Articulatory Control of HMM-based Parametric Speech Synthesis using Feature-Space-Switched Multiple Regression

Citation for published version:

Digital Object Identifier (DOI):
10.1109/TASL.2012.2215600

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Early version, also known as pre-print

Published In:
IEEE Transactions on Audio, Speech and Language Processing

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Abstract—In previous work we proposed a method to control the characteristics of synthetic speech flexibly by integrating articulatory features into a hidden Markov model (HMM) based parametric speech synthesiser. In this method, a unified acoustic-articulatory model is trained, and context-dependent linear transforms are used to model the dependency between the two feature streams. In this paper, we go significantly further and propose a feature-space-switched multiple regression HMM to improve the performance of articulatory control. A multiple regression HMM (MRHMM) is adopted to model the distribution of acoustic features, with articulatory features used as exogenous “explanatory” variables. A separate Gaussian mixture model (GMM) is introduced to model the articulatory space, and articulatory-to-acoustic regression matrices are trained for each component of this GMM, instead of for the context-dependent states in the HMM. Furthermore, we propose a task-specific feature tailoring method to ensure compatibility between state context features and articulatory features that are manipulated at synthesis time. The proposed method is evaluated on two tasks, using Japanese vowels and target vowels by varying articulatory inputs than our previous method. The new method achieves better performance when reconstructing the combined phoneme articulation and the state context features and articulatory features that are manipulated at synthesis time. The proposed method is evaluated on two tasks, using Japanese vowels and target vowels. Compared to our previous method, the new method is highly effective at producing a new vowel from appropriate articulatory representations which, even though no acoustic samples for this vowel are present in the training data, is shown to sound highly natural.

Index Terms—Speech synthesis, articulatory features, multiple-regression hidden Markov model, Gaussian mixture model

I. INTRODUCTION

In recent years, hidden Markov models (HMM) have been successfully applied to acoustic modelling for speech synthesis, and HMM-based parametric speech synthesis has becoming a mainstream speech synthesis method [1], [2]. In this method, the spectrum, F0 and segment durations are modelled simultaneously within a unified HMM framework [1]. At synthesis time, these features are predicted from the sentence HMM by means of a maximum output probability parameter generation (MOPPG) algorithm that incorporates dynamic features [3]. The predicted parameter trajectories are then sent to a parametric synthesiser to reconstruct the speech waveform. This method is able to synthesise highly intelligible and smooth speech sounds [4], [5]. Another significant advantage of this model-based parametric approach is that it makes speech synthesis far more flexible compared to the conventional unit selection and waveform concatenation approach.

Specifically, several adaptation and interpolation methods have been applied to control the HMM’s parameters and so diversify the characteristics of the generated speech [6]–[10]. However, this flexibility relies upon data-driven machine learning algorithms and is strongly constrained by the nature of the training or adaptation data that is available. In some instances, though, we would like to integrate phonetic knowledge into the system and control the generation of acoustic features directly when corresponding training data is not available. For example, this phonetic knowledge could be place of articulation for a specific phone, the differences in phone inventories between two languages, or physiological variations among different speakers. Unfortunately, it is difficult to achieve this goal because the acoustic features used in conventional HMM-based speech synthesis are typically the parameters that are required to drive a speech vocoder, which do not enable fine control in terms of the human speech production mechanism.

We have previously proposed a method to address this problem and to achieve flexible control over HMM-based speech synthesis by integrating articulatory features [11], [12]. Here, we use “articulatory features” to refer to the continuous movements of a group of speech articulators, for example the tongue, jaw, lips and velum, recorded by human articulography techniques such as electromagnetic articulography (EMA) [15], magnetic resonance imaging (MRI) [16] or ultrasound [17]. In this method, a unified acoustic-articulatory model is

1This is often referred to as maximum likelihood parameter generation (MLPG) in the literature. However, in accordance with the technical difference between “likelihood” (which interprets the probability distribution as a function of the model parameters given a fixed outcome) and “probability” (which interprets the probability distribution as a function of the outcome given fixed model parameters), the term “output probability” is used in this paper in place of “likelihood” to refer to the parameter generation criterion.

2In some literature, the term “articulatory features” may refer to the scores for pre-defined articulatory classes, such as nasality or voicing, which can be extracted from acoustic speech signals [13]. This kind of articulatory feature has also been applied to expressive speech synthesis in recent work [14].

This work is partially funded by the National Nature Science Foundation of China (Grant No. 60905010) and the National Natural Science Foundation of China - Royal Society of Edinburgh Joint Project (Grant No. 6111130120). The research leading to these results was partly funded from the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement 256230 (LISTA), and EPSRC grants EP/I027696/1 (Ultrax) and EP/J000256/2. Part of this work has been presented at Interspeech (Florence, Italy, August 2011) [34].

Z. Ling is with iFLYTEK Speech Lab, University of Science and Technology of China, Hefei, 230027 P.R.China. E-mail: zhling@ustc.edu.

K. Richmond and J. Yamagishi are with the Center for Speech Technology Research (CSTR), University of Edinburgh, Edinburgh, EH3 9AQ United Kingdom. E-mail: korin@cstr.ed.ac.uk; jyamagis@inf.ed.ac.uk.

trained and a piecewise linear transform is adopted to model the dependency of the acoustic features on the articulatory features. During synthesis, articulatory features are first generated from the trained model. The generation of acoustic features and the characteristics of synthetic speech can then be controlled by modifying these generated articulatory features in arbitrary ways, for example in accordance with phonetic rules. Experimental results have shown the potential of this method for controlling the overall characteristics of synthesised speech, as well as the identity of specific vowels [12].

The initial motivation for developing this method was in fact two-fold: to gain articulatory control, of course, but also to improve the accuracy of acoustic feature generation. Consequently, both these aims influenced the model structure we developed. However, in terms of optimising articulatory control alone, we hypothesise there are some shortcomings in this model structure which could be improved in order to achieve even better control. First, in short, the articulatory features are constrained to be generated from the unified acoustic-articulatory model, which makes the integration of phonetic knowledge into articulatory movement prediction somewhat inconvenient. Second, the transform matrices between articulatory and acoustic features are trained for each HMM state and are tied based on context using a decision tree as for other model parameters. This may prove problematic when articulatory features are modified by significant amounts during synthesis because the fixed transform matrix may no longer be appropriate for the new articulator positions. Third, the unified acoustic-articulatory model is trained without considering the specific task of articulatory control. The modified articulatory features could conflict with the context information used in model training and parameter generation.

To address these shortcomings, an improved method for articulatory control over HMM-based parametric speech synthesis is proposed in this paper. As the first improvement, a multiple-regression hidden Markov model (MRHMM) [18] is introduced to replace the unified acoustic-articulatory HMM used in our previous work. This makes it possible to integrate other forms of articulatory prediction model. The MRHMM was initially proposed to improve the accuracy of acoustic modelling for automatic speech recognition (ASR) by utilising auxiliary features that are correlated with the acoustic features [18]. The auxiliary features that have been used in this way include fundamental frequency [18], emotion and speaking style [19], for example. The MRHMM has also been applied to HMM-based parametric speech synthesis, with sentence-level style vectors being used as the explanatory variables [10]. In this paper, we propose to treat articulator movements as the external auxiliary features to help determine the distribution of acoustic features.

As a second improvement, we propose a feature-space regression matrix switching method for the MRHMM in order to address the restriction that comes with context-dependent regression matrix training for articulatory control. In this method, a separate Gaussian mixture model (GMM) is introduced to model the articulatory space, and the regression matrices are estimated for each mixture component in this GMM instead of for each HMM state. This idea is similar to the switching system in the field of control systems, e.g. [20], where impedance parameters are switched according to the contact configuration during the assembly process.

Finally, as a third improvement, a strategy of task-specific context feature tailoring is presented to avoid potential conflicts arising between state context information and the articulatory features that are generated and modified at synthesis time.

The remainder of this paper is organised as follows. Section 2 gives a brief overview of the unified acoustic-articulatory modelling method proposed in our previous work. Section 3 describes our proposed novel method in detail. Section 4 presents the experiments we have conducted and their results, and Section 5 gives the conclusions we draw from this work.

II. UNIFIED ACOUSTIC-ARTICULATORY MODELLING

A. Model training

Our previous work took the general framework of HMM-based parametric speech synthesis and integrated articulatory features into the conventional model for acoustic features by expanding the observed feature vectors [12]. Let \( X = [x_1^T, x_2^T, ..., x_T^T] \) and \( Y = [y_1^T, y_2^T, ..., y_T^T] \) denote the parallel acoustic and articulatory feature vector sequences of the same length \( T \). For each frame, the feature vectors \( x_t \in \mathbb{R}^{3D_X} \) and \( y_t \in \mathbb{R}^{3D_Y} \) consist of static parameters and their velocity and acceleration components, where \( D_X \) and \( D_Y \) are the dimensions of static acoustic features and static articulatory features respectively. The detailed definition of these dynamic features may be found in [12]. The feature production model used in this method is illustrated in Fig. 1. A piecewise linear transform is added to the parameters of the HMM (transform matrices \( A_j \)) to represent the dependency between the acoustic features and the articulatory movements. During model training, an HMM \( \lambda \) is estimated by maximising the likelihood function of the joint distribution \( P(X, Y | \lambda) \), which can be written as

\[
P(X, Y | \lambda) = \sum_q \pi_{q_0} T \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(x_t, y_t), \quad (1)
\]

\[
b_{j}(x_t, y_t) = b_j(x_t | y_t) b_j(y_t), \quad (2)
\]

\[
b_{j}(y_t) = \mathcal{N}(y_t; \mu_{Y_j}, \Sigma_{Y_j}), \quad (3)
\]

\[
b_{j}(x_t, y_t) = \mathcal{N}(x_t; A_j y_t + \mu_{X_j}, \Sigma_{X_j}), \quad (4)
\]

where \( q = \{q_1, q_2, ..., q_T \} \) is the state sequence shared by the two feature streams; \( \pi_j \) and \( a_{ij} \) represent initial state probability and state transition probability; \( b_j(\cdot) \) is the state observation probability density function (PDF) for state \( j \); \( \mathcal{N}(\cdot; \mu, \Sigma) \) denotes a Gaussian distribution with a mean vector \( \mu \) and covariance matrix \( \Sigma \); and \( A_j \in \mathbb{R}^{3D_X \times 3D_Y} \) is the linear transform matrix for state \( j \). This matrix is context-dependent and tied to a given regression class using a decision tree, and hence a globally piecewise linear transform is achieved. The model parameters can be estimated using the EM algorithm, as described in [12].
B. Parameter generation

A flowchart summarising the generation of acoustic features with articulatory control is shown in Fig. 2. The MOPPG algorithm, which embodies explicit constraints inherent in the dynamic features [3], is employed to generate articulatory and acoustic features from the trained model. In order to control the characteristics of the synthetic speech flexibly, the generated articulatory features may be modified according to phonetic knowledge to reproduce acoustic parameters that reflect those changes appropriately. The detailed formulae for this parameter generation process were introduced in [12].

C. A discussion on articulatory control over synthesis

In previous experiments we have shown the method of unified acoustic-articulatory modelling with cross-stream dependency described above can achieve effective control over the characteristics of the synthesised speech [12]. However, we also note the degree of control that is possible with that method has not yet fully met our expectations. For example, one experiment in [12] demonstrated control over vowel identity through modification of tongue height. However, though the experiment proved this modification to be effective and convincing, it was necessary to raise or lower the tongue position by approximately 1.0 cm to achieve a clear transition from vowel /e/ to /æ/ or /æ/ [12]. This range of modification is larger than the differences in tongue height among these three vowels that we observe in the recorded database, as shown in Fig. 3. In fact, as mentioned in Section I, we can identify three aspects of the structure of the model presented in [12] that in theory restrict or limit the scope of articulatory control that is possible. We shall consider these three factors in more detail next.

The first limitation arises from the fact that the articulatory features are generated from the unified acoustic-articulatory HMM, which is trained context-dependently and contains a large number of parameters. At synthesis time, there are two ways to effect articulatory control: by manipulating either i) the articulatory PDF parameters or ii) the generated articulatory feature trajectories. Both these approaches have inherent advantages and disadvantages. On one hand, for example, it is relatively straightforward to modify the mean vectors of Gaussian PDFs (e.g. to add an offset to the appropriate articulatory PDF mean parameters to change the target position of the tongue). On the other hand, it is less obvious how to manipulate covariance matrices directly according to phonetic
rules, or indeed how to modify mean and variance parameters to obtain exactly the articulatory trajectories that are desired after processing with the MOPPG algorithm. Meanwhile, the second approach of modifying the generated articulatory trajectories instead also becomes problematic for example if the phonetic rules are not applied globally but to some specific phones, because extra smoothing algorithms are necessary to ensure the continuity and naturalness of the modified articulatory trajectories. With such difficulties in mind, it is interesting to note that there exist other forms of generative model, such as target approximation models [21], [22], which offer a model structure that is more compact and easier to control than the HMM used for articulatory prediction in [12]. Thus, to make it more convenient to control an HMM-based synthesiser via articulation, it seems prudent to consider separating the model for predicting articulatory movements from the unified acoustic-articulatory HMMs.

The second limitation lies in the way the articulatory-acoustic relationship is modelled. As mentioned in Section II-A, a globally piecewise linear model is used to represent this relationship, in the form of a number of (tied) state-dependent linear transform matrices \(A_j\) in (4). For small articulatory modifications, the local linear relationship dictated by state index \(j\) is likely to remain appropriate. However, with larger changes, as the modified articulatory features are moved further from their initial starting point, it becomes less reasonable to assume that the same linear relationship will be appropriate. In fact, it may be that a significantly different local linear transform becomes more appropriate, but the model structure in [12] is unable to react to such changes in the generated articulatory features, and is instead unfortunately constrained to use the same fixed transform matrices dictated by the state-dependent context features.

Finally, as the third limitation, it should also be noted in (4) that not only are the articulatory-to-acoustic transform matrices \(A_j\) fixed according to the state context features, but so too are the acoustic distribution parameters \(\mu_{X_i}\) and \(\Sigma_{X_j}\). Hence, modifying the generated articulatory trajectories at synthesis time (using either approach above) risks introducing a conflict with this HMM state context information. In some instances, for example, we might wish to modify the generated articulatory features to a relatively large extent, so as to change the identity of a vowel or to generate a new, significantly different speaking style. However, in attempting such large modifications we introduce a conflict, since the modified articulatory features will be incompatible with the other state-dependent model parameters in (4), which will still correspond to the context features of the acoustic unit before the articulatory modification.

### III. Feature-Space-Switched MRHMM for Articulatory Control of Speech Synthesis

In order to overcome the shortcomings in our previous approach, an improved method to gain articulatory control over HMM-based synthesis is proposed in this paper. Fig. 4 gives a flowchart illustrating how acoustic features are generated in this new method. In summary, this method proposes to use a feature-space-switched MRHMM (FSS-MRHMM) for acoustic modelling. Unlike our previous approach, the articulatory features are used as external (or exogenous) explanatory variables for regression. Meanwhile, instead of tying the regression matrices in the MRHMM in a state-dependent way, we tie them within the articulatory space, which ultimately allows adaptive regression matrix switching in response to articulatory modification. Finally, subsets of context features for context-dependent model training are specially selected, or “tailored”, so as to avoid conflict between context-dependent model parameters and modified articulatory features at synthesis time. The details of this proposed method will be discussed in greater depth next.

#### A. MRHMM for HMM-based parametric speech synthesis

As illustrated in Fig. 4, the unified acoustic-articulatory HMM for acoustic modelling in Fig. 2 is replaced by an MRHMM together with a separate external articulatory prediction model. At this stage, we are focussing on the acoustic modelling part; the external articulatory prediction model is not within the primary scope of this paper, and will instead be the subject of future work. For the experiments presented in this paper, we have chosen to use a baseline articulatory prediction method that was readily available to us, and which is described further in Section IV.

We shall begin by briefly reviewing the MRHMM approach to acoustic modelling. This model was initially proposed to model acoustic features better by utilising auxiliary features [18]. Its feature production model is shown in Fig. 5. The difference between this model and standard HMMs is that an auxiliary feature sequence \(Y\) is introduced to supplement the state sequence \(q\) for determining the distribution of the acoustic feature sequence \(X\). In this paper, the auxiliary feature sequence \(Y\) is comprised of the articulatory trajectories. Mathematically, the distribution of \(X\) in the conventional MRHMM can be written [18] as

\[
P(X | \lambda, Y) = \sum_{q} \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(x_t | y_t), \tag{5}
\]

\[
b_j(x_t | y_t) = \mathcal{N}(x_t; A_j \xi_t + \mu_j, \Sigma_j), \tag{6}
\]

where \(X, Y, \pi_j, a_{ij}, b_j(\cdot)\) and \(\mathcal{N}(\cdot; \mu, \Sigma)\) have the same definition as in (1)-(4); \(q = \{q_1, q_2, ..., q_T\}\) is the state
considered.

The relationship between the articulatory features and the F0 features is not

acoustic feature sequence

butions [2]. Given auxiliary feature sequence

A

class using a decision tree. Eq. (6) is similar to (4) in the

unified acoustic-articulatory modelling, whereby

A

represents the mean of the transform residuals. Research on

speech production informs us that the relationship between

acoustic and articulatory features is complex and nonlinear

in form. Here, a piecewise linear transform is adopted to

approximate this nonlinear relationship. The effectiveness of

this approximation has been demonstrated in previous work

[12], [23], [24]. Eq. (6) is also similar to the state PDF

in cluster adaptive training (CAT) of HMMs [25], where each

column of

A

corresponds to the mean vector of one cluster and

ξ

corresponds to the cluster weight vector. The difference is

that

ξ

in the MRHMM is observable, whereas the cluster

weight vector in CAT needs to be estimated for each speaker

(or other factor).

To build the MRHMM-based parametric speech synthesis

system in this paper, the procedures of standard HMM-

based synthesiser training [2] are first followed in order to

initialise model parameters by maximising

\[ P(X|\lambda) \]

without using articulatory features. The acoustic features consist of F0

and spectral parameters extracted from the waveforms of the

training set. A multi-space probability distribution (MSD) [26] is

applied for F0 modelling to address the problem that F0 is

only defined for voiced speech segments. Context-dependent

HMMs are trained using richly-defined contexts that include

detailed phonetic and prosodic features [2]. A decision-tree-

based model clustering technique that uses the minimum

description length (MDL) criterion [27] is adopted to deal

with the data-sparcity problem and to estimate the parameters

of models whose context description is missing in the training

set. Then, the estimated mean vector and covariance matrix

for each state are used as the initial values of

μ

and

Σ

in an MRHMM. The regression matrix

A

is initialised as a zero matrix. These parameters are iteratively updated to maximise

\[ P(X|\lambda, Y) \]

by introducing articulatory features and using the EM algorithm3. The detailed formulae are to be found in [18].

Next, a state alignment to the acoustic features is performed

using the trained MRHMM in order to train context-dependent

PDF parameters for state duration prediction [1].

At synthesis time, the maximum output probability criterion

[3] is adopted to generate acoustic features. For the purpose of simplification, only the optimal HMM state

sequence is considered. First, the optimal state sequence

\[ q^* = \{q_1, q_1^*, \ldots, q_T^*\} \]

is predicted using the trained duration distributions [2]. Given auxiliary feature sequence

Y, the optimal acoustic feature sequence

X^* is generated by maximising

\[ P(X|\lambda, Y, q^*) = \prod_{t=1}^{T} \mathcal{N}(x_t; A_{q^*_t} \xi_t + \mu_{q^*_t}, \Sigma_{q^*_t}). \] (7)

This can be solved using the conventional MOPPPG algorithm

3For training the MRHMMs, \( X \) only contains the spectral feature stream. The relationship between the articulatory features and the F0 features is not considered.

\[ b_j(x_t|y_t) = \sum_{k=1}^{M} P(x_t, m_t = k|y_t, q_t = j, \lambda, \lambda^{(G)}), \] (8)

\[ = \sum_{k=1}^{M} \zeta_{k}(t) P(x_t|y_t, q_t = j, m_t = k, \lambda, \lambda^{(G)}), \] (9)

where

\[ m_t \]

denotes the mixture index of \( \lambda^{(G)} \) for the articulatory feature vector at frame\( t \); the HMM state sequence

\( q \) and the GMM mixture sequence \( m = \{m_1, m_2, \ldots, m_N\} \) are
reasonably assumed to be independent of each other, so that
\[ P(m_t = k | y_t, q_t = j, \lambda, \lambda^{(G)}) = P(m_t = k | y_t, \lambda^{(G)}) = \zeta_k(t). \]

(10)

For each Gaussian mixture, the dependency between the acoustic features and the auxiliary articulatory features is represented by
\[ P(x_t | y_t, q_t = j, m_t = k, \lambda, \lambda^{(G)}) = N(x_t; A_k \xi_t + \mu_j, \Sigma_j), \]

(11)

where \( A_k \in \mathcal{R}^{3D \times (3D+1)} \) is the regression matrix for the \( k \)-th mixture of \( \lambda^{(G)} \). Note that an extra Gaussian mixture component index sequence \( m_t \) is introduced to determine the regression matrix for each frame, whereas (6) uses state index \( j \) to determine the regression matrix. Furthermore, we can interpret \( \zeta_k(t) \) as a weight that varies according to the articulatory features, and which changes how each transform matrix is weighted, or “blended” together, according to (9). It is in this way that “soft” regression classes are achieved.

A similar model structure can be found in subspace GMM modelling [28], where all HMM states share the same GMM structure and the state-dependent subspace vectors play the role of external articulatory features in our method.

To train the HMM parameter set \( \{A_k, \mu_j, \Sigma_j\}^4 \), we substitute (8)-(11) into (5) and get
\[ P(X|\lambda, Y) = \sum_q \sum_m P(X,q,m|\lambda,Y), \]

(12)

where
\[ P(X,q,m|\lambda,Y) = \pi_{q0} \prod_{t=1}^{T} a_{q_{t-1}q_t} \zeta_{m_t}(t). \]
\[ N(x_t; A_m \xi_t + \mu_q, \Sigma_q). \]

(13)

The EM algorithm is adopted to estimate the parameter set that maximises (12). The auxiliary function is defined as
\[ Q(\lambda, \lambda') = \sum_q \sum_m P(X,q,m|\lambda,Y) \log P(X,q,m|\lambda',Y) \]
\[ = \sum_{j=1}^{N} \sum_{k=1}^{M} \sum_{t=1}^{T} \gamma_j(t) \zeta_k(t) \log N(x_t; A_k \xi_t + \mu_j', \Sigma_j') + K, \]

(15)

where \( K \) is a constant term that is independent of the model parameter set; \( \gamma_j(t) = P(q_t = j|\lambda,X,Y) \) is the state occupancy probability of MRHMM state \( j \) at time \( t \); \( N \) is the total number of HMM states.

In order to re-estimate the transform matrix \( A_k' \) for each GMM mixture, we set \( \partial Q(\lambda, \lambda') / \partial A_k' = 0 \) and get
\[ \sum_{j=1}^{N} \sum_{t=1}^{T} \gamma_j(t) \zeta_k(t) \Sigma_j^{-1}(x_t - \mu_j) \xi_t^T \]
\[ = \sum_{j=1}^{N} \sum_{t=1}^{T} \gamma_j(t) \zeta_k(t) \Sigma_j^{-1} A_k' \xi_t \xi_t^T. \]

(16)

This equation can be simplified as
\[ Z = \sum_{t=1}^{T} V(t) A_k' D(t), \]

(17)

where
\[ Z = \{z_{it}\} = \sum_{j=1}^{N} \sum_{t=1}^{T} \gamma_j(t) \zeta_k(t) \Sigma_j^{-1} (x_t - \mu_j) \xi_t^T, \]

(18)

\[ V(t) = \text{diag} \{v_{it}^{(t)}\} = \sum_{j=1}^{N} \gamma_j(t) \Sigma_j^{-1}, \]

(19)

\[ A_k' = \{a_{ip}'; \xi_t \}, \]

(20)

\[ D(t) = \{d_{pl}^{(t)}\} = \zeta_k(t) \xi_t \xi_t^T, \]

(21)

According to (17), each element in \( Z \) can be calculated as
\[ z_{it} = \sum_{t=1}^{T} \sum_{p=1}^{3D+1} v_{ip}^{(t)} a_{ip}' d_{pl}^{(t)} = \sum_{p=1}^{3D+1} a_{ip}' \sum_{t=1}^{T} v_{it}^{(t)} d_{pl}^{(t)}. \]

(22)

Therefore, the transform matrix \( A_k' \) can be updated line by line. For the \( i \)-th line,
\[ a_{i}' = G(i)^{-1} z_{i}, \]

(23)

where \( z_i = [z_{i1}, z_{i2}, \ldots, z_{i(3D+1)}]^T; a_{i}' = [a_{i1}', a_{i2}', \ldots, a_{i(3D+1)}', G(i) = \{g_{ip}\} \) and \( d_{pl} = \sum_{t=1}^{T} v_{it}^{(t)} d_{pl}^{(t)} \).

The re-estimation formulae for the other model parameters can be derived by setting \( \partial Q(\lambda, \lambda') / \partial \lambda = 0 \), such that
\[ \mu_j' = \frac{\sum_{k=1}^{M} \sum_{t=1}^{T} \gamma_j(t) \zeta_k(t) (x_t - A_k' \xi_t)}{\sum_{t=1}^{T} \gamma_j(t)} \]
\[ \Sigma_k' = \frac{1}{\sum_{t=1}^{T} \gamma_j(t) \sum_{k=1}^{M} \sum_{t=1}^{T} \gamma_j(t) \zeta_k(t)} \Sigma_k \]
\[ \cdot (x_t - \mu_j' - A_k' \xi_t)(x_t - \mu_j' - A_k' \xi_t)^T, \]

(24)

(25)

At synthesis time, the parameter generation criterion in (7) is modified to
\[ P(X|Y, \lambda, q) = \prod_{t=1}^{T} \sum_{k=1}^{M} \zeta_k(t) N(x_t; A_k \xi_t + \mu_q, \Sigma_q), \]

(26)

where \( \zeta_k(t) \) is calculated based on the input articulatory features \( Y \). This is an MOPPG problem with mixtures of Gaussians at each frame. We can solve it either by using an EM-based iterative estimation method [3] (thus retaining the effect of “soft” clustering at synthesis time) or by considering only the optimal mixture sequence as a simplification.

C. Task-specific context feature tailoring

In a context-dependent MRHMM, the motivation for using context information and for introducing auxiliary features is the same. The aim is to improve the accuracy of acoustic modelling by taking into account external factors that could affect the distribution of acoustic features. When applying an MRHMM to automatic speech recognition, the auxiliary features supplement the context information to influence the
acoustic distribution at each HMM state. These auxiliary features are observable and fixed at decoding time. However, for MRHMM-based parametric speech synthesis with articulatory control, the articulatory features are generated at synthesis time and may be manipulated to reflect any phonetic knowledge we might wish to impart. This introduces the potential for conflict between the manipulated articulatory features and the context features, as discussed in Section II.C. Although the feature-space-switched MRHMM in Section III.B can determine the regression matrices without using context information, \( \mu_j \) and \( \Sigma_j \) in (11) are still dependent upon context.

Ultimately, the purpose of using articulatory features here is not to refine the distribution of acoustic features for a given context description, but to partially replace the function of the context features in order to gain flexibility in determining the distribution of acoustic features. Therefore, to avoid any conflict, we propose a strategy of task-specific context feature tailoring. Under this strategy, the full set of context features is divided into a base subset and a control subset. Only those features in the base subset are used for the context-dependent model training, whereas the control subset contains the context information that can be substituted by the articulatory features. In this way, we aim to ensure the context features and the articulatory features are compatible and complementary. Deciding which features to put in the base subset depends on the specific task in hand. Several examples of sets of context features tailored for specific tasks are given in Table I.

Generally, the more context features that can be replaced by adding articulatory features, the greater the flexibility we stand to gain in terms of articulatory control. The extreme case would be to discard all context information and build an articulatory-to-acoustic mapping at feature sequence level to gain complete control over the generation of acoustic features using articulatory inputs. However, it should be noted that performance will depend heavily on the consistency and scope of the articulatory features available. For example, it would be impossible to control the degree of nasality using EMA data in which a sensor coil had not been placed on the velum. Relatively recent work shows the accuracy of purely articulatory-to-acoustic mappings is still unsatisfactory [29], and this suggests that the articulatory features captured using current articulography techniques may not yet provide a description of the articulatory process that is fully adequate. But the aim of our work is to achieve the desired flexibility without degrading the naturalness of the synthetic speech significantly. Hence, the proposed context-tailoring approach represents a compromise between using the full set of context features and building a pure articulatory-to-acoustic mapping, and effectively boils down to finding a trade-off between naturalness and flexibility. In principle, higher quality and naturalness can be achieved if more context features are reserved in the base subset. Conversely, keeping fewer context features in the base subset can give greater flexibility in terms of articulatory control over the synthetic speech. In time, it is possible more elaborate articulatory control will become achievable with the development of new articulography and data processing techniques. But for now any limitations inherent in the articulatory data available make it more difficult to move context features into the control subset and still retain full naturalness in the synthetic speech.

### IV. Experiments

#### A. Database

The same multi-channel articulatory database used in our previous work [12] was adopted for the experiments of this paper. This database has been released with a free licence for research use. As far as we know, it provides the largest amount of data from a single speaker, and with the best sensor position consistency, compared to any other articulatory corpus that is publicly available [30]. It contains acoustic waveforms recorded concurrently with EMA data using a Carstens AG500 electromagnetic articulograph. A male British English speaker was recorded reading around 1300 phonetically balanced sentences. The waveforms used were in 16kHz PCM format with 16 bit precision. Six EMA sensors were placed on the speaker’s articulators, 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 head from the front). Because the movements in the z-axis were small, only the x- and y-coordinates of the six sensors were used in our experiments, making a total of 12 static articulatory features.
in each frame. The static acoustic features were composed of F0 and 40-order frequency-warped LSPs [5] plus an extra gain dimension, which were derived using STRAIGHT [31] analysis. The frame shift was set to 5ms.

B. Vowel modification task

1) Experimental conditions: As a first step to evaluating controllability in the various systems described above, we chose the task of changing the perceptual identity of one vowel type into another. This control would potentially be useful for computer-assisted language learning applications, or human perception experiments in phonetic research, for example. Five acoustic models were trained and compared in this experiment. Descriptions of these models are provided in Table II. Selecting this group of systems allowed us to evaluate the three major aspects of the proposed approach, while ensuring perceptual tests remained within practical limitations. We selected 1200 sentences from the database for model training, and the remaining 63 sentences were used as a test set. A five-state, left-to-right HMM structure with no skips was adopted to model the acoustic features. Diagonal covariance matrices were used for all five systems. The STD-F system was trained following the conventional HMM-based parametric speech synthesis approach [2]. The UNI-FC system was identical to the FD system in [12], where 100 context-dependent transform matrices were used to model the relationship between articulatory and acoustic features.

The MR-FC, MR-FF, and MR-TF systems were trained as described in Section III. As for the UNI-FC system, the regression matrices in these three systems were defined as three-block matrices corresponding to the static, velocity, and acceleration components of the feature vector in order to reduce the number of parameters that needed to be estimated. The task-specific context feature tailoring for the MR-TF system followed the scheme listed in the first row of Table I, where vowel ID is used as the control subset of the context features. For the MR-FC system, the number of regression matrices was set to 100 in order to match the UNI-FC system. In the MR-FF and MR-TF systems, the optimal numbers of GMM mixture components for the feature-space regression matrix switching were determined using the minimum description length criterion [27]. Here, the description length is defined as

$$D(\lambda) \equiv - \log P(\mathbf{X}|\lambda, \mathbf{Y}) + \frac{1}{2} D(\lambda) \log G + C$$  \(27\)

where $\log P(\mathbf{X}|\lambda, \mathbf{Y})$ is the log likelihood function of the model for the training set; $D(\lambda)$ is the dimensionality of the model parameters; $G$ is the total number of observed frames in the training set; $C$ is a constant. Considering the three-block matrix structure of $A_k$, $D(\lambda) = 3MD_X(D_Y + 1) + C_D$, where $C_D$ is a constant that is independent from the number of mixtures $M$ for each system. Ignoring the constant components in (26), we calculated the average description length per frame on the training set as

$$\bar{D}(\lambda) = - \frac{1}{T} \log P(\mathbf{X}|\lambda, \mathbf{Y}) + \frac{3}{2T} M D_X(D_Y + 1),$$  \(28\)

where $T$ is the number of frames in training feature sequence $\mathbf{X}$. The results for the MR-FF and MR-TF systems with $M = 8, 16, 32, 64, 128$ are shown in Fig. 7, from which we see that $M = 64$ leads to the minimum description length for both systems. Thus, we used 64 Gaussian mixtures for $\lambda^{(G)}$ and trained the regression matrices of the MRHMM for each mixture component in these two systems. For the MR-FF and MR-TF systems, only the optimal mixture sequence was considered when solving (26) during parameter generation.

In our previous work [12], monosyllabic words were embedded into a carrier sentence to conduct the vowel identity modification experiment. However, these sentences were composed artificially and had no corresponding natural acoustic and articulatory recordings. Therefore, we found it was not possible to guarantee the appropriateness of the input articulatory features for synthesising the target vowels and to calculate the generation error of the modified acoustic features objectively. Instead, the 63 sentences in the test set of the
2) Objective evaluation: The difference between the generated acoustic features after vowel modification and the natural recordings of these test sentences was adopted as an objective measure to evaluate the performance of each system in the vowel identity modification task. Root mean square error (RMSE) between two LSP sequences [12] was used to quantify this difference. To simplify the calculation of RMSE, the LSPs were generated using state durations derived from state alignment against the natural speech performed using each system. The resulting LSP RMSEs for different systems and different types of phone are shown in Fig. 8, where the label Ref. denotes the acoustic features generated using the STD-F system and the original context features without vowel replacement. The natural articulatory inputs were derived from the articulatory channel of the recorded database. The generated articulatory features were predicted based on the original phone transcription of the test sentences. For the UNI-FC system, the articulatory components of the unified acoustic-articulatory HMMs were used to generate the articulatory features according to the method proposed in our previous work [12]. As mentioned in Section III.A, the articulatory prediction model in Fig. 4 is not the emphasis of this paper. Thus, we adopted our HMM-based articulatory movement prediction method [32] for the three MRHMM-based systems. This method is similar to conventional HMM-based parametric speech synthesis. Context-dependent HMMs are trained using only the articulatory features of the training set, which consist of static, velocity and acceleration components. At synthesis time, articulatory movements are predicted from the input text using the trained models and the MOPPG algorithm. Here, full context features were used for training the articulatory HMM, and the generated articulatory features were synchronised with the acoustic features at state boundaries.

In Fig. 8, the RMSEs observed when modifying /ɛ/ to non-/ɛ/ vowels are of most interest. Comparing the results of the STD-F and Ref. systems in Fig. 8, we find that replacing all vowels to /ɛ/ increases the prediction error for the acoustic features greatly, especially for the non-/ɛ/ vowels (from 0.607 to 1.031). This is clearly to be expected, since the acoustic parameter generation is wholly dictated by the context information in standard HMM-based speech synthesis and different vowels have significantly different acoustic realisation. Using the articulatory features corresponding to the target phone transcription, the prediction errors of the UNI-FC system are much smaller than the STD-F system, especially for the non-/ɛ/ vowels (from 1.031 to 0.877 with natural articulatory features, and to 0.896 with generated articulatory features). This demonstrates the effectiveness of our previous approach using unified acoustic-articulatory HMMs [12] for vowel identity modification. The performance of the MR-FC system is close to the UNI-FC system when either natural or generated acoustic features are used. This is reasonable because the model structures of these two systems are similar; the only difference is that the likelihood of the articulatory features is not part of the model training criterion for the MR-FC system, as shown in Fig. 5 and (5). On the other hand, the vowel identity modification results of both the UNI-FC and MR-FC systems are still unsatisfactory because the RMSEs of these two systems for the non-/ɛ/ vowels are significantly higher than for the Ref. system which utilises the target phone transcription for synthesis.

Meanwhile, Fig. 8 also shows that both the feature-space-switched MRHMM model structure and the task-specific context feature tailoring method proposed in this paper further improve the performance of the MR-FC system in vowel identity modification. When natural articulatory features are

---

Fig. 8. LSP RMSEs for different systems and types of phone in the vowel identity modification experiment. “vowels-ex-/ɛ/” indicates all vowels excluding the source vowel /ɛ/ in the modification. The label Ref. represents the acoustic features generated using the STD-F system and the original context features without vowel replacement. For the UNI-FC, MR-FC, MR-FF, and MR-TF systems, the a) natural and b) generated articulatory features were used respectively.

Some examples of the synthetic speech used in the vowel modification experiment can be found at http://staff.ustc.edu.cn/~zhling/MRHMM-EMA/demo.html.
used as the explanatory variables of multiple regression, the 
LSP RMSE observed when modifying /e/ to non-/e/ vowels 
decreases from 0.853 (MR-FC) to 0.782 (MR-FF) and 0.614 
(MR-TF) respectively. The LSP RMSEs for the MR-TF system 
are almost the same as for the Ref. system, which means that 
the target vowels can be synthesised as accurately as with 
standard HMM-based speech synthesis by modifying the /e/ 
source vowels using appropriate articulatory inputs. Compar-
ing Fig. 8 a) and Fig. 8 b), we find that the performance 
of all the MRHMM-based systems degrades when the natural 
articulatory features are replaced with the generated ones. This 
means the appropriateness of input articulatory features plays 
an important role in our proposed method. The performance 
of the HMM-based articulatory movement prediction method 
used in this experiment still needs improvement because the 
generated trajectories are over-smoothed, due to the aver-
aging effects of HMM modelling and parameter generation 
algorithm. A detailed analysis on this articulatory movement 
prediction method can be found in [32]. In the MR-TF system, 
the average RMSE of EMA feature prediction is 1.107 mm 
and the average correlation coefficient between the natural 
and the predicted EMA features is 0.8037. We also examined 
the relationship between EMA prediction error and the LSP 
RMSE of the MR-TF system for different phones. The results 
are shown in Fig. 9. We can observe a positive correlation 
between these two error types, with a correlation coefficient 
of 0.431. Therefore, improving the accuracy of articulatory 
movement prediction is essential in order to achieve better 
controllability over the synthetic speech.

3) Subjective evaluation: In addition to using the objective 
error metrics described above, we have also conducted forced-
choice listening tests to evaluate performance on the vowel 
modification task subjectively. Six groups of systems were 
compared, and the definition of the systems in each group 
is presented in Fig. 10. Fifteen sentences were selected from 
the test set and synthesised by both systems in each test group. 
Each of these pairs of synthetic sentences were evaluated in 
random order by at least twenty native English listeners in 
listening booths. The listeners were asked to identify which 
sentence in each pair sounded more natural. We calculated 
the average preference scores with 95% confidence intervals 
for the six pairs of systems and Fig. 10 shows the results in 
detail. From Figs. 10 a) and b), we see that the naturalness 
of the UNI-FC and MR-FC systems is much worse than that 
of the Ref. system, which means the modification from /e/ 
to non-/e/ vowels is not achieved ideally in our baseline 
system. However, Fig. 10 c) shows that there is no significant 
difference in naturalness between the Ref. system and the MR-
TF system with natural articulatory inputs. The effectiveness
of our proposed methods, including feature-space regression matrix switching and task-specific context feature tailoring, are proved by Figs. 10 d) and e) respectively. However, using generated articulatory features degrades the naturalness of the MR-TF system significantly as shown in Fig. 10 f). These findings are consistent with the conclusions drawn from the objective evaluation results shown in Fig. 8.

C. Vowel creation task

Having identified the proposed MR-TF system as the best system in the first task, we devised a second task to further demonstrate the controllability offered by this system. This was a vowel creation task, whereby the aim was to create a new vowel without observing acoustic data for it in the training data set. This is potentially useful for applications such as building voices for different accents of a language, or cross-language speaker adaptation, for example. We simulated the scenario of vowel creation by selecting a target vowel from the English phone set and removing all sentences containing this target vowel from the training set. Vowel /ɔ/ was selected as the target vowel in our experiment, and 809 sentences in the database which contain no instances of this vowel were selected for training. 50 sentences were selected randomly from the remaining 454 sentences to form a test set. The STD-F and MR-TF systems listed in Table II were trained using this specially designed training set. In the MR-TF system, the task-specific context feature tailoring was conducted in the same way as for the vowel modification task. Again, the optimum number of GMM mixture components was identified using the minimum description length criterion, and this was found to be 64 components.

The sentences in the test set were synthesised using the two systems. For the MR-TF system, both natural and generated articulatory features were evaluated. The HMM-based articulatory prediction model used in the vowel identity modification task, which was trained using the full database and full context features, was reused here. Acoustic feature prediction error for different types of phone was calculated and is shown in Fig. 11. From this figure, we see that the STD-F system has much higher LSP RMSE for /ɔ/ than for the other vowels and consonants, because the acoustic data for /ɔ/ was not available during training. In contrast, the MR-TF system can predict the acoustic features of the /ɔ/ vowel much more accurately, even though the acoustic features of this vowel were unseen at training time. This is an important and very promising result that clearly demonstrates the flexibility of the proposed model. Its accuracy at predicting LSP features for other vowels and consonants is very close to that of the STD-F method when the natural articulatory features are given. Similar to the observations made in the vowel identity modification task, using generated articulatory movements degrades the accuracy of the MR-TF system at predicting acoustic features.

A vowel identity perception test was also carried out to further evaluate the effectiveness of creating the target /ɔ/ vowel. Five monosyllabic words (“but”, “hum”, “puck”, “tun”, “dud”) containing the /ɔ/ vowel were selected and embedded within the carrier sentence “Now we’ll say ... again”. These sentences were synthesised using the STD-F system and the MR-TF system respectively. Because recordings of natural articulatory movements for these sentences were not available, the articulatory features generated from the HMM-based articulatory prediction model were adopted as an alternative in the acoustic feature generation procedure of the MR-TF system. For the purpose of comparison, we substituted the vowel /ɔ/ in the five monosyllabic words with /ɛ/, /ɪ/ and /ə/, and then synthesised the respective test sentences using the STD-F system. Thus, we created twenty-five stimuli for the vowel identity perception test. Thirty-two native English listeners were asked to listen to these stimuli and to write down the key word in the carrier sentence they heard. Then, we calculated the percentages for how the vowels were perceived. These results are shown in Fig. 12. We see that only 35% of the synthesised vowels /ɔ/ were perceived correctly using the STD-F system, due to the lack of acoustic training samples for this vowel. This percentage is above chance level because

---

6Again, samples of the synthetic speech used in the vowel creation experiment are available at http://staff.ustc.edu.cn/~zhling/MRHMM-EMA/demo.html.
the phonetic characteristics of the /s/ vowel were still taken into account when designing the question set for the decision-tree-based model clustering during context-dependent model training. Using the MR-TF system and the generated articulatory features, this percentage increased to 66.25%, which is close to the perception accuracy of synthesising vowel /r/ (68.75%) and /æ/ (66.25%) using the STD-F system. Again, this demonstrates the MR-TF system is able to generate a new vowel accurately from appropriate articulator settings, which further proves the flexibility of the articulatory control offered by this system.

V. CONCLUSION

In this paper, we have presented an improved acoustic modelling method for imposing articulatory control over HMM-based parametric speech synthesis. In contrast to the unified acoustic-articulatory modelling used in our previous work, we have employed the framework of the multiple regression HMM to model the influence of the articulatory features on the generation of acoustic features. In this way, the articulatory features can be predicted using a separate articulatory prediction model, in which it is easier to integrate phonetic knowledge than with an HMM. A method involving feature-space regression matrix switching and a strategy of task-specific context feature tailoring has been proposed to improve the performance of the conventional MRHMM in dealing with the manipulated articulatory features. We have used a database with parallel waveform and EMA data in our experiments to evaluate this novel approach. Our results have shown the proposed method can achieve better control in vowel identity modification than the unified acoustic-articulatory modelling with full context features and context-dependent transform tying. Furthermore, our experiments have proved this method is effective in creating a new vowel, for which there are no acoustic samples in the training set, from appropriate articulatory features.

So far, our experiments have focussed on either modifying or creating isolated vowels. To apply the proposed framework to control the characteristics of synthetic speech at the word, sentence, or speaker level is in principle also possible, though there are certain issues that must be addressed in order to do so. First, for example, is the relationship between those speech characteristics we wish to control and the articulatory features that are available; in order to control some aspect of speech, it must be readily represented in terms of the features available. A second important prerequisite is an adequate module for articulatory movement prediction. Not only must this generate articulator trajectories that are plausible and accurate for whole utterances, but also it must allow convenient control to change the generated trajectories. As a final example, when attempting to impose articulatory control over longer spans, the issue of maintaining synchrony between the states of the acoustic models and the externally generated articulatory inputs becomes more prominent. As discussed in Section II.C, the HMM-based articulatory movement prediction method used in the experiments here is not convenient for sophisticated or extensive articulatory manipulation. Therefore, to investigate better models for articulator movement prediction will be a key task in our future work. Preliminary results of ongoing work in this direction have been presented in [33], where a target-filtering approach was adopted to predict the trajectories of articulator movements. This will help move us closer to our ultimate goal, which is to apply the current articulatory control approach to practical scenarios such as cross-accent speaker adaptation (e.g. changing a British English accent to an American one) and simulating Lombard effects in synthesised speech in response to environmental noise conditions.

REFERENCES


Zhen-Hua Ling received the B.E. degree in electronic information engineering, M.S. and Ph.D. degree in signal and information processing from University of Science and Technology of China, Hefei, China, in 2002, 2005, and 2008 respectively. From October 2007 to March 2008, he was a Marie Curie Fellow at the Centre for Speech Technology Research (CSTR), University of Edinburgh, U.K.. From July 2008 to February 2011, he was a joint postdoctoral researcher at University of Science and Technology of China and iFLYTEK Co., Ltd., China. He is currently an associate professor at University of Science and Technology of China. His research interests include speech synthesis, voice conversion, speech analysis, and speech coding. He was awarded IEEE Signal Processing Society Young Author Best Paper Award in 2010.

Junichi Yamagishi is a senior research fellow and holds a prestigious EPSRC Career Acceleration Fellowship at the Centre for Speech Technology Research (CSTR) at the University of Edinburgh. He was awarded a Ph.D. by Tokyo Institute of Technology in 2006 for a thesis that pioneered speaker-adaptive speech synthesis and was awarded the Tegima Prize as the best Ph.D. thesis of Tokyo Institute of Technology in 2007. Since 2006, he has been at CSTR and has authored and co-authored around 100 refereed papers in international journals and conferences. His work has led directly to three large-scale EC FP7 projects and two collaborations based around clinical applications of this technology. He was awarded the Itakura Prize (Innovative Young Researchers Prize) from the Acoustical Society of Japan for his achievements in adaptive speech synthesis. He is an external member of the Euan MacDonald Centre for Motor Neurone Disease Research and the Anne Rowling Regenerative Neurology Clinic in Edinburgh.

Korin Richmond has been involved with human language and speech technology since 1991. This began with an M.A. degree at Edinburgh University, reading Linguistics and Russian (1991-1995). He was subsequently awarded an M.Sc. degree in Cognitive Science and Natural Language Processing from Edinburgh University in 1997, and a Ph.D. degree at the Centre for Speech Technology Research (CSTR) in 2002. This Ph.D. thesis (“Estimating Articulatory Parameters from the Acoustic Speech Signal”), applied a flexible machine-learning framework to corpora of acoustic-articulatory data, giving an inversion mapping method that surpasses all other methods to date.

As a research fellow at CSTR for over ten years, his research has broadened to multiple areas, though often with emphasis on exploiting articulation, including: statistical parametric synthesis (e.g. Researcher Co-Investigator of EPSRC-funded “ProbTTS” project); unit selection synthesis (e.g. implemented the “MULTISYN” module for the FESTIVAL 2.0 TTS system); and lexiconography (e.g. jointly produced “COMBILEX”, an advanced multi-accent lexicon, licensed by leading companies and universities worldwide). He has also contributed as a core developer of CSTR/CMU’s Festival and Edinburgh Speech Tools C++ library since 2002. Dr. Richmond’s current work aims to develop ultrasound as a tool for child speech therapy.

Dr. Richmond is a member of ISCA and IEEE, and serves on the Speech and Language Processing Technical Committee of the IEEE Signal Processing Society.