three key differences. First, unlike Hiroya and Mochida (2006), we evaluate using a large set of models defined in terms of a fine-grained set of linguistic context features. This can theoretically improve accuracy by modeling the characteristics of articulatory movements in differing environments. Second, in Hiroya and Mochida (2006), the state durations for the articulatory movement generation from HMMs were not predicted, but derived from the measured articulatory data by Viterbi alignment. In contrast, we use a statistical model to predict state durations from text and the influence of state duration prediction is studied in our experiments for the articulatory HMMs using different forms of context information. Third, we augment our system to model the dependence of the acoustic features on the associated articulatory features. This provides a unified acoustic-articulatory model which may be trained to predict articulatory features that are synchronized with an input acoustic signal.

In summary, several important aspects of HMM-based prediction of articulatory movements are studied in this paper:

1) **The effectiveness of context-dependent modeling.** As mentioned above, fine-grained linguistic features can be used here to define our model set because the text from which these are derived is given. It is necessary to evaluate the effect of introducing rich context features into the model definition, both when text is the only input and when acoustic input is also available.

2) **The role of supplementary acoustic input.** Due to the mechanism of speech production the acoustic signal is strongly correlated with articulatory movements. In this paper we analyze how acoustic input complements text in the prediction of articulatory movements. We compare prediction performance using: a) text input alone; b) audio input alone (i.e. the inversion mapping); and c) both text and audio input together.

3) **Appropriate model structures for unified acoustic-articulatory modeling.** In previous work, we have explored various model structures for an articulatorily controllable HMM-based speech synthesis system (Ling et al., 2009). However, the purpose of the current paper is to predict articulatory movements, and not to generate acoustic synthesis parameters as in our previous work. Hence, similar investigations into model structure are conducted in this paper.

In the remainder of the paper, Section 2 describes the HMM-based articulatory-movement prediction method in detail, Section 3 presents the results of our experiments, and Section 4 gives the conclusions we draw on the basis of these.

2. Method

2.1. Articulatory Movement Prediction from Text

The framework of the HMM-based method used to predict articulatory movements is shown in Fig. 1. To begin with, we consider the case of predicting articulation from text alone. To construct the training data set, articulatory movements of dimensionality D_X are recorded by human articulography. During training, a set of context-dependent HMMs λ are estimated to maximize the likelihood function $P(\boldsymbol{X}|\lambda)$. Here $\boldsymbol{X} = [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, ..., \boldsymbol{x}_N^\top]^\top$ is the observed articulatory feature sequence, $(\cdot)^\top$ denotes the matrix transpose and N is the length of the sequence. The observation feature vector $\boldsymbol{x}_t \in \mathcal{R}^{3D_X}$ for each frame consists of static articulatory parameters $\boldsymbol{x}_{S_t} \in \mathcal{R}^{D_X}$ and their velocity and acceleration components as

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

where

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

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

After initial context-dependent HMM training, a decision tree is trained using the minimum description length (MDL) criterion (Shinoda and Watanabe, 2000) to cluster the probability density functions of all HMM states. This is to mitigate problems of data sparsity and to formulate estimates for the parameters of models whose context description is missing in the training set. Next, a state alignment is derived using the trained HMMs. This is then used to train context-dependent state duration probabilities (Yoshimura et al., 1998) for state duration prediction.

To generate articulatory movements, the results of front-end linguistic analysis on the input text are used to determine the sentence HMM by consulting the clustering decision tree built during training. The MLPG algorithm (Tokuda et al., 2000)

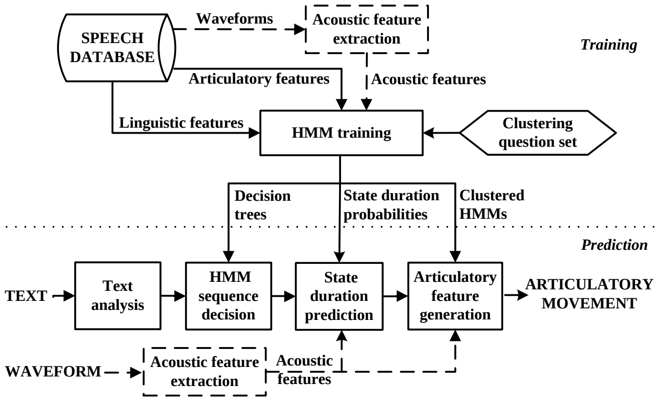


Fig. 1. Flowchart of the HMM-based text-to-articulatory movement prediction method. The dashed lines are used for the condition that acoustic waveforms are input with the text to guide prediction.

is then applied to generate the optimal articulatory trajectories using dynamic features, such that

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

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

where $\boldsymbol{X} = \boldsymbol{W}_X \boldsymbol{X}_S$; $\boldsymbol{X}_S = [\boldsymbol{x}_{S_1}^\top, \boldsymbol{x}_{S_2}^\top, ..., \boldsymbol{x}_{S_N}^\top]^\top$ is the static articulatory feature sequence; $\boldsymbol{W}_X \in \mathcal{R}^{3ND_X \times ND_X}$ 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

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

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

where the optimal state sequence

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

is determined from the trained state duration probabilities (Yoshimura et al., 1998) and $\boldsymbol{X}_S^*$ is calculated by setting $\partial \log P(\boldsymbol{W}_X \boldsymbol{X}_S|\lambda, \boldsymbol{q}^*)/\partial \boldsymbol{X}_S = \boldsymbol{0}$, as introduced in Tokuda et al. (2000).

2.2. Articulatory Movement Prediction with Acoustic Inputs

When acoustic waveforms are available with the input text, the predicted articulatory movements are required to be synchronized with the reference acoustic signal. A unified acoustic-articulatory model is necessary to represent the relationship between these two parameter streams. During training, HMMs λ for the combined acoustic and articulatory features are estimated to maximize the likelihood function of their joint distribution $P(\boldsymbol{X}, \boldsymbol{Y}|\lambda)$, where $\boldsymbol{X}$ and $\boldsymbol{Y} = [\boldsymbol{y}_1^\top, \boldsymbol{y}_2^\top, ..., \boldsymbol{y}_N^\top]^\top$ denote the parallel articulatory and acoustic observation sequences of length N respectively. At each frame the acoustic feature vector $\boldsymbol{y}_t \in \mathcal{R}^{3D_Y}$ is similarly composed of static features $\boldsymbol{y}_{S_t} \in \mathcal{R}^{D_Y}$ and their velocity and acceleration components as

$$\boldsymbol{y}_t = [\boldsymbol{y}_{S_t}^\top, \Delta \boldsymbol{y}_{S_t}^\top, \Delta^2 \boldsymbol{y}_{S_t}^\top]^\top \qquad (9)$$

where D_Y is the dimensionality of the static acoustic features.

Various structures may be adopted to model the joint distribution $P(\boldsymbol{X}, \boldsymbol{Y}|\lambda)$. In previous work on an articulatorily controllable HMM-based speech synthesis system (Ling et al., 2009), we investigated three aspects of model structure:

1) **Model clustering.** Model-clustering, using decision trees, is an important step in the training of context-dependent HMMs. We can choose either to cluster the acoustic model components and articulatory model components independently ("*separate clustering*") or to build a shared decision tree to cluster the models for both feature types simultaneously ("*shared clustering*").
2) **Cross-stream synchrony.** The acoustic and articulatory feature sequences can be assumed to be generated from different state sequences ("*asynchronous-state*") or from a single state sequence ("*synchronous-state*").
3) **Cross-stream dependency.** The generation of acoustic features can be assumed to depend only upon the current state ("*independent-feature*") or also depend upon the current articulatory features ("*dependent-feature*").

In this paper, the synchronous-state model structure is assumed. However, the other two aspects pertaining to model structure are investigated in our experiments below.

We model the dependency between the acoustic and articulatory features using a piecewise linear transform within the HMM states (Ling et al., 2009). Mathematically, we can write the joint distribution as

$$P(\boldsymbol{X}, \boldsymbol{Y}|\lambda) = \sum_{\forall \boldsymbol{q}} P(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{q}|\lambda) \qquad (10)$$

$$= \sum_{\forall \boldsymbol{q}} \pi_{q_0} \prod_{t=1}^{N} a_{q_{t-1}q_t} b(\boldsymbol{x}_t, \boldsymbol{y}_t) \qquad (11)$$

$$b_j(\boldsymbol{x}_t, \boldsymbol{y}_t) = b_j(\boldsymbol{x}_t) b_j(\boldsymbol{y}_t|\boldsymbol{x}_t) \qquad (12)$$

$$b_j(\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_{X_j}, \boldsymbol{\Sigma}_{X_j}) \qquad (13)$$

$$b_j(\boldsymbol{y}_t|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{A}_j \boldsymbol{x}_t + \boldsymbol{\mu}_{Y_j}, \boldsymbol{\Sigma}_{Y_j}) \qquad (14)$$

where $\boldsymbol{q} = \{q_1, q_2, ..., q_N\}$ denotes the state sequence shared by the two feature streams; π_j and a_{ij} represent initial state probability and state transition probability respectively; $b_j(\cdot)$ denotes the state observation probability density function (PDF) for state j; $\mathcal{N}(; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$; and $\boldsymbol{A}_j \in \mathcal{R}^{3D_Y \times 3D_X}$ is the linear transform matrix for state j to model the dependency of acoustic features on articulatory features. As the transform matrix is state-dependent, a piecewise linear transform is achieved globally. An Expectation-Maximization (EM) algorithm can be used to estimate the model parameters; the re-estimation formulae may be found in Ling et al. (2009).

To predict articulatory movements from text with supplementary acoustic inputs the same maximum-likelihood criterion as in Section 2.1 is followed, though (4) is modified so that

$$\boldsymbol{X}_S^* = \arg\max_{\boldsymbol{X}_S} P(\boldsymbol{W}_X \boldsymbol{X}_S | \lambda, \boldsymbol{Y}) \tag{15}$$

$$= \arg\max_{\boldsymbol{X}_S} \sum_{\forall \boldsymbol{q}} P(\boldsymbol{W}_X \boldsymbol{X}_S, \boldsymbol{q} | \lambda, \boldsymbol{Y}). \tag{16}$$

Again, we simplify the optimization for (16) by considering only the optimal state sequence. Therefore, we have

$$[\boldsymbol{X}_S^*, \boldsymbol{q}^*] \approx \arg\max_{\boldsymbol{X}_S, \boldsymbol{q}} P(\boldsymbol{W}_X \boldsymbol{X}_S, \boldsymbol{q} | \lambda, \boldsymbol{Y}). \tag{17}$$

An iterative update method that alternately optimizes the state sequence is adopted here to solve (17). Each iteration consists of two steps.

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

$$\boldsymbol{X}_{Si}^* = \arg\max_{\boldsymbol{X}_S} P(\boldsymbol{W}_X \boldsymbol{X}_S | \lambda, \boldsymbol{q}_{i-1}, \boldsymbol{Y}) \tag{18}$$

$$= \arg\max_{\boldsymbol{X}_S} P(\boldsymbol{W}_X \boldsymbol{X}_S, \boldsymbol{Y} | \lambda, \boldsymbol{q}_{i-1}) \tag{19}$$

where $i \in \{1, 2, ...\}$ denotes the i-th iteration and $\boldsymbol{q}_0$ is calculated by Viterbi alignment on $\boldsymbol{Y}$ using an isolated acoustic model. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are assumed to be independent given the state sequence, (18) can be solved using the conventional MLPG algorithm (Tokuda et al., 2000), and $\boldsymbol{Y}$ cannot affect the prediction of $\boldsymbol{X}_S$ at all (other than through the shared state-sequence in 2) below). Once the dependent-feature model structure is adopted, as in (14), the joint distribution in (19) can be rewritten as

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

where

$$\boldsymbol{U}_X^{-1} = \mathrm{diag}[\boldsymbol{\Sigma}_{X_{q_1}}^{-1}, \boldsymbol{\Sigma}_{X_{q_2}}^{-1}, ..., \boldsymbol{\Sigma}_{X_{q_N}}^{-1}] \tag{21}$$

$$\boldsymbol{M}_X = [\boldsymbol{\mu}_{X_{q_1}}^\top, \boldsymbol{\mu}_{X_{q_2}}^\top, ..., \boldsymbol{\mu}_{X_{q_N}}^\top]^\top \tag{22}$$

$$\boldsymbol{U}_Y^{-1} = \mathrm{diag}[\boldsymbol{\Sigma}_{Y_{q_1}}^{-1}, \boldsymbol{\Sigma}_{Y_{q_2}}^{-1}, ..., \boldsymbol{\Sigma}_{Y_{q_N}}^{-1}] \tag{23}$$

$$\boldsymbol{M}_Y = [\boldsymbol{\mu}_{Y_{q_1}}^\top, \boldsymbol{\mu}_{Y_{q_2}}^\top, ..., \boldsymbol{\mu}_{Y_{q_N}}^\top]^\top \tag{24}$$

$$\boldsymbol{A} = \mathrm{diag}[\boldsymbol{A}_{q_1}, \boldsymbol{A}_{q_2}, ..., \boldsymbol{A}_{q_N}] \tag{25}$$

and K is a constant value. Therefore, by setting $\partial P(\boldsymbol{W}_X \boldsymbol{X}_S, \boldsymbol{Y} | \lambda, \boldsymbol{q}_{i-1}) / \partial \boldsymbol{X}_S = \boldsymbol{0}$, we have

$$\begin{aligned}\boldsymbol{X}_{Si}^* = {} & (\boldsymbol{W}_X^\top (\boldsymbol{U}_X^{-1} + \boldsymbol{A}^\top \boldsymbol{U}_Y^{-1} \boldsymbol{A}) \boldsymbol{W}_X)^{-1} \\ & \cdot \boldsymbol{W}_X^\top (\boldsymbol{U}_X^{-1} \boldsymbol{M}_X + \boldsymbol{A}^\top \boldsymbol{U}_Y^{-1} (\boldsymbol{Y} - \boldsymbol{M}_Y)). \end{aligned} \tag{26}$$

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

$$\boldsymbol{q}_i^* = \arg\max_{\boldsymbol{q}} P(\boldsymbol{q} | \lambda, \boldsymbol{W}_X \boldsymbol{X}_{Si}^*, \boldsymbol{Y}). \tag{27}$$

This can be solved with a Viterbi alignment using the trained HMMs on the feature sequence pair $(\boldsymbol{W}_X \boldsymbol{X}_{Si}^*, \boldsymbol{Y})$. The updated optimal state sequence $\boldsymbol{q}_i^*$ is then used to generate articulatory features according to (21)-(26) in the next iteration.

3. Experiments

3.1. Database

In our experiments, we have used a data set comprised of articulatory movements recorded concurrently with the corresponding acoustic waveforms. A Carstens AG500 electromagnetic articulograph was used to record 1,263 phonetically balanced sentences, which were read by a male British English speaker. The waveforms were in 16kHz PCM format with 16 bit precision. Six EMA sensors were used, located at the *tongue dorsum* (T3), *tongue body* (T2), *tongue tip* (T1), *lower incisor* (LI), *upper lip* (UL), and *lower lip* (LL) of the speaker. This is illustrated in Fig. 8(a). 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 sensors were used in our experiments, making a total of 12 static articulatory features at each sample instant.

3.2. System Construction

To create context-dependent HMMs, we first labeled the database using tools from Unilex (Fitt and Isard, 1999) and Festival (Taylor et al., 1998). Phone boundaries were determined automatically using HTK (Young et al., 2002). 1,200 sentences were selected for training and the remaining 63 sentences were used as a test set. A 5-state, left-to-right model structure with no skips was adopted to train phone HMMs. A single Gaussian distribution with diagonal covariance was used for each HMM state. Our training and prediction implementation was based upon the HTS toolkits (Zen et al., 2007). In addition to simple monophone models, two forms of context-dependent HMMs were trained and evaluated in our experiments:

1) Quinphone model. The context features for each model comprised the identity of the current phone, together with those of the preceding and follow two neighbouring phones.
2) Fully context-dependent model. In addition to the phone identities used in the quinphone models, a broad set of linguistic and prosodic features were adopted, similar to those used in HMM-based TTS systems (Tokuda et al., 2004). A full list of the specific context features used is given in Table 1.

Table 1
The linguistic context features used for fully context-dependent model training.

the identity of the current and neighbouring 4 phones (phone before the previous, previous, current, next, phone after the next);
the position of the current phone in the current syllable;
the number of phones in the {previous, current, next} syllable;
whether the {previous, current, next} syllable is stressed or not;
whether the {previous, current, next} syllable is accented or not;
the position of the current syllable in the current word;
the number of syllables in the {previous, current, next} word;
the number of {stressed, accented} syllables in the current {word, phrase};
the distance between the current syllable and the neighbouring {stressed, accented} syllable;
the part-of-speech of the {previous, current, next} word;
the position of the current {syllable, word} in the current phrase;
the number of {syllables, words} in the {previous, current, next} phrase;
the number of content words in the current phrase;
the distance between the current word and the neighbouring content word;
the boundary tone of the current phrase;
the position of the current phrase in the utterance;
the number of {syllables, words, phrases} in the utterance.

3.3. *Articulatory Movement Prediction from Text*

In this experiment, only articulatory features and linguistic context labels were used for training, and no acoustic signals were used during articulatory movement prediction (as in Section 2.1). Three systems were trained, one with monophone models, one with quinphone models, and one with fully context-dependent models. RMS error calculated for the 63 test sentences (with silence segments excluded) and averaged over all 12 EMA features was used as an objective measure to evaluate the accuracy of articulatory movement prediction. To facilitate the calculation of the error for each utterance, the state duration prediction in (8) was solved under the constraint of setting the total length of generated articulatory frames to be the same as the duration of the natural utterance (Yoshimura et al., 1998).

Results for the three systems are shown in Fig. 2. A *t*-test informs us that the differences among these three systems are significant ($p < 0.05$). From these results, we see the context-dependent modeling approach which is commonly used in HMM-based speech synthesis is also an effective method to predict articulatory movements from text. Compared with monophone models, using quinphone models improves the accuracy of articulatory feature prediction significantly, as it can account for the coarticulatory effects of nearby phones on the movement of articulators when producing a given phone.

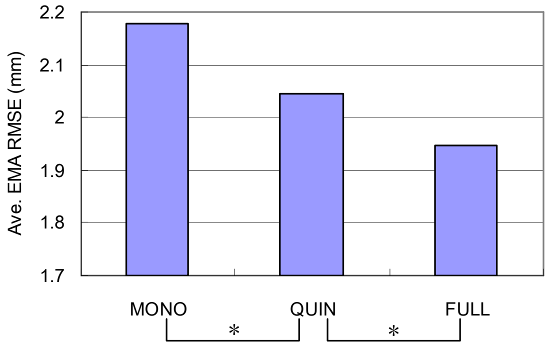


Fig. 2. RMS error of EMA features predicted from text using monophone (*MONO*), quinphone (*QUIN*), and fully context-dependent (*FULL*) models. “*” indicates the difference between two systems is significant.

The rich linguistic context features that were used in addition to the neighbouring phone identities (see Table 1) when training the fully context-dependent models are commonly believed to be correlated with the suprasegmental characteristics of speech, such as pitch and duration. However, Fig. 2 shows that the fully context-dependent models are also significantly better than quinphone models for the prediction of articulatory movements. We conducted a further experiment to explore the reasons for this difference. Similar to Hiroya and Mochida (2006), the result of Viterbi alignment on the natural articulatory recordings by the monophone, quinphone, and fully context-dependent models was adopted to replace the duration prediction in (8) when generating the articulatory movements using the three models respectively. The average RMS error of predicted articulatory movements for the 63 test sentences was then calculated, as shown in Fig. 3. Compared with the results in Fig. 2, we see that RMS error is greatly reduced for all three systems when natural state durations are provided. Furthermore, the difference between the quinphone models and the fully context-dependent models is not significant any more. This implies the superiority of the fully context-dependent models over the quinphone models in Fig. 2 lies in better duration prediction. This is reasonable since the fully context-dependent models take context features related to prosody into account to train the duration probabilities.

Although we have used a different data set here, which inhibits direct comparison, we nevertheless note these RMSE results for the same task of predicting articulation from text compare very well with other methods and results previously reported, such as Blackburn and Young (2000) and Hiroya and Mochida (2006), especially when context-dependent models are used.

3.4. *Articulatory Movement Prediction with Acoustic Inputs*

3.4.1. *Without Cross-stream Dependency Modeling*

Unified acoustic-articulatory HMMs were trained to predict articulatory movements, using acoustic features as input to supplement the text. Frequency-warped LSFs of order 40 plus an extra gain dimension were derived with a 5ms frame shift from the spectral envelope provided by STRAIGHT (Kawahara et al., 1999) analysis on the acoustic waveforms. These spectral parameters and the logarithmized F0 of each frame were used as

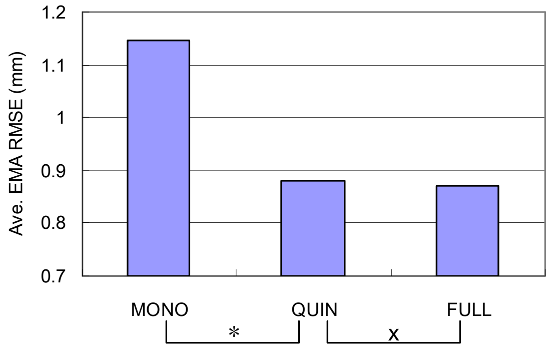


Fig. 3. RMS error of EMA features predicted from text using monophone (*MONO*), quinphone (*QUIN*), and fully context-dependent (*FULL*) models when the natural state segmentations are given. "*" indicates the difference between two systems is significant and "x" indicates the difference is insignificant.

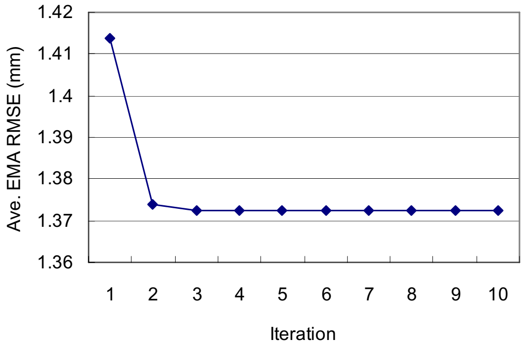


Fig. 4. RMS error of EMA features predicted from text with acoustic inputs for monophone model ("*MONO*"). The x-axis refers to the number of iterations in the articulatory feature generation.

two separate acoustic feature streams and were combined with the articulatory features to train the unified acoustic-articulatory HMMs. A multi-space probability distribution (MSD) (Tokuda et al., 1999) was used to model the F0 stream.

First, five models were compared to evaluate the effectiveness of context-dependent modeling and different model clustering strategies:

- Monophone models with independent-feature model structure (*MONO*);
- Quinphone models with separate clustering and independent-feature model structure (*QUIN*);
- Quinphone models with shared clustering and independent-feature model structure (*QUIN-SC*);
- Fully context-dependent models with separate clustering and independent-feature model structure (*FULL*);
- Fully context-dependent models with shared clustering and independent-feature model structure (*FULL-SC*).

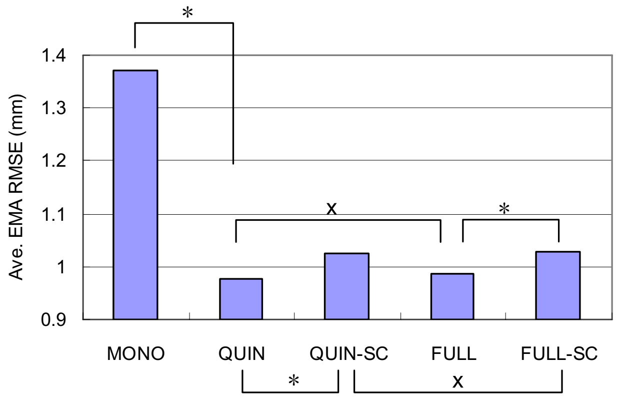


Fig. 5. RMS error of EMA features predicted from text with acoustic inputs for monophone model ("*MONO*"), quinphone model with separate clustering ("*QUIN*"), quinphone model with shared clustering ("*QUIN-SC*"), fully context-dependent model with separate clustering ("*FULL*"), and fully context-dependent model with shared clustering ("*FULL-SC*"). All systems adopt independent-feature model structures. "*" indicates the difference between two systems is significant and "x" indicates the difference is insignificant.

A fully context-dependent acoustic model was trained using the same acoustic features as for the unified acoustic-articulatory HMMs to get the initial state sequence $\boldsymbol{q}_0$ in (19) by Viterbi force-alignment on the acoustic inputs for all the five systems. Then iterative optimization on the predicted articulatory movements was conducted. The results of the iterative optimization using monophone models are shown in Fig. 4. Although the iterative updates do not guarantee to find the global optimum for (17), depending on the calculation of initial state sequence $\boldsymbol{q}_0$, we see that prediction error decreases as a result of optimizing the state sequence, with convergence after approximately 3 iterations. Thus, the number of iterations was set to 3 for all systems in the following experiments.

The performance of systems *MONO*, *QUIN*, *QUIN-SC*, *FULL*, and *FULL-SC* is compared in Fig. 5. A *t*-test at 95% confidence level was applied to analyze the significance of the difference between two systems. Comparing Fig. 2 with Fig. 5, we note that the supplementary audio features reduce the prediction error significantly. Table 2 shows RMS error for phone duration prediction for system *FULL* when only text input is used and system *QUIN* when both text and audio inputs are available. The reference phone durations are calculated by Viterbi alignment on the natural articulatory movements using corresponding models. Comparing the second row with the first row in this table, we can see that the phone durations obtained by Viterbi alignment on input acoustic features are far more accurate than those predicted from text using fully context-dependent distributions. This is due to the synchronous relationship between acoustic and articulatory features incorporated in the unified HMMs. We also see that the error in predicted durations for system *QUIN* with text and audio input are reduced further when the optimization of the state sequence is conducted iteratively. The advantage of context-dependent modeling is reaffirmed by comparing *MONO* (1.373mm) with *QUIN* (0.978mm) and *FULL* (0.987mm) in Fig. 5. However,

Table 2
RMS error of predicted phone duration for system *FULL* when only text input is used and system *QUIN* when both text and audio inputs are available. For *Text & Audio* input, the values in the brackets indicate the numbers of iterations in the articulatory feature generation and the phone durations of the generated articulatory movements after the 1-st iteration is determined by the Viterbi alignment on the input acoustic features.

Input	System	Phone duration RMSE (ms)		
		Consonants	Vowels	All
Text	*FULL*	35.19	40.49	37.38
Text & Audio	*QUIN (1)*	16.66	14.32	15.77
	QUIN (3)	12.22	8.80	10.99

in contrast to the results in Fig. 2, we find there are no significant differences between the systems using quinphone models and fully context-dependent models irrespective of whether separate or shared clustering is applied. In Section 3.3, we concluded that it is the better duration prediction that leads to the superiority of the fully context-dependent model over the quinphone model when only text inputs are available. However, when acoustic inputs are given and a synchronous-state model structure is used for the unified acoustic-articulatory HMMs, the state durations are not predicted using trained duration probabilities, but are decided by Viterbi alignment according to (27). Therefore, it is reasonable that the fully context-dependent models cannot outperform the quinphone models here. This is consistent with the results shown in Fig. 3.

Finally, Fig. 5 also makes clear that separate clustering is significantly better than shared clustering when either quinphone models or fully context-dependent models are used. Table 3 lists the sizes of trained decision trees for the EMA and LSF model clustering for different systems when the same MDL criterion is followed. In this table, we see that performing clustering separately results in a larger decision tree for the articulatory features and a smaller decision tree for the acoustic features than when models for both these features are clustered jointly ("shared"). This confirms our previous observation (Ling et al., 2009) that articulatory features provide better discrimination in terms of pronunciation variation than acoustic features. Shared clustering can improve the model tying topology for the acoustic features, but impairs that for the articulatory features. Therefore, separate clustering should be adopted when predicting articulatory movements.

3.4.2. With Cross-stream Dependency Modeling

The effect of cross-stream dependency modeling is evaluated next. Two more systems were trained:

- Quinphone models with separate clustering and a dependent-feature model structure, where a single global transform matrix $\boldsymbol{A}_j$ (see (14)) was used (*QUIN-GLB*).
- Quinphone models with separate clustering and dependent-feature model structure where the transform matrix $\boldsymbol{A}_j$ was tied for each leaf node of the model clustering decision tree for the acoustic features (*QUIN-REG*).

Table 3
A comparison of the number of leaf nodes contained in model-clustering decision trees for EMA and LSF features. (see Fig. 5 for a key to the labels)

System	EMA	LSF
QUIN	5926	2159
QUIN-SC	3300	3300
FULL	6358	2265
FULL-SC	3548	3548

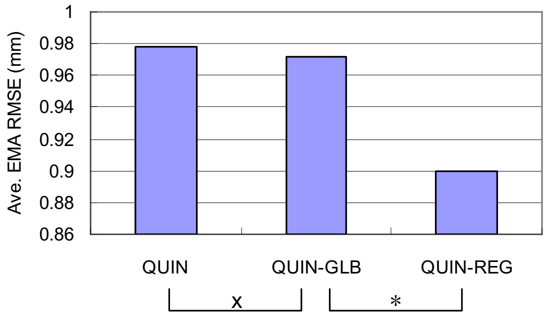


Fig. 6. RMS error of EMA features predicted from text with acoustic inputs for quinphone model (*QUIN*), quinphone model with global cross-stream dependency modeling (*QUIN-GLB*), and quinphone model with cross-stream dependency modeling using regression classes (*QUIN-REG*). All systems adopt the separate clustering model structure. "*" indicates the difference between two systems is significant and "x" indicates the difference is insignificant.

To train models *QUIN-GLB* and *QUIN-REG*, $\boldsymbol{A}_j$ was defined as a three-block matrix corresponding to static, velocity and acceleration components of the feature vector in order to reduce the number of parameters to be estimated. Only the dependency between the articulatory features and the spectral features was considered (i.e. any potential dependency between the articulatory features and the F0 stream was ignored). The results for systems *QUIN*, *QUIN-GLB*, and *QUIN-REG* are presented in Fig. 6. These results show that the addition of cross-stream dependency modeling does not reduce the prediction error if a single, global transform is applied. However, when $\boldsymbol{A}_j$ is set to be state-dependent using regression classes, the RMS error decreases from 0.978mm for system *QUIN* to 0.900mm for system *QUIN-REG*, which is statistically significant. This means using a piecewise linear transform is a more reasonable model for the dependency between LSFs and EMA movements than the global linear transform. This coincides with our previous study on integrating articulatory features into HMM-based speech synthesis (Ling et al., 2009).

Previously, it has been noted that certain articulators may be more key to the production of a given phone than others. Papcun et al. (1992) presented evidence for what they termed

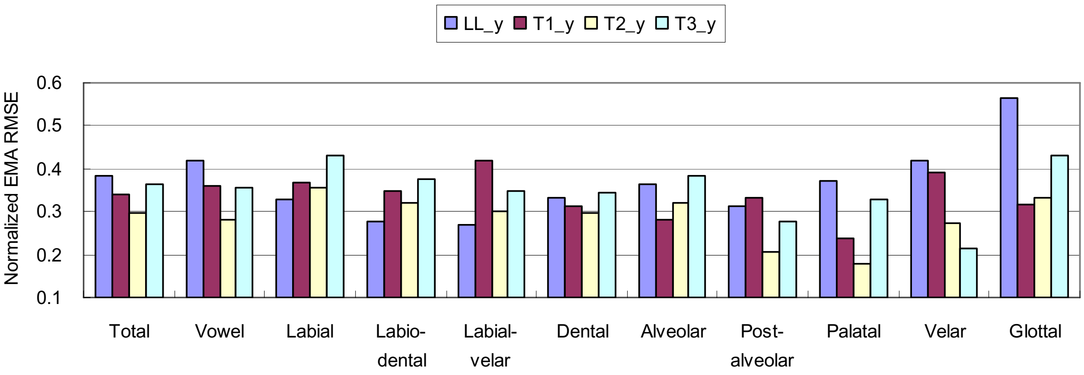


Fig. 7. Normalized RMS error for the y-coordinates of the LL, T1, T2 and T3 sensors for different phone types when both text and audio inputs are used in system *QUIN-REG*. RMS errors have been normalized by dividing by the global standard deviation for each EMA sensor coordinate separately.

Table 4
Definition of phone types. The phone symbols used here are in Unilex format (Fitt and Isard, 1999).

Vowel		*aa uu i ei @ ii ai a ou iy eir uw @r e oi oo ow o uh u i@ ur*
Consonant	Labial	*p b m m!*
	Labiodental	*f v*
	Labial-velar	*w*
	Dental	*dh th*
	Alveolar	*t d s z n n! l l! lw r*
	Postalveolar	*sh ch jh zh*
	Palatal	*y*
	Velar	*k g ng*
	Glottal	*h*

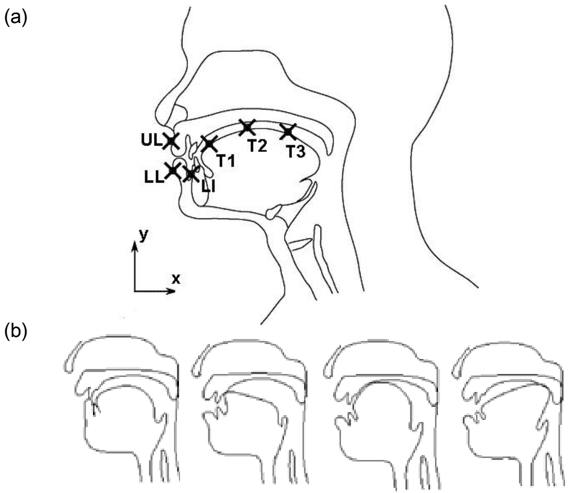
critical articulators. They demonstrated, for example, that the variance of trajectories of a point at the back of the tongue is significantly lower for phones for which this articulatory location is critical (i.e. for velar oral stops [k,g]) than for phones for which it is not (i.e. alveolar and bilabial stops [t,d,p,b]). In short, the implication is that the movements of articulators which 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. With this in mind, we have further analyzed RMS error for specific EMA sensor coordinates according to phone type. Fig. 7 shows normalized RMS error for the y-coordinates of the LL, T1, T2 and T3 sensors as predicted by system *QUIN-REG* according to the phone types listed in Table 4. Interestingly, we indeed find that the movements of *critical articulators* can 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 vocal tract. Fig. 7 shows that T2_y has the lowest prediction error (0.282) among the four EMA dimensions for type "Vowel", which is lower than the average T2_y prediction error of all phones (0.299).
- For consonants, the *critical articulators* depend upon a phone's place of articulation, e.g. the point where an obstruction occurs in the vocal tract. Fig. 8 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" correspond to the LL, T1, T2 and T3 sensors respectively. The clear pattern which emerges is that, for each consonant type, the *critical articulator* has the lowest prediction error among the four EMA dimensions. Furthermore, Fig. 7 shows that these EMA dimensions can be predicted more accurately for the corresponding consonant types than for the others.

3.5. Inversion Mapping

In this section, we compare the prediction of articulatory movements using concurrent text and audio inputs with the condition where only audio input is available, which is commonly known as the inversion mapping. An inversion mapping method using HMMs with cross-stream dependency modeling has been previously proposed (Hiroya and Honda, 2004), where the formula for articulatory movement prediction is the same as (26), with a state sequence $\boldsymbol{q}$ decoded from the acoustic feature stream using automatic speech recognition (ASR). In the experiment here, the iterative optimization approach introduced in Section 2.2 was applied to achieve the inversion mapping. The key difference is that, whereas in Section 2.2 the initial state sequence $\boldsymbol{q}_0$ was calculated by Viterbi alignment when both text and acoustic features are given, here ASR decoding becomes necessary because only acoustic inputs are available.

Two acoustic HMMs, a monophone model and a triphone model, were trained to provide a phone recognizer. To facilitate training, the acoustic features were the same as those used in Section 3.4, which were composed of spectral and F0 streams.



(RMSE)	Labiodental	Alveolar	Palatal	Velar
LL_y	0.278	0.362	0.371	0.420
T1_y	0.346	0.281	0.239	0.392
T2_y	0.320	0.321	0.181	0.274
T3_y	0.374	0.384	0.327	0.215

Fig. 8. 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. The normalized RMS error values are copied from Fig. 7. We have underlined the EMA dimension which has the lowest prediction error among the four dimensions for each consonant type.

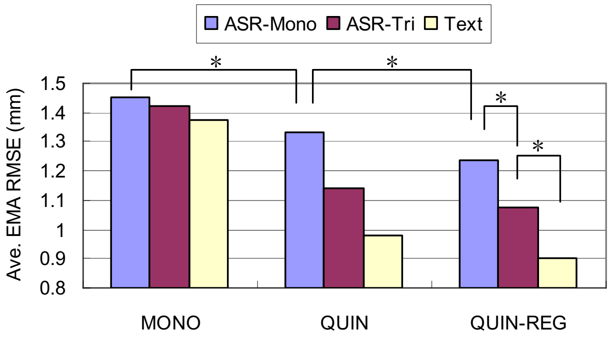


Fig. 9. RMS error for inversion mapping using monophone models (*MONO*), quinphone models (*QUIN*), and quinphone models with cross-stream dependency modeling (*QUIN-REG*). Labels “ASR-Mono” and “ASR-Tri” mean the initial state sequence is decoded using monophone and triphone acoustic models respectively; label “Text” indicates the initial state sequence is given by text analysis when text input is also available. “*” indicates the difference between two systems is significant.

Table 5
Phone recognition accuracy using monophone and triphone acoustic models.

ASR Model	*Monophone*	*Triphone*
Phone Accuracy (%)	59.12	71.49

As part of the training of the triphone models, decision tree-based model clustering was applied. The *HVite* tool in the HTS toolkit (Zen et al., 2007) was used to perform the decoding of acoustic features to give a phone sequence. A simple phone-loop grammar was used and no language model was applied. The phone recognition accuracy of the two models on the 63 test sentences is shown in Table 5.

Because ASR was performed using only a phone-loop grammar, and the recognition accuracy was not sufficiently high, the decoded phone sequence could not be reliably subjected to further analysis to extract further linguistic context features. Therefore, only the identities of neighbouring phones were available as context features, and the fully context-dependent models in Section 3.4 were not appropriate for the inversion mapping. Therefore, the systems *MONO*, *QUIN*, and *QUIN-REG* were compared in this experiment. The results of these three models are shown in Fig. 9.

From this figure, we see that

1) The performance of the phone recognizer plays an important role. The phone recognition accuracy obtained using triphone models is higher than with monophone models, which in turn results in lower RMS error for all *MONO*, *QUIN*, and *QUIN-REG* systems. Once both text and audio inputs are given, the correct phone sequences can be ascertained and the RMS error of predicted articulatory movements is lower than when using recognized phone sequences. This implies that the performance of this inversion mapping could be improved further should a better acoustic recognizer be available.
2) Because the identity of quinphone models depends on a greater number of phones, phone recognition errors can exert a greater adverse effect on the *QUIN* system in comparison to the *MONO* system. As shown in Fig. 9, the RMS error gap between the phone sequence from the triphone recognizer and the correct phone sequence is 0.049mm for system *MONO*, and 0.165/0.177mm for systems *QUIN*/*QUIN-REG*.
3) The benefit of cross-stream dependency modeling is reaffirmed by comparing the *QUIN* system with the *QUIN-REG* system in Fig. 9. When the triphone ASR model is used with the *QUIN-REG* system, an average RMS error of 1.076mm is achieved, which is the best result in our experiments for the inversion mapping. Compared with previously reported results, using both the same and different databases, such as Toda et al. (2008); Richmond (2007, 2009); Hiroya and Honda (2004); Zhang and Renals (2008), this result is strong.

3.6. A Summary of Articulatory Movement Prediction

The results of the above experiments on articulatory movement prediction with different inputs are summarized in Tables 6 and 7, where the RMS errors and the correlation coefficients of the predicted movements for the 12 EMA channels are listed. For each kind of input feature in the tables, the best results of the corresponding experiments are chosen, i.e. the system

Table 6
RMS error of EMA feature prediction using different inputs. The "_x" and "_y" indicate the x- and y-coordinates of each EMA sensor respectively. The two values in each column indicate the absolute RMS error (mm) and RMS error normalized by the standard deviation of each EMA feature dimension respectively.

	Text	*Audio*	*Text & Audio*
T3_x	2.061 / 0.876	1.352 / 0.575	1.229 / 0.523
T3_y	3.091 / 0.858	1.798 / 0.499	1.307 / 0.363
T2_x	2.269 / 0.866	1.478 / 0.564	1.264 / 0.482
T2_y	3.011 / 0.831	1.314 / 0.363	1.082 / 0.299
T1_x	2.488 / 0.820	1.321 / 0.435	1.095 / 0.361
T1_y	2.966 / 0.853	1.335 / 0.384	1.178 / 0.339
LI_x	0.916 / 0.885	0.633 / 0.611	0.600 / 0.580
LI_y	1.636 / 0.905	0.835 / 0.462	0.730 / 0.404
UL_x	0.517 / 0.853	0.358 / 0.591	0.331 / 0.546
UL_y	0.731 / 0.868	0.482 / 0.573	0.385 / 0.457
LL_x	1.167 / 0.875	0.742 / 0.557	0.614 / 0.461
LL_y	2.519 / 0.980	1.270 / 0.494	0.989 / 0.385
Average	1.948 / 0.873	1.076 / 0.509	0.900 / 0.433

Table 7
Correlation coefficients between the natural and predicted EMA features using different inputs. The "_x" and "_y" indicate the x- and y-coordinates of each EMA sensor respectively.

	Text	*Audio*	*Text & Audio*
T3_x	0.608	0.786	0.822
T3_y	0.661	0.837	0.908
T2_x	0.581	0.747	0.792
T2_y	0.668	0.906	0.932
T1_x	0.580	0.781	0.819
T1_y	0.602	0.874	0.899
LI_x	0.599	0.766	0.791
LI_y	0.582	0.858	0.883
UL_x	0.568	0.761	0.812
UL_y	0.627	0.787	0.864
LL_x	0.608	0.818	0.867
LL_y	0.514	0.825	0.875
Average	0.600	0.812	0.855

FULL shown in Fig. 2 for text input, the system *QUIN-REG* using the triphone-based phone recognizer in Fig. 9 for audio input, and the system *QUIN-REG* in Fig. 6 for concurrent text and audio inputs. In these tables, we see that both the linguistic information and the supplementary audio features contribute to the prediction of all EMA channels.

Fig. 10 compares the prediction of EMA trajectories using different inputs in the form of scatter plots. An example of predicted EMA trajectories is given in Fig. 11. From these figures, we see that when both the text and acoustic features are input, the predicted articulatory features achieve the highest consistency with the natural ones in both static positions and dynamic movements. The text input is useful because it provides the correct phone transcription and context information to determine the sentence HMM for articulatory movement prediction. The importance of acoustic features lies in its synchronous and dependent relationship with the articulatory movements, which is dictated by the human speech production mechanism. Comparing system *QUIN-REG* in Fig. 6 with system *QUIN* and *FULL* in Fig. 3, we can see that if both text and audio inputs are available, the accuracy of EMA feature prediction is very close to the condition where only text input is used and state durations are given by Viterbi alignment to natural EMA trajectories. This also confirms the effectiveness of acoustic features for the task of estimating articulatory movements.

Finally, we have calculated the average RMS error when different input combinations are used for the same phone types as in Table 4. These results are shown in Fig. 12. We see that both text and acoustic inputs help the prediction of EMA features for all classes of phone. The errors for "Labial", "Labial-velar", "Velar", and "Glottal" are larger than that for the "Vowel" class when both text and acoustic features are provided as input. Comparing the *Audio* input with the

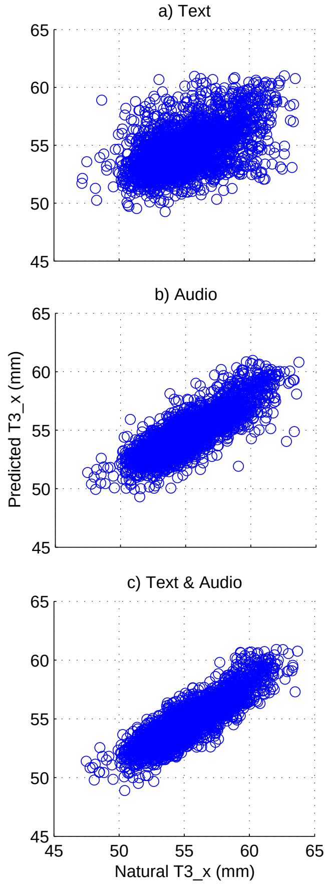


Fig. 10. Scatter plots for the x-coordinate of the T3 EMA sensor predicted using a) text, b) audio, and c) text and audio inputs. The x- and y-axes in these plots represent the natural and predicted T3_x positions respectively. Each circle in the plots corresponds to one frame in the test set.

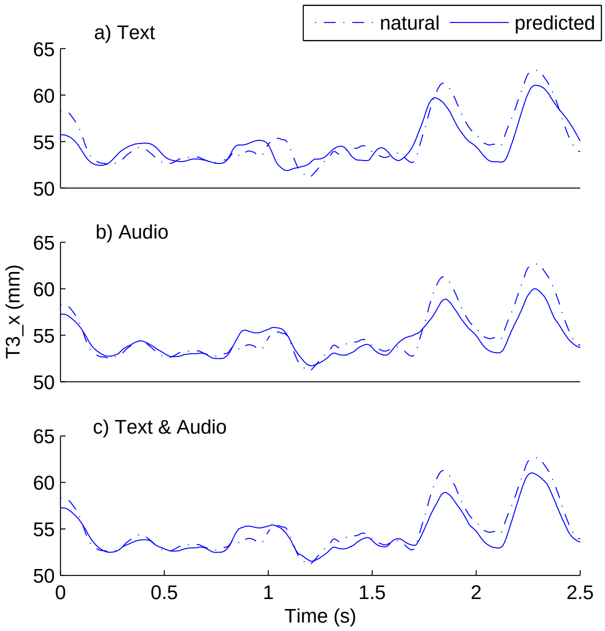


Fig. 11. Comparison between the natural and predicted x-coordinate movements of the T3 EMA sensor using a) text, b) audio, and c) text and audio inputs. The sentence is taken from the test set and the text is "For services to Lothian and Edinburgh Enterprise". Silence segments are excluded from the illustration.

Text & Audio input, we see that the consonants benefit more from the correct phone sequence than the vowels, especially for the "Postalveolar", 'Palatal", and "Velar" phone types.

4. Conclusion

In this paper, we have investigated several aspects of using an HMM-based method for predicting articulatory movements. When text is the sole input, articulatory movements are generated using an MLPG algorithm from context-dependent HMMs, which have been trained on the articulatory features. Fully context-dependent models, using rich context specifications similar to that used in TTS, outperform quinphone models, due to better modeling and prediction of state duration. For cases where an acoustic signal is available, we have introduced a unified acoustic-articulatory model and iterative optimization on state sequence to predict the articulatory movements. Our experiments have shown that quinphone models perform as well as fully context-dependent models when the acoustic signal is input with text. Furthermore, we observed the best performance using unified acoustic-articulatory HMMs with separate clustering, synchronous-state and a dependent-feature model structure. Supplementary acoustic input plays an important role in the prediction of articulatory movements. By Viterbi alignment with the input acoustic features, the predicted state durations for the articulatory movement generation are much more accurate than those predicted from the context-dependent duration probabilities for text input alone. If the acoustic features are input without text, we have found that the performance of the acoustic phone recognizer affects the inversion mapping significantly.

Finally, in terms of our intended future work in this area, we aim to look at reducing the amount of training data required for a specific speaker by applying speaker-independent modeling and model adaptation techniques. Among other benefits, this will reduce the impact of the inconvenience and cost of recording articulatory movements for any given speaker by EMA.

Acknowledgements

The authors thank Phil Hoole of Ludwig-Maximilian University, Munich for his great effort in helping record the EMA database. 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). Junichi Yamagishi is funded by EPSRC and an EC FP7 collaborative project called the *EMIME* project.

References

Baer, T., Gore, J. C., Boyce, S., Nye, P. W., 1987. Application of MRI to the analysis of speech production. Magnetic Resonance Imaging 5, 1–7.

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

Fitt, S., Isard, S., 1999. Synthesis of regional English using a keyword lexicon. In: Eurospeech. Vol. 2. pp. 823–826.

Hiroya, S., Honda, M., 2004. Estimation of articulatory movements from speech acoustics using an HMM-based speech production model. IEEE Trans. Speech Audio Process. 12 (2), 175–185.

Hiroya, S., Mochida, T., 2006. Multi-speaker articulatory trajectory formation based on speaker-independent articulatory HMMs. Speech Communication 48 (12), 1677–1690.

Kawahara, H., Masuda-Katsuse, I., de Cheveigne, A., 1999. Restructuring speech representations using pitch-adaptive time-frequency smoothing and an instanta-neous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. Speech Communication 27, 187–207.

Kiritani, S., 1986. X-ray microbeam method for the measurement of articulatory dynamics: Technique and results. Speech Communication 45, 119–140.

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

Papcun, G., Hochberg, J., Thomas, T. R., Laroche, F., Zachs, J., Levy, S., August 1992. Inferring articulation and recognising gestures from acoustics with a neural network trained on X-ray microbeam data. J. Acoust. Soc. Am. 92 (2), 688–700.

Richmond, K., 2007. Trajectory mixture density networks with multiple mixtures for acoustic-articulatory inversion. In: NOLISP. pp. 263–272.

Richmond, K., September 2009. Preliminary inversion mapping results with a new EMA corpus. In: Proc. Interspeech. Brighton, UK.

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

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

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

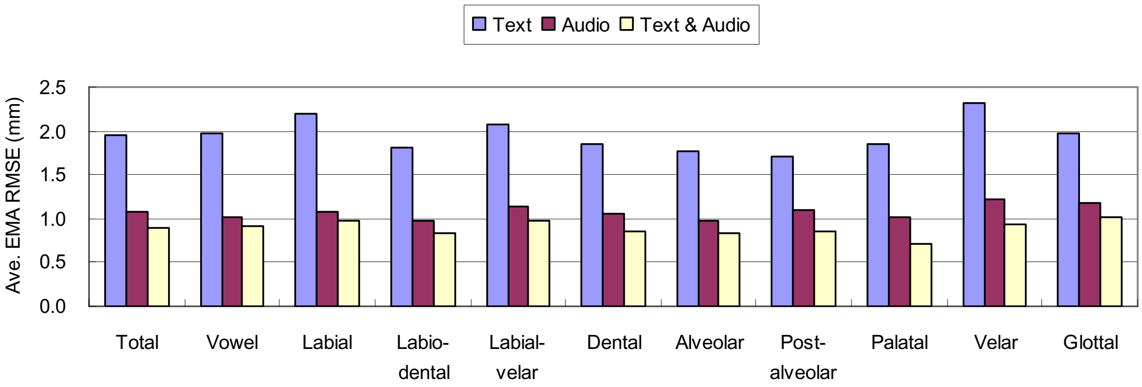


Fig. 12. Summary of RMS error of EMA feature prediction using different inputs for each phone type.

Taylor, P., Black, A. W., Caley, R., 1998. The architecture of the Festival speech synthesis system. In: 3rd ESCA Workshop in Speech Synthesis. pp. 147–151.

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

Tokuda, K., Masuko, T., Miyazaki, N., Kobayashi, T., 1999. Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. In: ICASSP. pp. 229–232.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T., 2000. Speech parameter generation algorithms for HMM-based speech synthesis. In: ICASSP. Vol. 3. pp. 1315–1318.

Tokuda, K., Zen, H., Black, A. W., 2004. HMM-based approach to multilingual speech synthesis. In: Narayanan, S., Alwan, A. (Eds.), Text to speech synthesis: New paradigms and advances. Prentice Hall.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 1998. Duration modeling in HMM-based speech synthesis system. In: ICSLP. Vol. 2. pp. 29–32.

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

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W., Tokuda, K., 2007. The HMM-based speech synthesis system (HTS) version 2.0. In: 6th ISCA Workshop on Speech Synthesis. pp. 294–299.

Zhang, L., Renals, S., 2008. Acoustic-articulatory modelling with the trajectory HMM. IEEE Signal Processing Letters 15, 245–248.