# 基于HMM的蒙古语语音合成技术研究

赵建东 高光来 飞 龙
(内蒙古大学计算机学院 呼和浩特 010021)

**摘 要** 基于隐马尔科夫模型的语音合成方法是当今语音合成的主流方法,它已被广泛应用于英语、汉语、日语等语音合成系统中。然而基于隐马尔科夫模型的蒙古语的语音合成技术研究还处于空白状态。首次将基于隐马尔科夫模型的语音合成方法用于蒙古语语音合成,并进行了语音合成实验。从最终合成系统的效果来看,合成的语音整体稳定流畅,可懂度高,而且节奏感比较强,主观平均得分为3.80。这为进一步研究基于隐马尔科夫模型的蒙古语语音合成技术奠定了基础。

**关键词** 隐马尔科夫模型,蒙古语,标注,语音合成
**中图法分类号** TP391 **文献标识码** A

## Research on HMM-based Mongolian Speech Synthesis

ZHAO Jian-dong GAO Guang-lai BAO Fei-long
(College of Computer Science, Inner Mongolia University, Hohhot 010021, China)

**Abstract** HMM-based speech synthesis method, as a mainstream method nowadays, has been widely applied to English, Chinese, Japanese, and so on. However, the research on HMM-based Mongolian speech synthesis is still in blank field. We applied the HMM-based speech synthesis method to Mongolian firstly, and did some experiments. From the evaluation results of the final Mongolian speech synthesis system, the synthesized Mongolian speech is stable, fluent, rhythmed and has high intelligibility. The mean opinion score of the synthesized Mongolian speech is 3.80. This laids the foundation for further research on the HMM-based speech synthesis technology.

**Keywords** HMM, Mongolian, Annotation, Speech synthesis

## 1 引言

蒙古语是内蒙古自治区的主体民族语言。在中国,除内蒙古自治区外,还有新疆、青海、甘肃、辽宁、吉林、黑龙江等省区使用蒙古语。在世界范围内,除中国外,蒙古国、俄罗斯的布力亚特共和国等国家和地区也使用蒙古语。研究蒙古语的语音合成技术,对少数民族地区教育、交通、通讯、自动化办公具有重要的意义。在蒙古语的语音合成方面,已有前人做了一些研究。敖其尔、巩政提出了一种波形拼接的蒙古语语音合成方法[1];萨其容贵用基音同步叠加法建立了多样板语音合成音库,对蒙古语的语音合成进行了研究[2];田会利对基于词干词缀的有限条词的蒙古语语音合成系统进行了研究[3];孟和吉雅对基于动词词干词缀的蒙古语语音合成方法进行了研究[4];敖敏对基于韵律的蒙古语语音合成进行了研究[5]。这些工作具有一定的价值,但它们都是基于大语料库的拼接合成技术,这种技术存在合成语音的效果不够稳定,语音库的构建周期太长以及合成系统的可扩展性差等缺陷,这些缺陷限制了它在多样化语音合成方面的应用。而基于隐马尔科夫模型HMM(Hidden Markov Model)的语音合成方法可以在短时间内,基本不需要人工干预的情况下自动构建出一个新系统,而且整个训练过程基本上不依赖于发音人、发音风格以及情感等因素[6]。现在,基于HMM的英语、汉语[7]、日语等语音合成系统已被广泛应用,但是基于HMM的蒙古语语音合成技术研究还处于空白状态。

本文首次将基于HMM的语音合成方法应用到蒙古语的语音合成,我们的主要工作是:构建蒙古语语料库,结合蒙古语特性设计了上下文属性集以及相应的用于模型聚类的属性问题集,最后实现了基于HMM的蒙古语语音合成系统。通过实验发现合成的语音质量整体稳定流畅,而且节奏感比较强,达到了预期目标,为进一步深入研究基于HMM的蒙古语的语音合成技术奠定了基础。本文第2节介绍了蒙古语语料库的构建;第3节对基于HMM的蒙古语语音合成方法进行整体介绍;第4节给出了基于HMM的蒙古语语音合成系统上下文属性集和对应问题集的设计;第5节为实验及其评价结果;最后为结束语。

## 2 蒙古语语料库的构建

### 2.1 蒙古语的基本特点

蒙古文字是一种拼音文字,它有35个字母,其中包括8

到稿日期:2013-05-11 返修日期:2013-08-02 本文受国家自然科学基金项目(61263037),内蒙古自然科学基金重大项目(2011ZD11)资助。
**赵建东**(1988—),男,硕士生,主要研究方向为语音合成、语音识别,E-mail:djzshadow@126.com;**高光来**(1964—),男,硕士,教授,主要研究方向为文字识别、语音识别、图像识别、计算智能和数据挖掘;**飞 龙**(1985—),男,博士,主要研究方向为蒙古文信息处理、语音识别、语音合成、语音检索。

个元音、17 个基本辅音和 10 个借词辅音[8]。蒙古文的拼写规则是以词为单位竖写，词与词之间用空格分开，采取从上到下的书序，从左到右的行序。蒙古文字母在词中存在变形，即在一个蒙古文单词中，蒙古文字母在上、中、下位置不同将导致写法也不同。同时蒙古文字母中形同音不同的现象比较普遍。鉴于蒙古文中元音与辅音的形式变化多样问题，本文对蒙古文进行处理时采用了拉丁转写的方法。这样有助于蒙古文的校正、统计和研究。本文使用的蒙古文与拉丁字母的对照如表 1 所列。

表 1　蒙古文拉丁对应表

| 元音 | 拉丁转写 | 辅音 | 拉丁转写 | 辅音 | 拉丁转写 | 辅音 | 拉丁转写 |
|---|---|---|---|---|---|---|---|
| ᠠ | a | ᠩ | N | ᠬ | x | ᠺ | k |
| ᠡ | e | ᠨ | n | ᠲ | t | ᠻ | K |
| ᠢ | i | ᠪ | b | ᠳ | d | ᠴ | C |
| ᠤ | q | ᠫ | p | ᠴ | c | ᠽ | z |
| ᠥ | v | ᠬ | h | ᠵ | j | ᠵ | H |
| ᠦ | o | ᠭ | g | ᠶ | y | ᠷ | R |
| ᠧ | u | ᠮ | m | ᠷ | r | ᠯ | L |
| ᠨᠧ | E | ᠯ | l | ᠸ | w | ᠽ | Z |
| | | ᠰ | s | ᠹ | f | ᠺ | Q |

蒙古语标准读音是以内蒙古方言的正蓝旗蒙古语为代表的察哈尔土语作为标准。蒙古文以音节为发音单位，蒙古文的音节都是以元音为中心构成的。蒙古文常见的口语音节有如下几种：(1)元音；(2)辅音＋元音；(3)元音＋辅音；(4)辅音＋元音＋辅音；(5)元音＋辅音＋辅音；(6)辅音＋元音＋辅音＋辅音[9]。

### 2.2　蒙古语语料库标注

对蒙古语语料库进行标注前首先将蒙古文通过规则转到拉丁转写，然后将拉丁转写通过词典自动翻译成所对应的发音音素。转写例子如下：

蒙古语：

拉丁转写：yeren　tabvn　jqgqs

发音音素：yiresn　tabas1n　jwls

通常蒙古语口语中常用的音素有 62 个，由于某些音素的发音极其相似，自然人分辨起来都存在一定困难，因此，我们按蒙古语语音合成的需要将发音相近的音素做了一些合并。合并后的标注音素总共为 59 个，包括 34 个元音、23 个辅音、1 个句末停顿(sil)和 1 个句中停顿(sp)。标注音素表如表 2 所列。

表 2　标注音素表

| 元音 | | | | 辅音 | | | 停顿 |
|---|---|---|---|---|---|---|---|
| a | as2 | vl | Oll | b | r | k | sp |
| e | es | ul | vi | p | j | z | sil |
| Il | ws | ae | ui | W1 | q | c | |
| i | os | oe | va | n | x | Zl | |
| w | al | Yl | vae | s | y | Cl | |
| v | el | ael | ue | d | g | | |
| o | Ill | oel | Yll | t | h | | |
| u | il | Ell | | m | Nl | | |
| as1 | wl | ol | | l | f | | |

标注的 4 层为：音素层、语调层、音节层、韵律层。其中音素层、音节层和韵律层主要是为了得到音素、音节、短语和句子的相互关系及其对应的音频边界，语调层是为了获得重读和边界调信息。图 1 给出了一个标注示例，其中 H* 表示对应的音节重读，L-L%表示句尾是降音，L-H%表示句尾是升音，1 表示对应的是蒙古语的一个韵律词，3 表示对应的是一个韵律短语的结束词，4 表示一个完整的句子结束的词。一般在标记长句时 1、3、4 会同时存在，但是在短句子中，通常只存在 1、4，或者只存在 4。

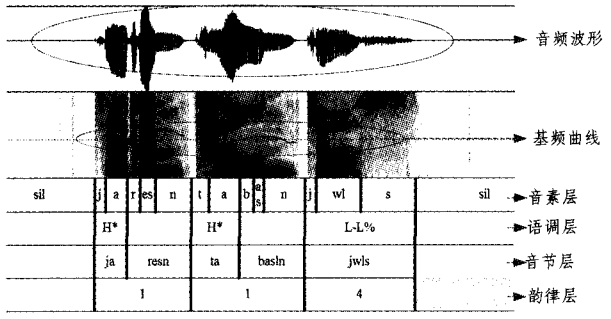


图 1　蒙古语语料库的标注示例

## 3　基于 HMM 的蒙古语语音合成概述

图 2 是基于 HMM 的蒙古语语音合成系统的基本流程图，它可以分成两个阶段：训练阶段和合成阶段。

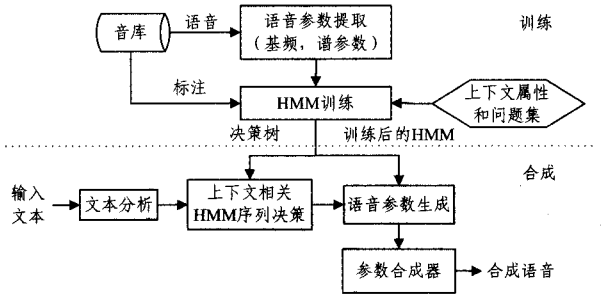


图 2　基于 HMM 的语音合成系统的流程

训练阶段主要包括预处理和隐马尔科夫模型训练。在预处理阶段，首先需要对音库中的语音数据进行分析，以提取出一些相应的语音参数(基频和谱参数)。根据提取得到的语音参数，隐马尔科夫模型的观测向量可以分成谱和基频两个部分。其中，采用连续概率分布 HMM 对谱参数部分进行建模，而对基频部分则采用多空间概率分布 HMM[10] 进行建模。对隐马尔科夫模型进行训练前，另外一个重要的工作就是设计上下文属性集和用于决策树聚类的问题集，即根据先验知识来选择一些对谱、基频和时长这些声学参数有一定影响的上下文属性，并设计与上下文属性相应的问题集以用于上下文相关模型聚类。

预处理完成后就是整个隐马尔科夫模型的训练过程，其训练步骤依次为模型的初始化、声韵母的 HMM 训练、扩展上下文相关模型的训练、聚类后模型的训练以及时长模型的训练，最后得到的训练结果包括谱、基频和时长参数的聚类隐马尔科夫模型以及各自的决策树。

合成阶段主要分为 3 个步骤，首先，输入的文本经过文本分析后转换成上下文相关的单元序列。然后，利用训练得到

的决策树对每一个单元进行决策，得到对应的聚类状态模型，并形成聚类状态模型序列。最后，根据参数生成算法[11]，利用参数的动态特性来生成目标的声学参数序列，并且通过STRAIGHT[12]合成器得到最终的合成语音[13]。

## 4 蒙古语上下文属性集和问题集设计

在HMM的训练阶段，得到训练好的单音素模型后，首先根据上下文属性集合进行模型扩展，然后对扩展后的模型进行训练。由于所有的上下文属性组合数远远大于训练数据的数目，对于每一个上下文相关模型，通常其对应的训练数据非常有限(一到两个)。因此，我们要通过比较分析来删除一些与蒙古语相关性低的上下文属性。只选择那些与蒙古语相关性高，并且对训练模型有利的上下文属性。根据最终选择的蒙古语上下文属性集生成的用于上下文相关模型训练的文件格式如下：

p1－p2＋p3@p4～p5

/A:a1～a2/B:b1～b2_b3@b4＃b5＄b6

/C:c1～c2_c3@c4＃c5/D:d1～d2/E:e1～XX_e2

/F:f1/G:g1～XX_g2/H:h1～h2_h3@h4

/I:i1～i2_i3@i4＃i5/J:j1～j2/K:k1～k2～k3/L:l1

以上用于模型训练的文件格式中各标记的含义如表3所列。

表3 上下文属性标注格式解析

| 标记 | 对应上下文属性 | 标记 | 对应上下文属性 |
|---|---|---|---|
| p1 | 前音素 | p2 | 当前音素 |
| p3 | 后音素 | k3 | 句子包含短语数 |
| p5 | 当前音素在音节中的后向绝对位置 | a1 | 前音节是否重读(0:不重读,1:重读) |
| a2 | 前音节包含音素数 | b2 | 当前音节包含音素数 |
| b1 | 当前音节是否重读(0:不重读,1:重读) | b3 | 当前音节在单词中的前向绝对位置 |
| b4 | 当前音节在单词中的后向绝对位置 | b5 | 当前音节在短语中的前向绝对位置 |
| b6 | 当前音节在短语中的后向绝对位置 | c1 | 当前短语中前向重读音节个数 |
| c2 | 当前短语中后向重读音节个数 | c3 | 当前音节在短语中到前一个重读音节的距离 |
| c4 | 当前音节在短语中到后一个重读音节的距离 | d1 | 后音节是否重读(0:不重读,1:重读) |
| c5 | 当前音节中元音音素名 | d2 | 后音节包含音素数 |
| e1 | 前单词包含音节数 | e2 | 前单词是否短停 |
| f1 | 当前单词包含音节数 | g1 | 后单词包含音节数 |
| g2 | 后单词是否短停 | h1 | 前短语包含音节数 |
| h2 | 当前短语包含单词数 | i1 | 当前短语包含音节数 |
| h3 | 当前单词在短语中的前向绝对位置 | h4 | 当前单词在短语中的后向绝对位置 |
| i2 | 当前短语包含单词数 | i5 | 当前短语的边界调 |
| i3 | 当前短语在句子中的前向绝对位置 | i4 | 当前短语在句子中的后向绝对位置 |
| j1 | 后短语包含音节数 | j2 | 后短语包含单词数 |
| k1 | 句子包含音节数 | k2 | 句子包含单词数 |
| p4 | 当前音素在音节中的前向绝对位置 | i1 | 当前音节后的边界类型(1:音节,2:单词,3:短语,4:句子) |

由于训练数据稀疏，模型的参数在训练后基本上都“过拟合”到那一两个数据上。对此，我们还要采用基于决策树的聚类方法对上下文相关模型进行聚类，以提高模型的鲁棒性以及模型复杂度和训练数据量之间的均衡性。表4列出了本文设计的用于蒙古语语音合成系统的部分问题集。

表4 蒙古语语音合成系统部分问题集

| 问题 | 含义 |
|---|---|
| L_Unit | 前接单元 |
| C_Unit | 当前单元 |
| R_Unit | 后接单元 |
| QS L_YY | 前接元音 |
| QS L_FY | 前接辅音 |
| QS L_Tone | 前音节重读 |
| QS L_AN_L0conLp | 前音节包含音素数 |
| QS C_YYinL0 | 当前音节中元音音素 |
| QS C_AHP_L0inL1 | 当前音节在单词中前向绝对位置 |
| QS C_AHP_L1inL3 | 当前单词在短语中前向绝对位置 |
| QS C_RWL | 当前音节后的停顿类型 |
| QS C_W_PAUSE | 当前单词是否短停 |
| QS R_AN_L3conL1 | 后短语包含单词数 |

基于决策树的模型聚类过程如图3所示，整个流程主要分3步：首先，把所有模型都放在根节点，作为当前待分裂节点；其次，遍历所有的问题，进行尝试分裂并计算得分，取得分最大的问题作为最终分裂问题；最后，对分裂后的节点重复以上操作，直到所有叶子节点不能分裂为止。

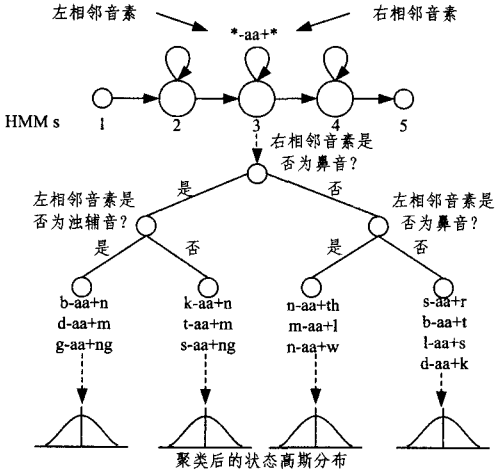


图3 基于决策树的模型聚类

## 5 基于HMM的蒙古语语音合成系统实现及评价

要实现基于HMM的蒙古语语音合成系统，首先要建立用于训练和测试的蒙古语语料库。实验中用于训练的有1098句音素覆盖均衡的蒙古语句子，时长约1.5小时。音频为一名20岁蒙古族女生用蒙古语标准读音在具有安静环境的录音室录制，采样率为44.1kHz，在训练时将采样率转换为16kHz。标注时采用本文第2节介绍的标注方法，标注工具为praat[14]。训练前将标注文本转换为训练需要的数据格式，然后用HTK[15]和SPTK[16]将数据进行了训练。最终在工具包HTS[17]的基础上实现了基于HMM的蒙古语的语音合成系统MTTS(Mongolian Text to Speech)，其界面如图4所示。

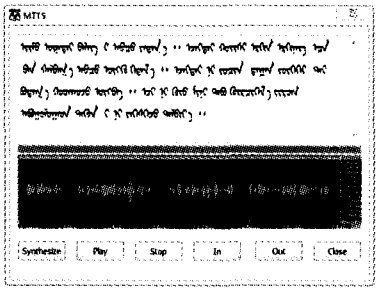


图4 MTTS界面

(下转第104页)

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