文章编号:1004-9037(2014)02-0204-07

基于隐马尔科夫模型的中文发音动作参数预测方法

蔡明琦 凌震华 戴礼荣

(中国科学技术大学电子工程与信息科学系,合肥,230027)

摘要:发音动作参数描述发音过程中唇、舌、颚等发音器官的位置与运动。本文对给定文本与语音情况下中文发音动作参数的预测方法进行了研究。首先,设计并实现了基于电磁发音仪的发音动作参数采集与预处理方法,通过头部运动规整与咬合面规整保证了发音动作参数的可靠性;其次,将隐马尔科夫模型应用于中文发音动作参数预测,采用包含声学参数与发音动作参数的双流模型结构实现从声学参数到发音动作参数的映射,并且分析对比了建模过程中不同上下文属性、模型聚类方式以及流间相关性假设对于中文发音动作参数预测性能的影响。实验结果表明,当采用三音素模型、双流独立聚类并且考虑流间相关性的情况下,可以获得最优的预测性能。

关键词:隐马尔科夫模型;发音器官;电磁发音仪;发音动作参数

中图分类号:TP391 **文献标志码**:A

Hidden-Markov-Model-Based Articulatory Movement Prediction for Chinese

Cai Mingqi, *Ling Zhenhua*, *Dai Lirong*

(Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, 230027, China)

Abstract: Articulatory features represent the quantitative positions and continuous movements of articulators during the production of speech. These articulators include the tongue, lips, jaw, velum and so on. This paper presents an investigation into articulatory feature prediction for Chinese when text and audio inputs are given. First, a method of recording and preprocessing articulatory features captured by electromagnetic articulography (EMA) is designed. By head movement and occlusal surface normalization, the reliability of articulatory features is guaranteed. Then, unified acoustic-articulatory hidden Markov models (HMMs) are introduced to predict Chinese articulatory features and achieve the inversion mapping from acoustic to articulatory features. Several aspects of this method are analyzed, including the effectiveness of context-dependent modeling, the difference among model clustering methods and the influence of cross-stream dependency modeling. The results show that the optimized performance is achieved using unified acoustic-articulatory triphone HMMs with separate clustering of acoustic and articulatory model parameters and a dependent-feature model structure.

Key words: hidden Markov model (HMM); articulatory organ; electromagnetic articulography (EMA); articulatory features

引 言

语音是从肺部呼出的气流通过声门、声道等各种器官作用而发出的。声道的形状主要由唇、颚、舌等的位置决定。不同的声道形状决定了不同的发音[1]。人们用发音动作参数描述发音器官在发音过程中的位置及运动,这些发音器官包括舌、下颚、嘴唇等。发音动作参数可以通过多种技术来采集,例如X射线微束影像[2]、磁共振成像[3]、超声波[4]、图像采集外部发音器官运动[5]及电磁发音仪(Electro magnetic articulography, EMA)[6]等。

基金项目:国家自然科学基金青年基金(60905010)资助项目。
收稿日期:2012-05-15;**修订日期**:2012-08-16

发音动作参数不仅可以有效地描述语音特征，而且相对于声学参数还具有以下优势：

(1)因为发音器官的物理运动能力有限，所以发音动作参数相对于声学参数变化缓慢且平滑，更适合使用隐马尔科夫模型(Hidden Markov model，HMM)进行建模。

(2)对语音中存在的某些现象，发音动作参数可以进行更直接的解释。例如，语音中的第二共振峰从高到低的变化，可以通过发音动作参数解释为舌位从前往后的运动。

(3)发音动作参数直接记录发音器官的位置，它们不受声学噪音的影响且较少受录音环境的影响。因此发音动作参数相对于声学参数更加鲁棒[7]。

基于发音动作参数的以上优点，已有研究人员将发音动作参数应用到语音识别与语音合成的方法研究中，例如将发音动作参数作为语音识别的额外特征参数以降低识别错误率[8]，在语音合成中融合发音动作参数以提高合成语音的自然读与灵活可控性[9]等。

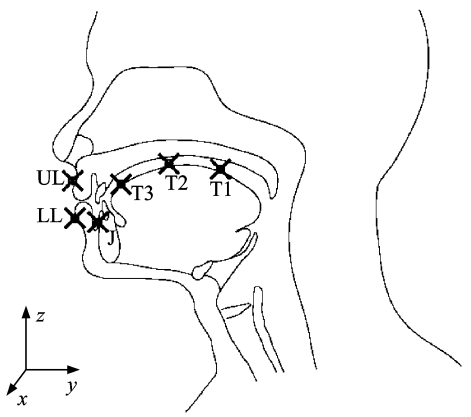
此外，在给定文本或者语音输入时的发音动作参数预测也是发音动作参数研究的热点之一，其潜在的应用场景包括语音驱动的人脸动画系统、语言学习中的发音位置问题检测、基于调音的语音合成方法中的发音器官运动预测等。目前发音动作参数预测方法按照输入主要分为两类：(1)输入文本：利用时间对齐的音素序列及高斯分布描述音素中点发音动作参数的分布，通过一个协同发音模型预测发音动作参数[10]；利用目标逼近模型进行发音动作参数预测[11]；基于 HMM 的发音动作参数预测[12]。(2)输入语音：基于高斯混合模型的声学-发音动作参数映射，并使用最大似然估计准则考虑动态参数[13]；利用人工神经网络和最大似然参数生成(Maximum likelihood parameter generation，MLPG)算法训练一个轨迹模型[14]。由于缺少中文发音动作参数数据库，目前少有对中文发音动作参数的研究。

本文对基于 HMM 的中文发音动作参数预测方法进行研究。在模型训练阶段，利用电磁发音仪完成了中文连续语流的发音动作参数采集、处理与数据库制作，构建了包含声学与发音动作参数的双流 HMM 模型来表征两种参数之间的关系[12]；在预测阶段，利用输入的文本及声学参数，基于最大似然准则实现发音动作参数的预测。此外，本文还研究了建模过程中不同的上下文属性、模型聚类方式、流间相关性假设以及转换矩阵绑定方式对于中文发音动作参数预测性能的影响。

1　中文连续语流 EMA 数据库构建

利用 EMA 可以便捷、准确、实时地采集发音动作参数。本文采用 NDI 公司的 Wave System 设备录制中文发音人连续语流的发音动作参数及语音波形，并经过预处理制作成中文连续语流 EMA 数据库。由于使用 EMA 采集发音动作参数，因此后续介绍中“发音动作参数”也用“EMA 参数”来表示。

本文设计的中文数据库包括音素平衡的 390 句中文语句，由一名普通话女发音人在隔音密闭专业录音室里采用 AKG 领夹式麦克风朗读录制。使用 NDI 公司的 Wave System 设备平行录制语音波形与 EMA 参数。波形录制使用 16 kHz 采样，16 bit 量化的 PCM 格式。通过在发音人的各发音器官放置小的传感器，并利用电磁信号对发音过程中各传感器进行定位来实现 EMA 数据的采集。实验中分别在感兴趣的 6 个发音器官位置放置了传感器，其位置如图 1 所示。利用 Wave System 设备，可以采集每个传感器在发音过程中的空间三维位置。



UL—上唇; LL—下唇; J—下颚; T3—舌尖; T2—舌中; T1—舌背。

图 1　EMA 传感器位置示意图

Fig. 1　Placement of EMA receivers in database

由于 EMA 参数是由 EMA 传感器直接记录的位置信息，在对 EMA 参数进行 HMM 建模前必须对其进行预处理。预处理主要分为两个步骤：头部运动规整和咬合面规整。

1.1　头部运动规整

原始的 EMA 数据记录的是发音器官相对于固定参考系的位置信息，而实际感兴趣的信息是发

音器官相对于发音人头部的运动信息。因此,需要对 EMA 数据进行规整以消除头部运动的影响。本文利用 NDI 公司 Wave System 提供的一个 6D 参考传感器,并将这个参考传感器放置在说话人鼻梁处(认为鼻梁在发音时始终与头部保持相对静止),可以较为便捷地得到其他传感器发音器官相对此传感器的头部规整后的 EMA 数据。

1.2 咬合面规整

将发音人牙齿自然咬合时所形成的平面定义为咬合面,如图 2 所示,在一块硬纸板上安置 A,B,C 三个传感器(直线 AB 垂直于 BC),让发音人自然咬住硬纸板来测量发音人的咬合面。咬合面规整就是将原始以鼻梁参考点为中心的 xyz 坐标系变换成 $x'y'z'$ 坐标系,其中 $x'y'$ 平面为咬合面、$y'z'$ 平面为垂直于咬合面的头部中轴面。利用咬合面对发音动作参数进行规整可以使发音动作参数物理意义更明显,并且可以较好保证不同发音人 EMA 参数的可比性。

做完头部运动规整的 EMA 数据,每个传感器分别有 x,y,z 三维数据,如图 1 所示,其中 x 表示左右方向位移、y 表示前后方向位移、z 表示上下方向位移。在图 2 中,假设 M 点为需要规整的点,T 为点 M 在咬合面的投影,S 为 TS 在直线 BC 上的垂足。将 MT,TS 的长度作为 z',y'的模。由于所有传感器均安置在发音人的头部中轴面上,所以 x'的模很小可以忽略。z',y'的正负符号信息由 BM 与咬合面的法向量及 AB 直线夹角决定。经过咬合面规整,每个传感器所对应 EMA 数据由三维降为两维。

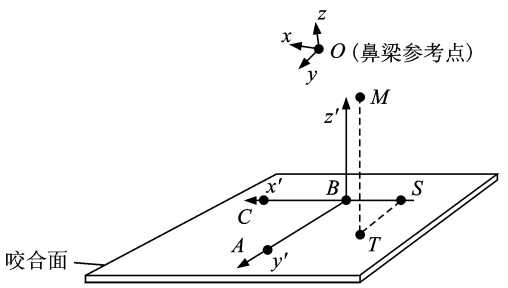


图 2 咬合面规整过程示意图

Fig. 2 Schematic diagram for occlusal surface normalization

2 应用 HMM 的中文发音动作参数预测

将 HMM 用于中文发音动作参数预测,其框架类似于基于 HMM 的参数语音合成系统[15]。首先需要训练统一的声学-发音动作参数 HMM 模型以表示声学参数与发音动作参数之间的关系;在生成过程中,利用最大似然准则和动态参数约束生成最优发音动作参数[12]。

2.1 发音动作参数预测方法

整个发音动作参数预测系统主要分为训练和预测两部分[7]。系统框架如图 3 所示。

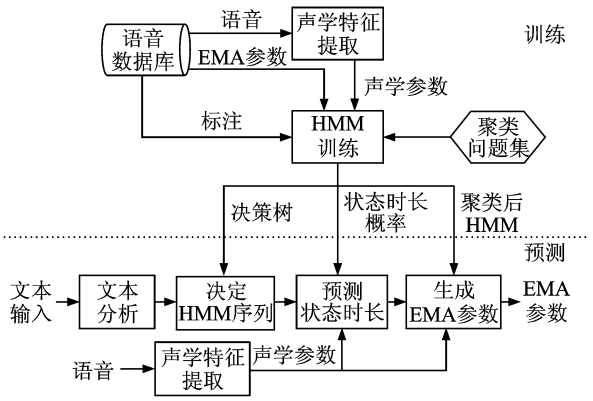


图 3 基于 HMM 的发音动作参数预测系统

Fig. 3 HMM-based articulatory movement prediction system

在训练过程中,通过最大化声学与发音动作参数的联合分布似然函数 $P(\boldsymbol{X},\boldsymbol{Y}|\lambda)$,得到一组上下文相关的 HMM 模型 λ,其中 $\boldsymbol{X}=[\boldsymbol{x}_1^{\mathrm{T}},\boldsymbol{x}_2^{\mathrm{T}},\cdots,\boldsymbol{x}_N^{\mathrm{T}}]^{\mathrm{T}}$ 和 $\boldsymbol{Y}=[\boldsymbol{y}_1^{\mathrm{T}},\boldsymbol{y}_2^{\mathrm{T}},\cdots,\boldsymbol{y}_N^{\mathrm{T}}]^{\mathrm{T}}$ 表示平行长度为 N 的发音动作特征与声学特征观测序列,$(\cdot)^{\mathrm{T}}$ 表示矩阵转置。每帧发音动作特征观测向量 $\boldsymbol{x}_t\in\mathbf{R}^{3D_X}$、声学特征观测向量 $\boldsymbol{y}_t\in\mathbf{R}^{3D_Y}$ 包含静态参数 $\boldsymbol{x}_{S_t}\in\mathbf{R}^{D_X}$,$\boldsymbol{y}_{S_t}\in\mathbf{R}^{D_Y}$ 与其一阶及二阶差分,如式(1~6)。

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

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

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

$$\boldsymbol{y}_t=[\boldsymbol{y}_{S_t}^{\mathrm{T}},\Delta\boldsymbol{y}_{S_t}^{\mathrm{T}},\Delta^2\boldsymbol{y}_{S_t}^{\mathrm{T}}]^{\mathrm{T}} \tag{4}$$

$$\Delta\boldsymbol{y}_{S_t}=0.5\boldsymbol{y}_{S_{t+1}}-0.5\boldsymbol{y}_{S_{t-1}} \tag{5}$$

$$\Delta^2\boldsymbol{y}_{S_t}=\boldsymbol{y}_{S_{t+1}}-2\boldsymbol{y}_{S_t}+\boldsymbol{y}_{S_{t-1}} \tag{6}$$

初始化上下文相关的 HMM 训练后,用最小描述长度(Minimum description length, MDL)准则和上下文属性问题集训练一棵决策树,利用该决策树对 HMM 进行聚类[16],这样可以解决由数据稀疏引起的过拟合问题。在对发音动作参数与声学参数进行基于决策树的模型聚类时,可以对两种参数分别构建决策树(独立聚类);也可以为这两种

参数构建一棵共享的决策树（共享聚类）。然后使用训练得到的上下文相关 HMM 进行状态切分并且训练状态的时长概率模型[17]。通过上述训练流程，最后训练得到的模型包括谱、基频、时长及发音动作参数的聚类 HMM 以及各自的决策树。

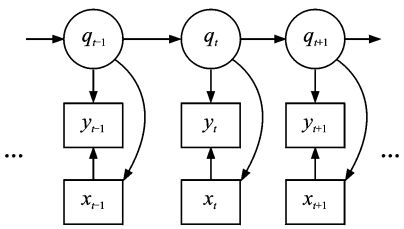
预测过程中，首先利用前端文本分析得到的结果和决策树确定 HMM 序列，然后利用 MLPG 算法生成最优发音动作参数[18]如下

$$\boldsymbol{X}_S^* = \arg\max_{\boldsymbol{X}_S} P(\boldsymbol{W}_X\boldsymbol{W}_S \mid \lambda, \boldsymbol{Y}) = \arg\max_{\boldsymbol{X}_S} \sum_{\forall q} P(\boldsymbol{W}_X\boldsymbol{W}_S, q \mid \lambda, \boldsymbol{Y}) \tag{7}$$

式中：$\boldsymbol{X}_S = [\boldsymbol{x}_{S_1}^{\mathrm{T}}, \boldsymbol{x}_{S_2}^{\mathrm{T}}, \cdots, \boldsymbol{x}_{S_N}^{\mathrm{T}}]^{\mathrm{T}}$ 为发音动作参数的静态参数；$\boldsymbol{W}_X \in \mathbf{R}^{3ND_X \times ND_X}$ 是发音动作参数的扩展矩阵，如式(1～3)；$q = \{q_1, q_2, \cdots, q_N\}$ 表示状态序列。

2.2　流间相关性建模

因为声学信号是由发音器官的运动引起的，所以声学参数与发音动作参数是彼此相关的。因此在对声学参数与发音动作参数建模时，应考虑这种相关性。根据发音的物理机制，本文选择采用状态同步系统[7]，状态同步系统假设声学参数和发音动作参数是由相同的状态序列生成的。在状态同步系统的基础上，对声学参数和发音动作参数之间的依赖关系进行直接建模。此时声学参数的生成不仅依赖于当前的上下文相关音素的声学模型，还依赖于当前帧对应的发音动作参数。特征生成模型结构如图 4 所示。



x—发音参数特征；y—声学特征；q—状态序列。

图 4　特征生成模型结构

Fig. 4　Feature production model for combined acoustic and articulatory modeling

在之前的工作中，作者采用一无偏置的线性变换来对声学参数与发音动作参数的依赖关系进行直接建模[9,12]。本文在此基础上改进为一有偏置的线性变换对声学参数与发音动作参数的依赖关系进行建模，并且考虑该线性变换的分回归类绑定以减少需要估计的模型参数数目。因此，声学参数与发音动作参数的联合分布可以写成

$$P(\boldsymbol{X}, \boldsymbol{Y} \mid \lambda) = \sum_{\forall q} P(\boldsymbol{X}, \boldsymbol{Y}, q \mid \lambda) = \sum_{\forall q} \pi_{q0} \prod_{t=1}^{N} a_{q_{t-1}q_t} b(\boldsymbol{x}_t, \boldsymbol{y}_t) \tag{8}$$

$$b_j(\boldsymbol{x}_t, \boldsymbol{y}_t) = b_j(\boldsymbol{x}_t) b_j(\boldsymbol{y}_t \mid \boldsymbol{x}_t) \tag{9}$$

$$b_j(\boldsymbol{x}_t) = N(\boldsymbol{x}_t; \mu_{\boldsymbol{X}_j}, \Sigma_{\boldsymbol{X}_j}) \tag{10}$$

$$b_j(\boldsymbol{y}_t \mid \boldsymbol{x}_t) = N(\boldsymbol{y}_t; \boldsymbol{A}_{\gamma_j} \xi_t + \mu_{\boldsymbol{Y}_j} \Sigma_{\boldsymbol{Y}_j} \tag{11}$$

式中：$q = \{q_1, q_2, \cdots, q_N\}$ 表示两种特征共享的状态序列；π_j 和 a_{ij} 分别表示初始状态概率和状态转移概率。$b_j(\cdot)$ 表示状态 j 的观测概率密度函数；$N(;\mu, \Sigma)$ 表示均值向量和协方差矩阵分别为 μ 和 Σ 的正态分布。$\boldsymbol{A}_{\gamma_j} \in \mathbf{R}^{3D_Y \times (3D_X+1)}$ 表示状态 j 对应的线性转换矩阵，其中 $\gamma_j \in \{1, 2, \cdots, K\}$ 表示状态 j 转换矩阵对应的回归类；$\xi_t = [\boldsymbol{x}_t^{\mathrm{T}}, 1]^{\mathrm{T}} \in \mathbf{R}^{3D_X+1}$ 表示时间 t 对应的扩展发音动作参数向量。本文可以采用期望最大化（Expectation maximization, EM）算法进行模型参数的估计[9]，如式(12～17)所示。

$$\hat{\mu}_{\boldsymbol{X}_j} = \frac{\sum_{t=1}^{T} \gamma_j(t) \boldsymbol{x}_t}{\sum_{t=1}^{T} \gamma_j(t)} \tag{12}$$

$$\hat{\Sigma}_{\boldsymbol{X}_j} = \frac{1}{\sum_{t=1}^{T} \gamma_j(t)} \cdot \sum_{t=1}^{T} \gamma_j(t) (\boldsymbol{x}_t - \hat{\mu}_{\boldsymbol{X}_j})(\boldsymbol{x}_t - \hat{\mu}_{\boldsymbol{X}_j})^{\mathrm{T}} \tag{13}$$

$$\hat{\mu}_{\boldsymbol{Y}_j} = \frac{\sum_{t=1}^{T} \gamma_j(t) (\boldsymbol{y}_t - \hat{\boldsymbol{A}}_{\gamma_j} \xi_t}{\sum_{t=1}^{T} \gamma_j(t)} \tag{14}$$

$$\hat{\Sigma}_{\boldsymbol{Y}_j} = \frac{\sum_{t=1}^{T} \gamma_j(t) \boldsymbol{V}_{tj} \boldsymbol{V}_{tj}^{\mathrm{T}}}{\sum_{t=1}^{T} \gamma_j(t)} \tag{15}$$

$$\boldsymbol{V}_{tj} = (\boldsymbol{y}_t - \hat{\boldsymbol{A}}_{\gamma_j} \xi_t - \hat{\mu}_{\boldsymbol{Y}_j}) \tag{16}$$

$$\sum_{j, \gamma_j = k} \sum_{t=1}^{T} \gamma_j(t) \Sigma_{\boldsymbol{Y}_j}^{-1} (\boldsymbol{y}_t - \mu_{\boldsymbol{Y}_j}) \xi_t^{\mathrm{T}} = \sum_{j, \gamma_j = k} \sum_{t=1}^{T} \gamma_j(t) \Sigma_{\boldsymbol{Y}_j}^{-1} \hat{\boldsymbol{A}}_k \xi_t \xi_t^{\mathrm{T}} \tag{17}$$

式中：$\gamma_j(t)$ 表示第 t 帧属于状态 j 的占有概率。利用式（17）求解更新后的转换矩阵 $\hat{\boldsymbol{A}}_k \in \mathbf{R}^{3D_Y \times (3D_X+1)}$，由于 Σ_Y^{-1} 是对角阵，因此等式右边相当于对转换矩阵 $\hat{\boldsymbol{A}}_k$ 按行进行加权，所以转换矩阵 $\hat{\boldsymbol{A}}_k$ 每行之间是独立的，可以按行来求解转换矩阵 $\hat{\boldsymbol{A}}_k$。

2.3 参数生成及迭代更新

发音动作参数生成公式见式(7),下面简化这一优化过程,只考虑发音动作参数在最优状态序列下的情况,因此式(7)可简化为式(18)

$$[\boldsymbol{X}_S^*, q^*] \approx \arg\max_{\boldsymbol{X}_S, q} P(\boldsymbol{W}_X\boldsymbol{W}_S, q \mid \lambda, \boldsymbol{Y}) \quad (18)$$

采用迭代更新方法来交替更新发音动作参数与状态序列,每一次迭代包括两步[12]:

(1)在给定声学特征 $\boldsymbol{Y}$ 与状态序列 q 的情况下,优化发音动作参数 $\boldsymbol{X}_S$。

$$\boldsymbol{X}_{S_i}^* = \arg\max_{\boldsymbol{X}_S} P(\boldsymbol{W}_X\boldsymbol{X}_S \mid \lambda, q_{i-1}, \boldsymbol{Y}) = \arg\max_{\boldsymbol{X}_S} P(\boldsymbol{W}_X\boldsymbol{X}_S, \boldsymbol{Y} \mid \lambda, q_{i-1}) \quad (19)$$

式中:$i \in (1,2,\cdots)$表示第 i 次迭代,q_0 表示利用一个纯声学特征模型用 Viterbi 对齐算法对声学特征序列 $\boldsymbol{Y}$ 切分出的初始状态序列。如果假设 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 在给定状态序列下没有依赖关系,采用传统的 MLPG 算法可以直接求解式(19)。一旦在建模时考虑声学参数与发音动作参数之间的依赖关系,如式(11)和式(19)中的联合分布可以写成式(20)。

$$\begin{aligned}\log P(\boldsymbol{W}_X\boldsymbol{X}_S, \boldsymbol{Y} \mid \lambda, q_{i-1}) = {} & \boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{W}_X^{\mathrm{T}}\boldsymbol{U}_X^{-1}\boldsymbol{M}_X - \\ & \frac{1}{2}\boldsymbol{X}_S^{\mathrm{T}}\boldsymbol{W}_X^{\mathrm{T}}\boldsymbol{U}_X^{-1}\boldsymbol{W}_X\boldsymbol{X}_S + \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{U}_Y^{-1}\boldsymbol{A}\,\xi + \\ & \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{U}_Y^{-1}\boldsymbol{M}_Y - \frac{1}{2}\xi^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{U}_Y^{-1}\boldsymbol{A}\,\xi + \\ & \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{U}_Y^{-1}\boldsymbol{Y}) - \xi^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{U}_Y^{-1}\boldsymbol{M}_Y + K \end{aligned} \quad (20)$$

其中

$$\boldsymbol{U}_X^{-1} = \mathrm{diag}[\Sigma_{X_{q1}}^{-1}, \Sigma_{X_{q2}}^{-1}, \cdots, \Sigma_{X_{qN}}^{-1}] \quad (21)$$

$$\boldsymbol{M}_X = [\mu_{X_{q1}}^{\mathrm{T}}, \mu_{X_{q2}}^{\mathrm{T}}, \cdots, \mu_{X_{qN}}^{\mathrm{T}}]^{\mathrm{T}} \quad (22)$$

$$\boldsymbol{U}_Y^{-1} = \mathrm{diag}[\Sigma_{Y_{q1}}^{-1}, \Sigma_{Y_{q2}}^{-1}, \cdots, \Sigma_{Y_{qN}}^{-1}] \quad (23)$$

$$\boldsymbol{M}_Y = [\mu_{Y_{q1}}^{\mathrm{T}}, \mu_{Y_{q2}}^{\mathrm{T}}, \cdots, \mu_{Y_{qN}}^{\mathrm{T}}]^{\mathrm{T}} \quad (24)$$

$$\boldsymbol{A} = \mathrm{diag}[\boldsymbol{A}_{\gamma q_1}, \boldsymbol{A}_{\gamma q_2}, \cdots, \boldsymbol{A}_{\gamma q_N}] \quad (25)$$

$$\xi = \{\xi_{q_1}^{\mathrm{T}}, \xi_{q_2}^{\mathrm{T}}, \cdots, \xi_{q_N}^{\mathrm{T}}]^{\mathrm{T}} \quad (26)$$

式中:K 为常数项。由式(26),$\xi_t = [\boldsymbol{x}_t^{\mathrm{T}}, 1]^{\mathrm{T}}$,$\boldsymbol{A}_{\gamma_j}\xi_t = \widetilde{\boldsymbol{A}}_{\gamma_j}\boldsymbol{x}_t + b_j$ 及 $\boldsymbol{A}_{\gamma_j} = [\widetilde{\boldsymbol{A}}_{\gamma_j}, b_j]$,可以得到

$$\boldsymbol{A}\xi = \widetilde{\boldsymbol{A}}\boldsymbol{X} + \boldsymbol{B} \quad (27)$$

式中 $\widetilde{\boldsymbol{A}} = \mathrm{diag}[\widetilde{\boldsymbol{A}}_{\gamma q_1}, \widetilde{\boldsymbol{A}}_{\gamma q_2}, \cdots, \widetilde{\boldsymbol{A}}_{\gamma q_N}]$,$\boldsymbol{B} = [\boldsymbol{b}_{q1}^{\mathrm{T}}, \boldsymbol{b}_{q2}^{\mathrm{T}}, \cdots, \boldsymbol{b}_{qN}^{\mathrm{T}}]^{\mathrm{T}}$。因此可以将式(20)扩写,并且设 $\partial P(\boldsymbol{W}_X^{\mathrm{T}}\boldsymbol{X}_S, \boldsymbol{Y} \mid \lambda, q_{i-1})/\partial \boldsymbol{X}_S = 0$,可以求解出最优发音动作参数,见式(28)。

$$\boldsymbol{X}_{S_i}^* = (\boldsymbol{W}_X^{\mathrm{T}}(\boldsymbol{U}_X^{\mathrm{T}} + \widetilde{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{U}_Y^{-1}\widetilde{\boldsymbol{A}})\boldsymbol{W}_X)^{-1} \cdot \boldsymbol{W}_X^{\mathrm{T}}(\boldsymbol{U}_X^{-1}\boldsymbol{M}_X + \widetilde{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{U}_Y^{-1}(\boldsymbol{Y} - \boldsymbol{M}_Y - \boldsymbol{B})) \quad (28)$$

(2)给定 $\boldsymbol{X}_S^*$ 和 $\boldsymbol{Y}$ 优化状态序列 q

$$q_i^* = \arg\max_q P(q \mid \lambda, \boldsymbol{W}_X\boldsymbol{X}_{S_i}^*, \boldsymbol{Y}) \quad (29)$$

更新的状态序列 q_i^* 将用在下一次的迭代中。

3 实验结果和分析

实验使用一个中文女发音人连续语流 EMA 数据库,它同时包含语音波形和 EMA 参数,具体信息可参考第 2 节。本文采用 40 阶线谱对(Line spectral pair, LSP)和 1 阶增益作为频谱声学参数,使用经过咬合面规整的 12 维特征(6 个传感器,每个传感器两维)作为发音动作参数。选择 380 句作训练,剩余的 10 句用作测试。

3.1 上下文属性

为了研究上下文相关 HMM 训练过程中使用的上下文属性集对于发音动作参数预测系统的影响,本文训练了 3 个模型系统:单音素模型、三音素模型及完全上下文相关模型系统。这里,采用独立聚类的频谱模型与发音动作参数模型聚类方式,并且暂不考虑 2.2 节中提出的流间相关性建模。其中,三音素模型的上下文属性包含当前音素及前后各一个音素;完全上下文相关模型的上下文属性除了包含三音素模型中的音素特征,还包含一组广泛的语言韵律特征。表 1 列出了其中一部分上下文属性,表中 L0 表示音节,L1 表示韵律词,L3 表示韵律短语。

表 1 完全上下文相关模型训练中使用的部分上下文属性列表

Table 1 Some context descriptions used in full context dependent model

属性名称	属性描述
L_Tone	前接音节的音调
C_Tone	当前音节的音调
R_Tone	后接音节的音调
C_POS	当前词性
L_BoundType	当前音节的前边界类型
R_BoundType	当前音节的后边界类型
C_RelaPos_L0L1	L0 在 L1 中的相对位置
C_RelaPos_L1L3	L1 在 L3 中的相对位置

分别采用单音素模型、三音素模型和完全上下文相关模型,计算 10 句测试句生成 LSP 参数的均方根误差(Root mean square error, RMSE)作为客观评价标准。3 个系统的实验结果如图 5 所示,单音素模型系统的系能明显低于三音素模型、完全

上下文相关模型系统，因为后两种上下文模型都考虑了当前音素与前后音素的协同发音现象。完全上下文相关模型相对三音素模型增加的上下文属性主要体现的是对基频、时长等韵律参数的影响，因此对于提升发音动作参数的预测精度作用不大。后续的实验都将基于三音素模型进行。

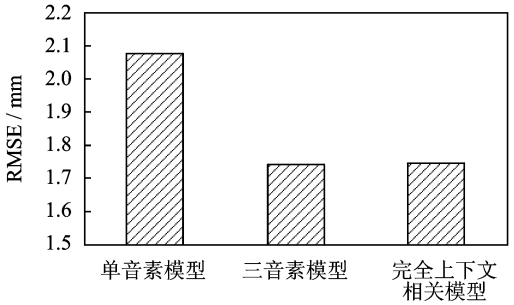


图 5　采用单音素模型、三音素模型与完全上下文相关模型时的发音动作参数预测客观测试结果

Fig. 5　Objective evaluation of articulatory RMSE on monophone model, triphone model and full context model

3.2　聚类方式

在本文的实验数据库上，分别采用共享聚类和独立聚类的决策树叶子节点数目如图 6 所示。采用独立聚类时，EMA 参数的决策树比采用共享聚类的决策树要大，这表明发音动作参数对比声学参数在发音变化上具有更好的区分性。

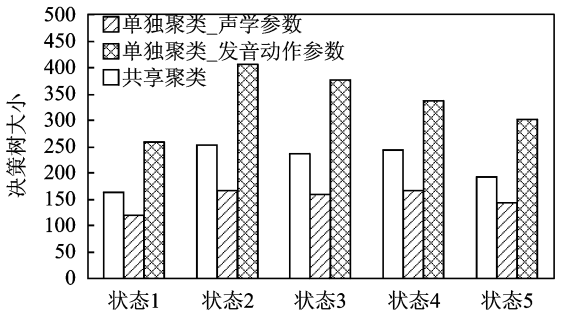


图 6　采用共享聚类与独立聚类方式的各状态决策树叶子节点数目对比

Fig. 6　Node numbers of decision trees on each state for shared clustering and separate clustering

共享聚类与独立聚类的客观测试对比试验结果如图 7 所示。采用独立聚类可以提高 EMA 参数的预测精确性。因此，之后的实验都将采用独立聚类的方式。

3.3　流间相关性建模

进一步验证 2.2 节提出的流间相关性建模方法对于发音动作参数预测性能的影响。为了考虑流间相关性建模中转换矩阵的数目对于系统的影响，采用回归类的方法对转换矩阵和决策树叶子节点进行绑定。因此，本文训练了 5 个系统进行回归类影响的分析，如表 2 所示。

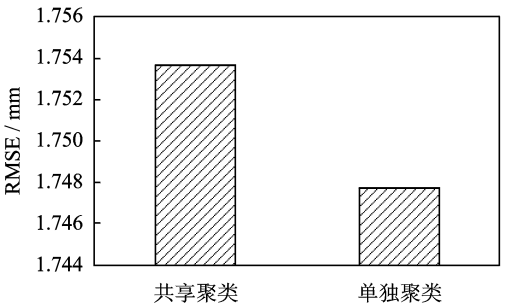


图 7　采用共享聚类与独立聚类时的发音动作参数预测客观测试结果

Fig. 7　Objective evaluation of articulatory RMSE on shared clustering system and separate clustering system

表 2　回归类方法实验的系统配置

Table 2　Configuration for different regression systems

系统名称	系统描述
NoDep	无流间相关性
Reg_100	考虑流间相关性，转换矩阵回归类数为 100
Reg_200	考虑流间相关性，转换矩阵回归类数为 200
Reg_400	考虑流间相关性，转换矩阵回归类数为 400
Reg_leaf_762	考虑流间相关性，对频谱模型聚类决策树中每个叶子节点单独训练转换矩阵（共 762 个）

实验结果如图 8 所示，可以看出加入声学参数与发音动作参数之间的依赖性可以明显提高预测的准确性。并且当增加转换矩阵的数目时，可以提

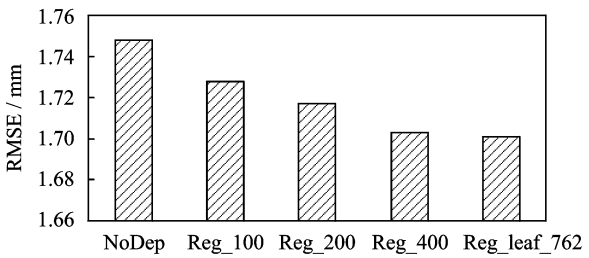


图 8　考虑流间相关性并采用不同绑定方式训练转换矩阵时的系统客观测试结果

Fig. 8　Objective evaluation of articulatory RMSE on different regression systems

高发音动作参数的预测准确性,在绑定到每个叶子节点时得到最优结果。

4 结束语

本文首先阐述了制作中文连续语流发音动作参数数据库及发音动作参数预处理方法。并且在中文数据库上进行了基于 HMM 的发音动作参数预测实验,对比了不同上下文模型、聚类方式对发音动作参数预测性能的影响,结果表明采用三音素模型与单独聚类的模型结构可以得到较好的结果。本文还采用有偏置的线性变换对流间相关性进行建模,并且对转换矩阵的回归类训练方法进行研究。实验表明,随着使用的转换矩阵回归类数目的增多,预测的发音动作参数误差明显下降。未来计划在声学参数与发音动作参数联合模型训练准则、引入非线性变换表征两种参数间依赖关系等方面开展进一步的研究工作。

参考文献:

[1] 赵力.语音信号处理[M].北京:机械工业出版社,2009:14-16.
Zhao Li. Speech signal processing[M]. Beijing: China Machine Press, 2009:14-16.

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

[3] Bare T, Gore J C, Boyce S, et al. Application of MRI to the analysis of speech production[J]. Magnetic Resonance Imaging, 1987,5:1-7.

[4] Akgul Y, Kambhamettu C, Stone M. Extraction and tracking of the tongue surface from ultrasound image sequences[J]. IEEE Comp Vision and Pattern Recog, 1998,123:298-303.

[5] Summerfield Q. Some preliminaries to a comprehensive account of audio visual speech perception[M]. Hillsdale, NJ England: Lawrence Evlbaum Associates, 1987:3-51.

[6] Schönle P W, Gröbe K, Wening P, et al. Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract[J]. Brain Lang, 1987,31:26-35.

[7] 凌震华.基于声学统计建模的语音合成技术研究[D].合肥:中国科学技术大学,2008.
Ling Zhenhua. Research on statistical acoustic model based speech synthesis[D]. Hefei: University of Science and Technology of China, 2008.

[8] Kirchhoff K, Fink G, Sagerer G. Conversation speech recognition using acoustic and articulatory input[C]//ICASSP. Istanbul, Turkey: IEEE, 2000: 1435-1438.

[9] Ling Zhenhua, Richmond K, Yamagishi J, et al. Integrating articulatory features into HMM-based parametric speech synthesis[J]. IEEE Transacions on Audio, Speech, and Language Processing, 2009,17(6):1171-1185.

[10] Blackburn C S, Young S. A self-learning predictive model of articulator movements during speech production[J]. Acoustical Society of America, 2000,107(3):1659-1670.

[11] Birkholz P, Kröger B J, Neuschaefer-Rube C. Model-based reproduction of articulatory trajectories for consonant-vowel sequences[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 10(5):1422-1433.

[12] Ling Zhenhua, Richmond K, Yamagishi J. An analysis of HMM-based prediction of articulatory movements[J]. Speech Communication, 2010, 52: 834-846.

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

[14] Richmond K. Trajectory mixture density networks with multiple mixtures for acoustic-articulatory inversion [C]//NOLISP. Berlin, Heidelberg: Springer-Verlag, 2007:263-272.

[15] Tokuda K, Zen H, Black A W. HMM-based approach to multilingual speech synthesis[M]. United States: Prentice Hall, 2004.

[16] Shinoda K, Watanabe T. MDL-based context-dependent sub-word modeling for speech recognition [J]. Journal of Acoustical Society of Japan (E), 2000,21(2):79-86.

[17] Yoshimura T, Tokuda K, Masuko T, et al. Duration modeling in HMM-based speech synthesis system[C]//ICSLP. Sydney, Australia: [s. n.], 1998,2: 29-32.

[18] Tokuda K, Yoshimura T, Masuko T, et al. Speech parameter generation algorithms for HMM-based speech synthesis[C]//ICASSP. Istanbul, Turkey: [s. n.], 2000,3:1315-1318.

作者简介:蔡明琦(1988-),男,博士研究生,研究方向:语音信号处理;凌震华(1979-),男,博士,副教授,研究方向:语音合成、语音编码、声音转换,E-mail:zhling@ustc.edu.cn;戴礼荣(1962-),男,博士,教授,研究方向:数字信号处理和模式识别。