基于双层码本的语音驱动视觉语音合成系统

贾熹滨 尹宝才 孙艳丰
(北京工业大学多媒体与智能软件技术北京市重点实验室 北京 100124)

摘 要 提出了一种基于双层码本的语音驱动视觉语音合成系统,该系统以矢量量化的思想为基础,建立语音特征空间到视觉语音特征空间的粗耦合映射关系。为加强语音和视觉语音的关联性,系统分别根据语音特征与视觉语音特征的相似性两次对样本数据进行自动聚类,构造同时反映语音之间与视觉语音之间相似性的双层映射码本。数据预处理阶段,提出一种能反映视觉语音几何形状特征与牙齿可见度的联合特征模型,并在语音特征LPCC及MFCC基础上采用遗传算法提取视觉语音相关的语音特征模型。合成的视频中图像数据与原始视频中图像数据的比较结果表明,合成结果能在一定程度上逼近原始数据,取得了很好的效果。

关键词 双层码本,视觉语音合成,视觉语音特征,语音特征
中图法分类号 TP391 **文献标识码** A

Bi-level Codebook Based Speech-driven Visual-speech Synthesis System

JIA Xi-bin YIN Bao-cai SUN Yan-fen
(Beijing Municiple Key Laboratory of Multimedia and Intelligent Software Technolgy,
Beijing University of Technology, Beijing 100124, China)

Abstract The paper proposed a bi-level codebook based speech-driven visual-speech synthesis system. The system uses the vector quantization principle to establish a coarse-coupling mapping relationship from the speech feature space to the visual speech feature space. In order to enhance the relationship between the speech and the visual speech, the system makes the unsupervising-clustering on the sample data according to the similarity of both the acoustic speech and the visual speech and constructs the bi-level mapping codebook reflecting the similarity of both the acoustic speech and the visual speech . At the stage of preprocessing, the paper proposed a joint feature model, which reflects the geometric character and the visibility of teeth. The paper also proposed an approach to extract the visual speech correlative speech feature from the speech features of LPCC and MFCC on the basis of genetic algorithm. The comparison results between the synthesis image sequences with the original one show that the synthesis one can approximate the original one and the result is good. In the future research, the restriction between the visual speech contexts should be considered to improve the smoothness of the synthesis results.

Keywords Bi-level codebook, Visual speech synthesis, Visual speech feature, Speech feature

1 引言

语言是自然人进行交流的主要方式,在人交谈中,除了声音外,人说话时的口形、表情、手势等都是加强语言理解的不可或缺的因素,因而有研究者提出将人类这种自然的交流方式用在人机交互中,提供包括声音语音、视觉语音、表情等多通道交互方式,提高人机交互的自然性和逼真性[1,2],特别与声音语音密切相关的口形,即视觉语音合成的研究成为重要方向之一[3-5]。

视觉语音合成技术概括地说就是如何用语音或文本去预测视觉语音,合成语音同步的语音动画[5,6],由于人说话行为中的声音语音和视觉语音具有完全不同的特征表示方法,因而建模两者的关联关系既是合成技术中的一个重点,也是一个难点。目前在国内外的相关研究中,神经网络、矢量量化、HMM等基于学习的方法被用于建模语音特征空间到视觉语音特征空间的映射关系[6-8],特别是隐马尔可夫模型,其由于具有很好的揭示语音动态特性的优点,因此被广泛采用,但这类方法的声视频映射模型相对复杂,计算复杂度较高。基于矢量量化的声视频映射建模方法是早期提出的方法,其模型简单,但由于难以揭示声视频复杂的关联关系,一直不占主导地位,近年来有研究者从构建视觉语音更相关的语音模型角度提出了一些解决方案,从而为基于矢量量化的方法开辟一种新的研究途径[9]。

本文以基于矢量量化方法的思想为基础,提出一种以视

到稿日期:2013-06-07 返修日期:2013-08-20 本文受国家自然科学基金(61070117),北京市自然科学基金(4122004)资助。
贾熹滨(1969-),女,博士生,副教授,主要研究方向为视觉图像理解,E-mail:jiaxibin@bjut.edu.cn(通信作者);尹宝才(1963-),男,教授,博士生导师,主要研究方向为数字多媒体技术、虚拟现实与图形学技术、多功能感知技术;孙艳丰(1964-),女,教授,博士生导师,主要研究方向为多功能感知。

觉语音帧为步长的语音类到图像类的粗耦合映射模型，通过样本学习建立同时反映语音之间及视觉语音之间相似性的声视频映射双层码本，在此基础上提出基于遗传算法提取视觉语音相关的语音特征的解决方案，用以加强有声语音和视觉语音的关联性。基于所学习的映射模型，在合成端，预测出待合成语音同步的口形图像类序列，进而利用 Viterbi 算法根据相邻帧口形图像距离最小原则实现对新输入语音序列同步的口形图像序列的预测。具体方法将在以下章节中介绍。

本文第 2 节对系统进行了概述；第 3 节阐述了本文所提出的视觉语音特征的建模方法、基于遗传算法(GA)的视觉语音相关的语音特征提取方法以及相关的实验结果；第 4 节介绍了所提出的基于双层码本的语音类/口形类映射模型的建模方法和基于该映射模型的部分合成结果；最后是结论以及下一步的改进点。

2 基于双层码本的语音驱动视觉语音合成系统概述

基于语音预测同步的视觉语音问题的核心问题之一就是建立从语音特征到视觉语音特征空间的映射模型，本文通过分析采集的说话过程视频序，采用基于样本学习的方法建模两者的映射关系。整个视觉语音合成系统框架如图 1 所示，在训练端(见图 1(a))，用无监督聚类算法在样本空间自动聚类语音数据，并对已标定对应关系的样本数据中的口形图像进行分类，生成映射图像类。为了提高图像类的聚类特性，本文对每个映射图像类集合，根据口形图像之间的相似性再次聚类，并反映射到语音特征空间构造第二层码本，建立语音类到口形图像类的映射模型。在合成端(见图 1(b))，以该映射模型为基础，预测出待合成语音同步的口形图像类序列，利用 Viterbi 算法，根据相邻图像距离最小原则，在对应图像类序列的候选图像中搜索出对应的口形序列。为加强语音特征与说话口形的关联关系，本文采用遗传算法对典型的语音特征 LPCC 和 MFCC 进行提取，建立口形相关的语音特征模型。

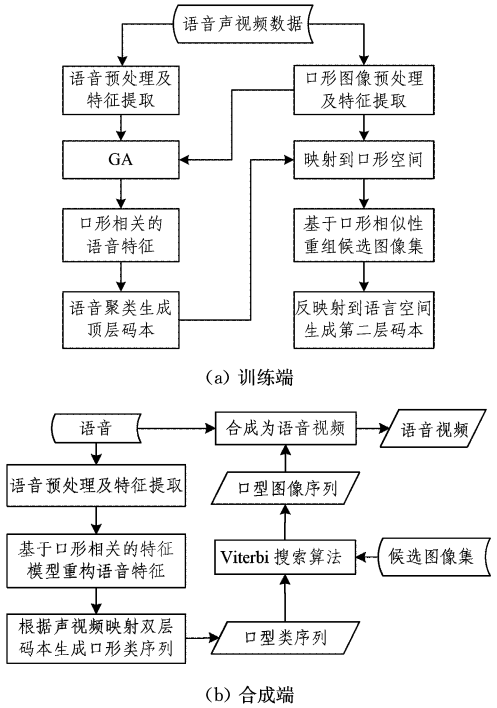

图 1 基于双层码本视觉语音合成系统框图

3 数据采集和预处理

为保证映射码本训练结果的一般性，选取语音音素分布均衡的段落，由特定人用普通话以正常语速读取，以视频采样频率为 25 帧/秒、语音每秒 32k 个采样点的方式采集。选用连续 900 帧图像及对应的语音作为训练样本，分别对语音和图像进行预处理。

3.1 视觉语音特征

考虑到自然人的发音过程不仅决定于口形，其牙齿和舌状态也是影响不同音素发音的影响因素，从外在表象看，体现在内唇纹理具有差异性，提出了一种反映唇形形状的几何特征 R 和反映内唇局部纹理特征 C 组成的联合特征 F 来表示口形图像。式(1)所示为连续视觉语音序列中第 k 帧图像的联合特征 F^k。

$$F^k=[\alpha\times R^k \quad \beta\times C^k] \tag{1}$$

其中，α,β 为折衷两种特征贡献率的加权系数，如式(2)所示。

$$\alpha=\frac{\max\limits_{k\in S}\{C^k\}}{\max\limits_{k\in S}\{C^k\}+\max\limits_{k\in S}\{R^k\}}$$
$$\beta=\frac{\max\limits_{k\in S}\{R^k\}}{\max\limits_{k\in S}\{C^k\}+\max\limits_{k\in S}\{R^k\}} \tag{2}$$

式中，S 为连续采集的视觉语音序列图像样本空间，k 表示样本空间的第 k 帧图像。R^k 为第 k 帧图像的几何特征，C^k 为第 k 帧图像的局部纹理特征，其具体定义方法如下所述。

几何特征向量 R 以 MPEG-4 标准中所定义的唇部 FDP 点为基础，如图 2 中的点，其分量为各 FDP 点之间的弦长函数，即 $r(t)=\sqrt{(x_i-x_j)^2+(y_i-y_j)^2}$ (i,j 分别表示第 i,j 个特征点，(x_i,y_i) 为第 i 个特征点坐标)。考虑到唇部的对称性，如点 2 和点 6，其所张成的特征分量存在着冗余成分，如点 2 与点 9 之间所计算的几何特征分量和点 6 与点 9 之间所计算的几何特征分量反映了相似的口形几何信息，去除类似的重复分量，本文最终确定 27 维的几何特征：$R=[r(1)\ r(2)\ \cdots\ r(m)]$，如图 2 所示。

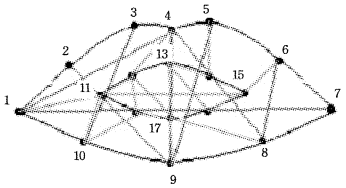

图 2 口形图像几何特征示意图

考虑到内唇纹理主要表现在牙齿可见度和舌状态，其与唇部最大的差异在于颜色，局部纹理特征向量 C 采用反映内唇颜色分布的矩建模，同时考虑到内唇状态的一致性，这里选用沿内唇中心垂直方向纹理一阶颜色矩：$C=[M_r\ M_g\ M_b]$，M_r,M_g,M_b 分别为口形位图图像的 3 个颜色分量 r,g,b 的内唇一阶颜色矩，如式(3)所示。

$$\left.\begin{aligned}M_r&=\frac{1}{N}\sum_{i=1}^{N}r(P_i)\\M_g&=\frac{1}{N}\sum_{i=1}^{N}g(P_i)\\M_b&=\frac{1}{N}\sum_{i=1}^{N}b(P_i)\end{aligned}\right\} \tag{3}$$

式中，N 为内唇特征点 13 和 17 之间沿 x 轴的像素点数，P_i 为第 i 个像素点，$r(P_i)$、$g(P_i)$、$b(P_i)$ 分别为像素点 P_i 在 RGB 空间的红色、绿色、蓝色分量的值。

图 3 为采用本文所提出的联合特征来表示视觉语音图像后，基于 K 均值方法自动聚类后的其中两类。可以看出两类中图像具有相似的口形形状，但具有不同的内唇纹理，这里主要体现在牙齿可见度，由此可见，基于该联合特征能够很好地把不同视觉语音图像有效地区分开，特别能对反映不同发音状态的具有相似口形形状但不同内唇纹理的视觉语音加以区分，满足了本系统对视觉语音描述的需要。

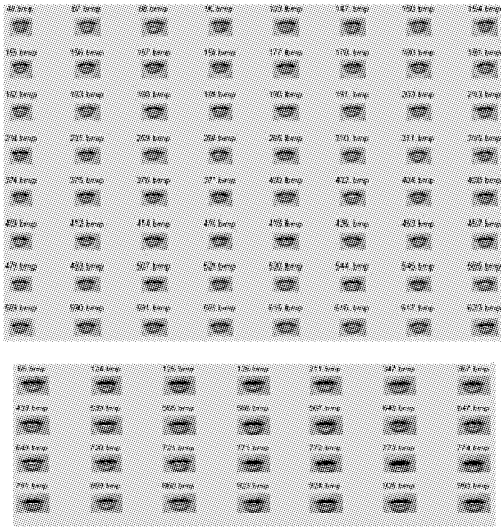

图 3　基于联合特征聚类的部分实验结果

3.2　视觉语音相关的语音特征

3.2.1　视频帧同步的联合语音特征

对于语音数据，本文以 640 个采样点为一帧、相邻帧重叠二分之一的方式进行分帧、加窗，分别提取 16 维 LPCC 和 12 维 MFCC 特征，建立视觉语音帧为步长的语音特征。由于语音采样频率高于视频采样频率，为匹配两者的速率，建立对应关系，本文按照视频帧的速率重组语音帧，即以一帧图像所对应的若干帧语音为一组，这里为 5 帧，组合其语音特征，构造与视频帧同步的联合语音特征，其选取方法如图 4 所示。

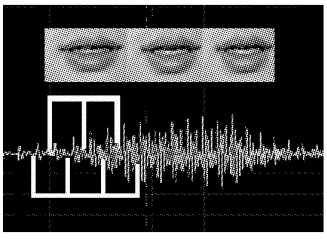

图 4　视觉语音帧同步语音帧选取

3.2.2　基于 GA 视频帧同步的联合语音特征提取

提出的用于音素识别的语音特征如线性预测系数 LPCC、Mel 谱系数 MFCC 被广泛地用于视觉语音合成系统中[6-10]。视觉语音反映的是说话过程中唇部的发音状态，和声音语音之间存在着一致性关系，但作为其中发音器官的一部分，声音语音特征不仅存在与视觉语音关联的成分，而且必然存在不相关因素，这一点可以从口形与语音的对应关系为非一对一的角度加以理解[11]。

为解决语音和可视语音非一一对应问题，提出在传统语音特征基础上，进一步对语音特征预测视觉语音的贡献加以分析，提取视觉语音相关的语音特征。语音识别领域中由于语音特征各维系数对语音的区分性能具有差异，因此实际使用通常会选其中若干维的经验，在提取视觉语音相关的语音特征时，也考虑去除语音特征分量中对预测视觉语音没有贡献的分量。这里提出基于遗传算法对典型的声学语音特征 LPCC 和感知特征 MFCC 各分量对预测视觉语音是否有贡献加以学习的方法，重构出视觉语音相关的语音特征。

本文采用基本遗传算法(SGA)，在样本空间中通过基因的选择、重组和变异，搜索出适应度函数最高的个体，并将其映射为该特定人说话口形相关的语音特征模型。基于 SGA 提取说话口形相关的语音特征模型重点解决两个问题，1)遗传基因的编解码问题：将物理空间的说话口形相关语音特征提取问题转化为遗传算法计算空间可处理的染色体编码；2)适应度函数的设计。

本系统的目标是从典型语音特征：LPCC、MFCC 中选择视觉语音相关的分量，去除冗余分量，本文采用二进制编码方法，假设语音特征向量中的分量与视觉语音相关，基因码对应位为“1”；若无关，则为“0”。基于上述方法，本文建立 28 位的二进制基因串来编码每一帧语音的 16 维 LPCC 和 12 维 MFCC，实现了特征提取的具体问题到遗传算法可计算空间的转换。

适应度函数是根据系统采用语音预测口形图像的目标要求来设计的，即在标定对应关系的视频序列空间中，基于语音特征自动聚类语音空间，并根据标定生成对应的视觉语音图像类，若基于语音特征聚类所生成的视觉语音类具有良好的类聚性，则该语音特征模型对视觉语音有很好的预测能力。因而这里选用评价基于语音模型分类语音后所生成视觉语音映射类的类聚性参数为适应度函数，本文采用类内散布矩阵 S_w 的迹 J_e(见式(4))为基础生成适应度函数 $fitness$(见式(5))。

$$J_e=tr[S_w]=\sum_{i=1}^{c}tr[S_i]=\sum_{i=1}^{c}\sum_{X\in D_i}\|X-m_i\|^2 \tag{4}$$

式中，D_i 表示自动聚类后第 i 个视觉语音映射类的图像集合，c 表示聚类后类的个数，X 为视觉语音的特征向量，m_i 为视觉语音特征向量的均值。

$$fitness==\frac{1}{J_e} \tag{5}$$

视觉语音相关的语音特征为世代进化过程中，使适应度函数最高的最佳编码所对应的模型。这里从样本库中选用其中 500 幅样本图像及对应的语音作为训练样本，进化过程中每代种群个数选为 100，进化代数为 50。根据最佳编码对应的语音模型与采用 LPCC、MFCC 和两者的联合特征相比，利用最佳编码对应的语音对象模型分类所生成的视觉语音类的类内离散度要小于单独采用 LPCC、MFCC 以及两者联合特征，见表 1，该数据表明利用视觉语音相关的语音特征分类语音数据，映射所生成的可视觉语音图像类，类内图像一致性更好，提高了基于语音预测可视语言的能力。

表 1　基于不同声音语音特征聚类衍生的视觉语音类内散度比较

	视觉语音相关语音特征	LPCC	MFCC	LPCC 与 MFCC 联合特征
类内散度	8.5144	9.2041	9.4111	9.2331

为了进一步验证和其他语音特征相比，基于 GA 所提取

的特征与口形具有更强的关联关系，这里以唇高数据为基础，分别对基于上述特征所预测的口形图像序列的唇高与对应原始图像唇高偏差的平均值进行了比较，如表 2 所列，基于提取特征所预测的图像与原始图像唇高的平均偏差要低于基于其他特征的。从这一角度也可以看出基于最佳编码对应的语音特征模型表示语音信号能更好地预测口形。

表 2　基于不同语音特征所预测视觉语音与原始视觉语音的平均偏差

	视觉语音相关语音特征	LPCC	MFCC	LPCC 与 MFCC 联合特征
偏差	5.25	6.36	6.84	6.25

4　基于双层码本的语音类/口形类映射模型

声学语音和视觉语音之间存在着一致性，但两者之间的对应关系又不是一一对应的，即同一口形可能对应着不同的发音，反之亦然[11]。因而为两者映射关系建立确定的数学模型是一个难点，本文绕过该难点，提出以视频帧为单位建立语音类和口形类的映射模型。即：对样本空间的语音信号，根据基于遗传算法所提取的口形相关的特征模型，计算视频帧对应的语音特征后，用 LBG 算法对语音训练样本进行无监督分类，并生成对应候选图像类。如式(6)所示，设 N 个样本聚类集合 $\aleph^N$，口形图像 y_i 根据同一帧语音对象 x_i 分类结果分配到相应的图像类 $\mathscr{R}_i^{N_i}$，并构成图像类集合 $\mathscr{R}^N$，以此作为对合成阶段新输入语音预测口形的候选图像。

$$\aleph^N=\{\aleph_1^{N_1},\aleph_2^{N_2},\cdots,\aleph_c^{N_c}\}$$

$$\text{其中}\aleph_i^{N_i}=\{x_i^k\}$$

$$\forall x_i^k\in\aleph_i^{N_i},\exists y_i^k\in\mathscr{R}_i^{N_i}$$

$$\mathscr{R}^N=\{\mathscr{R}_1^{N_1},\mathscr{R}_2^{N_2},\cdots,\mathscr{R}_c^{N_c}\},i=1,2,\cdots,c;k=1,2,\cdots,N_i \tag{6}$$

由于声学语音和视觉语音之间是一种多对多的对应关系，为了进一步加强训练中所获得的视觉语音图像类中候选图像的相似性，本文提出了构建双层码本。在语音类映射得到的图像类基础上，每一类中候选图像根据口形的相似性重新聚类为子类，子类中图像在样本空间反映射到语音集合，求出对应语音的聚类质心作为第二层码本，如图 5 所示。通过利用视觉语音图像的相似性将语音再次分类并反映射回语音空间，进一步加强了语音和视觉语音的关联性，提高了预测的效果。

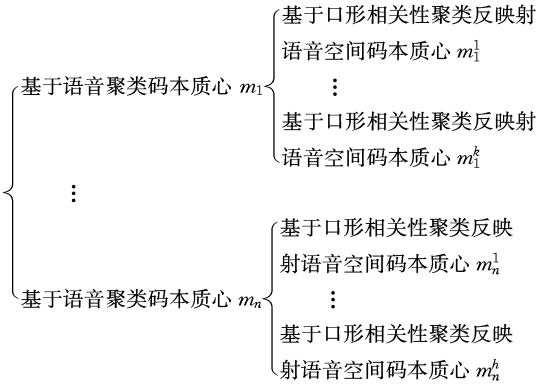

图 5　双层码本聚类质心结构示意图

5　实验结果分析

在合成端，待合成语音根据训练阶段所获的口形相关的语音模型及映射关系，可生成对应的口形图像类序列。为了进一步将图像类序列转化为平滑的口形序列，本文采用了 Viterbi搜索算法在候选图像中，根据最小距离原则确定对应的口形序列[12]。

本文从测试集选取了语音作为输入，利用该系统合成出语音同步的视频，并将所合成视频中的口形序列和原始口形序列进行了比较，如图 6 所示为其中一段视觉语音，可以看出合成图像序列与原始图像序列非常相近，特别在唇部的张合及圆唇状态非常相近，能够满足自然人对视觉语音的正确认知。同时其张合与圆唇程度的差异也反映了在下一步工作中在语音特征中增加一些反映声音强弱等的韵律特征，提高这方面的合成效果，同时部分跳跃现象如第 4、6、7 帧也反映了在下一步工作中应对视觉语音的上下文约束关系加以学习，以提高所合成视频序列的平滑性。

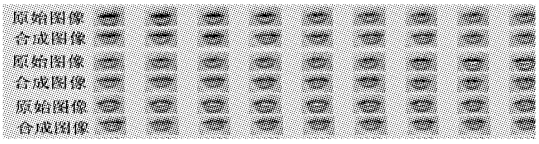

图 6　合成口形序列和原始口形序列的对比(部分数据)

结束语　基于双层码本的语音类到说话口形图像类的声视频映射模型在结构上更简洁，通过建立两者粗耦合的关联关系，绕过准确建模两者关系的困难，在类的概念意义上解决了语音与说话口形之间非一一对应关系的难以描述的问题。为了进一步解决单纯基于语音聚类特性生成的口形图像聚类特性不好的影响，提出了在每个说话口形图像类中再次根据图像相似性自动聚类，生成图像类子集，并反映射到语音空间，在基于语音特征聚类码本中生成下一层码本，构建双层码本，同时基于遗传算法提取视觉语音相关的语音特征等措施，在很大程度上加强了声视频语音之间的关联关系。对于视觉语音特征，本文提出了基于几何和局部纹理特征的联合特征，该特征相对单纯基于几何特征，有效提高了对不同视觉语音的可区分性，在低维空间提供了一种视觉语音图像的有效表示方法。本文采用样本重组方法，分别根据距离最小准则，在候选图像样本空间中搜索出平滑的口形序列。实验数据表明，基于双层码本预测的口形图像序列与原始说话口形图像序列非常接近。

对口形序列上下文的约束关系还有待在下一步工作中进一步完善，以提高合成视频的平滑性。

参 考 文 献

[1] Jia Jia, Zhang Shen, Meng Fan-bo, et al. Emotional audio-visual speech synthesis based on PAD[J]. IEEE Transactions on Audio, Speech and Language Processing, 2011, 19(3): 570-582

[2] 谢金晶，陈益强，刘军发. 基于语音情感识别的多表情人脸动画方法[J]. 计算机辅助设计与图形学学报，2005，20(4)：520-525

[3] Pandzic I S, Ostermann J, et al. User evaluation : synthetic talking faces for interactive services[J]. The Visual Computer, 1999, 15(7/8): 330-340

[4] Massaro D W, Ouni S, Cohen M M, et al. A multilingual embodied conversational agent[A]//Proceedings of 38th Annual Hawaii International Conference on System Sciences (HICCS'05) (CD-ROM, 10 pages) [C]. Los Alimitos, CA, IEEE Computer Society Press, 2005

[5] 王志明，陶建华. 文本-视觉语音合成综述[J]. 计算机研究与发

展,2006,43(1):145-152
[6] Gao W,Chen Y Q,et al. Learning and synthesizing mpeg-4 compatible 3-d face animation from video sequence[J]. IEEE Transactions on Circuits and Systems for Video Technology,2003,13(11):1119-1128
[7] Brand M. Voice puppetry [C]// Proceedings of ACM SIGGRAPH 1999. ACM Press/Addison-Wesley Publishing Co:New York,NY,USA,1999:21-28
[8] Morishima S, Harashima H. Speech-to-image media conversion based on VQ and neural network,ICASSP 91[C]//1991 International Conference on Acoustics,Speech and Signal Processing. 1991:2865-2868
[9] Gutierrez-Osuna R, Kakumanu P, Esposito A, et al. Speech-driven facial animation with realistic dynamics[J]. IEEE Transactions On Multimedia,2005,7(1):33-41
[10] Jiang J T,Alwan A,Bernstein L E,et al. Predicting face movements from speech acoustics using spectral dynamics[C]//IEEE International Conference on Multimedia and Expo. 2002:181-184
[11] Bregler C, Covell M, Slaney M. Video rewrite: driving visual speech with audio,SIGGRAPH'97[C]//ACM Press/Addison-Wesley Publishing Co:New York. NY,USA,1997:353-360
[12] Graf H P,Cosatto E. Sample-based synthesis of talking-heads [C]//The 8th IEEE Int'l Conf. Computer Vision. 2001:3-7

(上接第 82 页)

对基于 HMM 的蒙古语语音合成系统进行评价时,我们用 MTTS 对随机选取的 50 句蒙古语句子进行了合成实验,然后通过主观评价的方法由 2 位懂蒙古语的老师和 3 位懂蒙古语的同学对合成的蒙古语语音进行评价。结果表明,合成的语音整体稳定流畅,可懂度高,而且节奏感比较强。最后我们又采用主观平均分数 MOS(Mean Opinion Score)对合成的 50 句蒙古语语音进行打分。MOS 是目前使用得最广泛的一种主观评定方法,评分范围是 1 到 5 分,测试时要求听者按照表 5 所列的评分标准给出语音的得分[18]。

表 5 主观评分标准

MOS	质量	失真情况
5	优	十分自然,不觉察失真
4	良	比较自然,刚觉察失真
3	中	觉察失真,但可以接受
2	差	比较不自然,但不令人反感
1	劣	不能接受,令人反感

表 6 列出了 5 位评价者对 50 个合成蒙古语语音的 MOS 打分,从表 6 中很容易得到 5 位评价者对 50 个合成蒙古语语音的平均 MOS 打分为 3.80,接近于 4;而且 5 位评价者评分的方差仅仅为 0.008,可见 5 位评价者对 50 个合成蒙古语句子的评价是一致的。这说明基于 HMM 的方法进行蒙古语的语音合成是非常有效的。

表 6 主观评定结果

评价者	1	2	3	4	5
MOS	3.9	3.7	3.9	3.8	3.7

结束语 本文首次将基于 HMM 的语音合成方法应用在蒙古语上,初步实现了基于 HMM 的蒙古语的语音合成系统,并且进行了实验。实验结果表明,合成的蒙古语语音整体稳定流畅,可懂度高,节奏感比较强,能达到 3.80 的 MOS 得分。这为我们进一步深入研究基于 HMM 的蒙古语语音合成奠定了基础。然而,我们仅做了一些初始的工作,还有很多方面可以优化。在下一步的工作中,我们将重点针对基于 HMM 的蒙古语语音合成系统的前端韵律预测、语料库的完善等方面进行处理。

参 考 文 献

[1] 敖其尔,巩政. 一种波形拼接的语音合成实验[C]//第三届全国人机语音通讯学术会议. 重庆,1994:408-412
[2] 萨其容贵. 蒙古语语音合成技术的研究[D]. 呼和浩特:内蒙古大学,2005
[3] 田会利. 基于词干词缀的有限条词的蒙古语语音合成系统的研究[D]. 呼和浩特:内蒙古大学,2007
[4] 孟和吉雅. 基于动词词干词缀的蒙古语语音合成方法[J]. 内蒙古大学学报:自然科学版,2008,39(6):693-697
[5] 敖敏. 基于韵律的蒙古语语音合成研究[D]. 呼和浩特:内蒙古大学,2012
[6] Zen Hei-ga, Takashi N, Junichi Y, et al. The HMM-based Speech Synthesis System (HTS) Version 2.0[C]//6th ISCA Workshop on Speech Synthesis. Bonn,2007:294-299
[7] 井晓阳,罗飞,王亚棋. 汉语语音合成技术综述[J]. 计算机科学,2012,39(11A),386-390
[8] 确精扎布,陈壮,何正安,等. GB 25914—2010 传统蒙古文名义字符、变形显现字符和控制字符使用规则[S]. 北京,中国标准出版社,2010
[9] 清格尔泰. 蒙古语语法[M]. 呼和浩特:内蒙古人民出版社,1991:65-66,76-77
[10] Tokuda K,Masuko T,Miyazaki N,et al. Hidden Markov models based on multi-space probability distribution for pitch pattern modeling[C]//IEEE International Conference on Proceedings of the Acoustics, Speech, and Signal Processing. Arizona, 1999:229-232
[11] masuko T,Tokuda K,Kobayashi T,et al. Speech synthesis from HMMs using dynamic features[C]//IEEE International Conference on Proceedings of the Acoustics, Speech, and Signal Processing. Atlanta,1996:389-392
[12] Kawahara H, Masuda-Katsuse I, deCheveigne A. Restructuring speech representations using pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds[J]. Speech Communication,1999,27(3/4):187-207
[13] 吴义坚,王仁华. 基于 HMM 的可训练中文语音合成[J]. 中文信息学报,2006,20(4):75-81
[14] Paul B,David W. Praat:doing phonetics by computer[OL]. http://www.fon.hum.uva.nl/praat/,2005
[15] CUED. Hidden Markov Model Toolkit (HTK)[OL]. http://htk.eng.cam.ac.uk/,2009
[16] Satoshi I,Takao K. Speech Signal Processing Toolkit[OL]. http://sp-tk.sourceforge.net/,2012
[17] HTS working group. HMM-based Speech Synthesis System (HTS)[OL]. 2012. http://hts.sp.nitech.ac.jp/
[18] Wikipedia. Mean opinion score [OL]. http://en.wikipedia.org/wiki/Mean_opinion_score,2013