Feature-space transform tying in unified acoustic-articulatory modelling of articulatory control of HMM-based speech synthesis

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Proc. Interspeech

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.
Feature-Space Transform Tying in Unified Acoustic-Articulatory Modelling for Articulatory Control of HMM-based Speech Synthesis

Zhen-Hua Ling¹, Korin Richmond², Junichi Yamagishi²

¹iFLYTEK Speech Lab, University of Science and Technology of China, P.R.China
²CSTR, University of Edinburgh, United Kingdom
zhling@ustc.edu, korin@cstr.ed.ac.uk, jyamagis@inf.ed.ac.uk

Abstract

In previous work, we have proposed a method to control the characteristics of synthetic speech flexibly by integrating articulatory features into hidden Markov model (HMM) based parametric speech synthesis. A unified acoustic-articulatory model was trained and a piecewise linear transform was adopted to describe the dependency between these two feature streams. The transform matrices were trained for each HMM state and were tied based on each state’s context. In this paper, an improved acoustic-articulatory modeling method is proposed. A Gaussian mixture model (GMM) is introduced to model the articulatory space and the cross-stream transform matrices are trained for each Gaussian mixture instead of context-dependently. This means the dependency relationship can vary with the change of articulatory features flexibly. Our results show this method improves the effectiveness of control over vowel quality by modifying articulatory trajectories without degrading naturalness.

Index Terms: speech synthesis, articulatory features, hidden Markov model, Gaussian mixture model

1. Introduction

The hidden Markov model (HMM)-based parametric speech synthesis method has made significant progress in recent years [1, 2]. This method is able to synthesize highly intelligible and smooth speech sounds [3, 4]. In our previous work, we have proposed a method to improve the flexibility of HMM-based speech synthesis by integrating articulatory features [5, 6]. Here, we use “articulatory features” to refer to the continuous movements of a group of speech articulators, such as the tongue, jaw, lips and velum, recorded by human articulography techniques. In this method, a unified acoustic-articulatory model with cross-stream dependency is trained. During synthesis, the characteristics of synthetic speech can be controlled flexibly by modifying the generated articulatory features according to arbitrary phonetic rules. Experimental results have shown the effectiveness of this method in controlling the overall character of synthesized speech and the quality of specific vowels [6].

A piecewise linear transform was used to describe the dependency of acoustic feature production on the movement of articulatory features in our previous work [5, 6]. Like other model parameters in the unified acoustic-articulatory HMM, the transform matrices were trained for each HMM state and were tied based on context using a decision tree. Therefore, the cross-stream dependency was entirely determined by the context information of input text. This could become problematic when the articulatory features were modified using phonetic rules during synthesis because the transform matrix was expected to adapt to the new articulatory configuration. In this paper, a feature-space transform tying method is proposed to solve this problem. A Gaussian mixture model (GMM) is adopted to model the articulatory space and the cross-stream transform matrices are estimated for each Gaussian component instead of for each HMM state (thus depending on its context information). This paper is organized as follows. Section 2 gives a brief overview of our baseline acoustic-articulatory modelling method. Section 3 describes our proposed method in detail. Section 4 introduces the results of our experiments and Section 5 presents the conclusions we draw from this work.

2. Baseline

In our baseline method, the general framework of HMM-based speech synthesis was followed to integrate articulatory features into the conventional modelling of acoustic features [6]. 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 sequence of the same length $T$. For each frame, the feature vector $x_t \in \mathbb{R}^{D_X}$ and $y_t \in \mathbb{R}^{D_Y}$ 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. In model training, an HMM $\lambda$ is estimated by maximizing the likelihood function of the joint distribution $P(X, Y | \lambda)$. A piecewise (state-wise) linear transform is added to the model parameters to represent the dependency between the generation of acoustic features and the articulatory movements. The joint distribution can be written as

$$
P(X, Y | \lambda) = \sum_q \sum_{a_0} \sum_{b_0} \sum_{\pi} \sum_{A_q} \prod_{t=1}^{T} \alpha_{a_t-1 \rightarrow a_t}(x_t, y_t),$$

(1)

$$
b_j(x_t, y_t) = b_j(x_t^T, y_t) = \begin{cases} b_j(x_t, y_t) = b_j(x_t^T, y_t) = N(x_t; \mu_j, \Sigma_j), \\
b_j(x_t, y_t) = b_j(x_t^T, y_t) = N(x_t; \mu_j, \Sigma_j), \end{cases}$$

(2)

(3)

(4)

where $q = \{q_1, q_2, ..., q_N\}$ is the state sequence shared by the two feature streams; $a_t$ and $a_t$ represent initial state probability and state transition probability; $b_j(\cdot)$ is the state observation probability density function (PDF) for state $j$; $N(\cdot; \mu, \Sigma)$ denotes a Gaussian distribution with a mean vector $\mu$ and a covariance matrix $\Sigma$; $A_q \in \mathbb{R}^{D_X \times D_Y}$ is the linear transform matrix for state $j$. This matrix is context-dependent, hence a globally piecewise linear transform can be achieved. The model parameters can be estimated using the EM algorithm [6].

During synthesis, the acoustic and articulatory features are simultaneously generated from the trained models using maximum-likelihood parameter generation (MLPG) algorithm that considers explicit constraints of the dynamic features. In order to control the characteristics of synthetic speech flexibly, the generated articulatory features can be modified based...
on phonetic knowledge to reproduce acoustic parameters that reflect those changes appropriately [6].

3. Proposed Method

3.1. Model Structure

As discussed above, the transform matrix $A_t$ in Eq.(4) is tied across states according to context information. This may lead to the incorrect representation of feature dependency when the generated articulatory features are modified during synthesis. In this paper, we improve the model structure so that the transform matrix can be determined by the articulatory features instead of the context information. Here, a GMM model $\lambda^{(G)}$ of $M$ mixtures is trained in advance using only the articulatory stream of training data to represent the articulatory space. Then, the transform matrices are trained for each mixture component of $\lambda^{(G)}$. Mathematically, we rewrite Eq.(4) as

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

$$= \sum_{k=1}^{M} \zeta_k(t) P(x_t|y_t, q_t = j, m_t = k, \lambda^{(G)}),$$

where $m_t$ denotes the mixture index of $\lambda^{(G)}$ for articulatory feature vector at frame $t$; the HMM state sequence $q$ and the GMM mixture sequence $m = \{m_1, m_2, ..., m_N\}$ are assumed to be independent, i.e.

$$P(m_t = k|y_t, q_t = j, \theta, \lambda^{(G)}) = \zeta_k(t).$$

For each Gaussian mixture, the dependency between the acoustic and articulatory features is represented as

$$P(x_t|y_t, q_t = j, m_t = k, \theta, \lambda^{(G)}) = \mathcal{N}(x_t; A_k \xi_k + \mu_{x_k}, \Sigma_{x_k}),$$

where $\xi_k = [y_t^T, 1]^T \in \mathbb{R}^{3D_Y+1}$ is the expanded articulatory feature vector and $A_k \in \mathbb{R}^{3D_x \times (3D_Y+1)}$ is the transform matrix for the $k$-th mixture of $\lambda^{(G)}$. Fig.1 compares the feature production models used in our baseline and proposed methods. We see that an extra Gaussian mixture sequence $m_t$ is introduced to determine the cross-stream transform matrix for each frame. We can interpret $\zeta_k(t)$ as a weight that varies according to $y_t$, and which changes how each transform matrix is weighted, or “blended” together, according to Eq.(6).

3.2. Model training

To train the HMM parameter set $\{A_k, \mu_{x_k}, \Sigma_{x_k}, \mu_{y_j}, \Sigma_{y_j}\}$, we substitute Eq.(2), (3), (6), (8) into Eq.(1) and get

$$P(X, Y|\lambda) = \sum_{q} \sum_{m} P(X, Y, q, m|\lambda),$$

where

$$P(X, Y, q, m|\lambda) = \prod_{t=1}^{T} a_{q_{t-1}, q_t} \zeta_{m_t}(t) \mathcal{N}(y_t; \mu_{y_{q_t}}, \Sigma_{y_{q_t}}) \mathcal{N}(x_t; A_{m_t} \xi_t + \mu_{x_{m_t}}, \Sigma_{x_{m_t}}).$$

The EM algorithm is adopted to estimate the parameter set that maximizes Eq.(9). The auxiliary function is defined as

$$Q(\lambda, \lambda') = \sum_{q} \sum_{m} \sum_{t} P(X, Y, q, m|\lambda) \log P(X, Y, q, m|\lambda')$$

$$= \sum_{j=1}^{N} \sum_{k=1}^{M} \sum_{t=1}^{T} \gamma_j(t) \zeta_k(t) \left( \log \mathcal{N}(y_t; \mu_{y_j}, \Sigma_{y_j}) + \log \mathcal{N}(x_t; A_k \xi_t + \mu_{x_k}, \Sigma_{x_k}) \right) + K,$$

where $K$ is a constant term that is independent of the model parameter set; $\gamma_j(t)$ is the occupancy probability of state $j$ at time $t$; $N$ is the total number of HMM states.

In order to re-estimate the transform matrix $A_{k_t}$ for each GMM mixture, we set $\partial Q(\lambda, \lambda') / \partial A_{k_t} = 0$ and get

$$\sum_{j=1}^{N} \sum_{t=1}^{T} \gamma_j(t) \zeta_k(t) \zeta_k(t) \Sigma_{x_j}^{-1} (x_t - \mu_{x_j}) \xi_t^T$$

$$= \sum_{j=1}^{N} \sum_{t=1}^{T} \gamma_j(t) \zeta_k(t) \Sigma_{x_j}^{-1} A_{k_t} \xi_t \xi_t^T.$$ (13)

This equation can be simplified as

$$Z = \sum_{t=1}^{T} V(t) A_{k_t} D(t),$$

where

$$Z = \{z_{it}\} = \sum_{j=1}^{N} \sum_{t=1}^{T} \gamma_j(t) \zeta_k(t) \Sigma_{x_j}^{-1} (x_t - \mu_{x_j}) \xi_t^T,$$ (15)

$$V(t) = \text{diag} \left\{ v_{t_i}^{(t)} \right\} = \sum_{j=1}^{N} \gamma_j(t) \Sigma_{x_j}^{-1}.$$ (16)
For the \( \sum \) be derived by setting synthetic speech. The detailed steps are introduced as follows:

1. The re-estimation formulae for other model parameters can be derived by setting \( \partial Q(\lambda, \Lambda)/\partial \lambda = 0 \) as

   \[
   \mu'_{X_j} = \frac{1}{\sum_{l=1}^{T} \gamma_j(t)} \sum_{l=1}^{T} \gamma_j(t) (x_t - A_k^j x_l) (x_t - A_k^j x_l)^\top, \tag{20}
   \]

   \[
   \Sigma'_{X_j} = \frac{1}{\sum_{l=1}^{T} \gamma_j(t)} \sum_{l=1}^{T} \gamma_j(t) \mu'_j (x_t - \mu'_j) (x_t - \mu'_j)^\top, \tag{21}
   \]

   \[
   \mu'_{Y_j} = \frac{1}{\sum_{l=1}^{T} \gamma_j(t)} \sum_{l=1}^{T} \gamma_j(t) (y_t - A_k^j y_l) (y_t - A_k^j y_l)^\top, \tag{22}
   \]

   \[
   \Sigma'_{Y_j} = \frac{1}{\sum_{l=1}^{T} \gamma_j(t)} \sum_{l=1}^{T} \gamma_j(t) \mu'_j (y_t - \mu'_j) (y_t - \mu'_j)^\top. \tag{23}
   \]

3.3. Parameter generation with articulatory control

Similar to our previous work [6], the maximum likelihood criterion is adopted and only the optimal HMM state sequence is considered in the parameter generation. The generated articulatory features can be modified to control the characteristics of synthetic speech. The detailed steps are introduced as follows:

1) Generate the optimal state sequence \( \mathbf{q}^* \) using the trained duration distributions [2].

2) Generate the optimal articulatory features \( \mathbf{Y}^* \). In order to simplify the calculation, only the articulatory stream in the HMM is used, i.e., to maximize

   \[
   P(\mathbf{Y}^* | \lambda, \mathbf{q}^*) \approx \prod_{t=1}^{T} N(y_t; \mu_{Y,t}, \Sigma_{Y,t}). \tag{24}
   \]

   This can be solved using the conventional maximum likelihood parameter generation (MLPG) algorithm [1].

3) Modify the articulatory features by designing function \( f(\cdot) \) based on phonetic rules and get \( \tilde{Y} = f(Y^*) \).

4) Generate the optimal acoustic features \( \mathbf{X}^* \) according to the modified articulatory features by maximizing

   \[
   P(\mathbf{X}^* | \tilde{Y}, \lambda, \mathbf{q}^*) = \prod_{t=1}^{M} \sum_{k=1}^{K} \zeta_k(t) N(x_t; A_k \mathbf{X}_{t-1}^* + \mathbf{X}_{k^*}, \Sigma_{X,k^*}). \tag{25}
   \]
labic words (“bet”, “hem”, “peck”, “ten”, “dead”) with vowel /r/ were selected and embedded into the carrier sentence “Now we’ll say ... again”. We tried to modify the vowel /r/ to be perceived as the vowels /l/ and /n/ by manipulating the generated articulatory trajectory during synthesis. In our previous work [6], we only modified the EMA dimensions corresponding towards /l/ by the context information beforehand. This work is partially funded by the National Nature Science Foundation of China (Grant No. 60905010). 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 grant EP/I027696/1.

6. References


