基于 HMM 的傣语语音合成系统设计与实现

方　媛　杨　鉴　陈志琼　王　昱

摘　要　傣语是傣族人民使用的语言，目前仅在中国云南就有近 120 万人口使用。研究傣语语音合成对推进傣语信息化具有重要意义。本文以开发应用系统为目标，研究了基于 HMM 框架的傣语语音合成系统的实现方法，并详细阐述了语料收集与挑选、录音、文本归一化、自动分词、文本罗马化与标注、上下文属性与问题集设计、HMM 训练以及产生合成语音等模块的实现方法。实验结果表明，采用本文的合成方法，其合成语音有较好的可懂度，而其自然度还有待进一步提高。

关键词　傣语，语音合成，文本归一化，分词，文本罗马化，HMM 训练

Design and Implementation of an HMM-based Dai Speech Synthesis System

FANG Yuan　YANG Jian　CHEN Zhiqiong　WANG Yu

Abstract Dai is the language used by the Dai people. There are nearly 1.2 million people in Yunnan China who use Dai language. The research of Dai speech synthesis is of great significance to promote the informatization of Dai. This paper takes the development of application system as the target and studies implementation of Dai speech synthesis system based on HMM framework. It elaborates the speech corpora's collection and selection, recording, text Normalization, Dai word segmentation, text Romanization and labeling, the design of the context-dependent label files, the training of HMM and producing synthetic speech. The results show that, the intelligibility of the synthesized speech is good, while its naturalness needs to be improved.

Key words Dai Language, Speech Synthesis, Text Normalization, Word Segmentation, Text Ro-manization, HMM Training

1. 引言

傣语是傣族人民所使用的语言，属汉藏语系壮侗语族壮傣语支。傣语是源于古印度字母系统的一种拼音文字，伴随着佛教的传入而产生。傣语分为西双版纳傣语、德宏傣语、金平傣语、红金傣语等四种方言。本文选用西双版纳傣语作为研究语言。西双版纳傣语是一种带声调的语言，并采用拼音文字。其音节主要由声母、韵母、声调组成。傣语声母分为高低两组共 42 个，韵母 91 个，在区分舒声调和促声调的前提下共有 9 个声调。其音节构成特点为 C + V + C 或者 C + V，且只有 -p，-t，-k，-ŋ，-m，-n 能做尾辅音［2］。

语音合成近年在技术和应用方面都取得了长足进展。随着信息技术的发展以及对语音信号和语音合成技术本身的不断深入理解，合成语音的各项指标包括自然度、可懂度和音质都得到明显的改善，并在越来越多的实际系统中得到应用。目前，语音合成技术已经在自动应答呼叫中心、公共交通平台、汽车导航以及智能电子设备等方面得到广泛应用。而基于隐马尔可夫模型（Hidden Markov Model，HMM）的统计参数模型合成方法在标注语料数据相对缺乏的情况下，能保证更稳定的合成音质和效果［3］。

语音合成的发展为人们的生活提供了很多便利。而目前的语音合成语言主要针

对汉语、英语等常用语言，对少数民族语言进行语音合成的研究相对缺乏。傣语虽然作为一种民族语言，但是其使用人数众多。在全球大约有 6600 万人口使用傣泰语，在中国云南有 120 多万人使用傣语。研究傣语语音合成系统对推进傣语信息化具有重要意义。本文基于隐马尔可夫模型（Hidden Markov Model，HMM）的统计参数模型合成方法，设计并实现了傣语语音合成系统。

2. 基于 HMM 的语音合成

隐马尔可夫模型（HMM）在语音信号处理中的应用已经有 20 多年了。近年来，HMM 被广泛应用到语音合成领域。

基于 HMM 的统计参数语音合成方法提供了一整套不需人工干预的基于数据驱动的语音合成方法，图 1 为其基本框架，主要包括训练和合成两个部分。

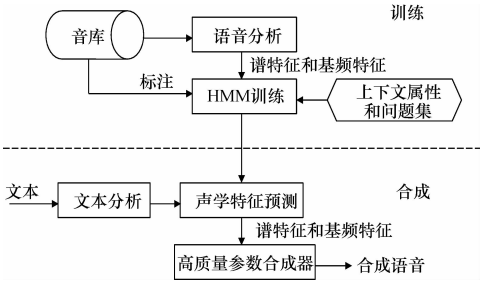

图 1　基于 HMM 语音合成基本框架

在训练过程中，我们通过分析语音波形进行声学特征提取，主要为谱特征和基频特征参数。由于 STRAIGHT 合成器可以提供高精度的声道估计和声源估计［1］，很多基于统计参数合成系统采用 STRAIGHT 合成器。同时相关的文本被转换成对应的标注数据；标注数据主要包括音段切分和韵律标注。基于提取的声学特征和转换后的标注数据，我们利用统计学习算法进行 HMM 训练。

在合成时，对输入文本通过文本分析模块进行上下文属性分析，并利用训练得到的模型进行声学特征预测，最后通过高质量参数合成器合成出语音。

3. 傣语语音合成系统的设计与实现

在设计与实现傣语语音合成过程中，我们是基于 HMM 语音合成基本框架来进行的。整个合成过程主要包括：语料收集与挑选，录音，归一化，分词，文本罗马化与标注，上下文属性和问题集设计以及 HMM 训练与合成。其框架如图 2 所示。

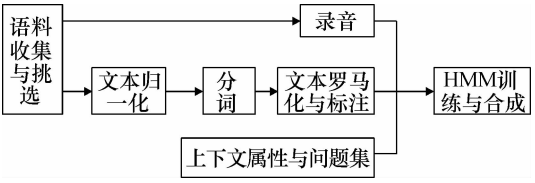

图 2　傣语语音合成框架

3.1 语料收集与挑选

根据语音合成框架，我们在训练声学模型时，需要有 Wav 文件以及与之对应的发音文稿，因而需要先构建一个傣语文本语料库，然后按某一准则挑选我们所需要的发音文稿语料。

本文使用 Teleport Ultra 工具，从西双版纳新闻网站下载并整理语料，构建了一个 60MB 的傣语文本语料库。对于收集的语料，我们编写程序，以音节覆盖率最大化为挑选准则，进行发音语料的挑选。

在设计挑选算法时，我们是以挑选傣语语料中完整句子为目的，在此基础上，我们主要考虑三个原则：第一，挑选的句子尽可能多地覆盖可能出现的所有音节；第二，针对某一语句，其出现音节在入选的发音语料中已经有足够的覆盖次数后，该语句不再入选发音语料；第三，句子越长，其入选发音语料的可能性越大，但是也和发音语料中音节覆盖率相关。考虑上述原则，我们设计出如下评估函数：

$$G = \sum_i g_i \tag{1}$$

其中，G 为待选句子的总得分，g_i 为

句子中第 i 个音节的得分，g_i 的计算公式如下：

$$g_i = \begin{cases} 200, & w_i = 0 \\ 20 - w_i, & w_i < 20 \\ 0, & w_i \geqslant 20 \end{cases} \quad (2)$$

其中，w_i 为第 i 个音节在发音语料中的出现次数。在式（2）中，我们假设如果待入选句子的第 i 个音节未在发音语料中出现则该音节得分为 200，因为我们在发音语料中需覆盖尽可能多的新音节，因而新音节的初始得分应该较高，同时我们假设如果待入选句子的第 i 个音节已经在发音语料中出现 20 次及以上，则该音节得分为 0。

通过采用上述算法，最终我们构建了一个规模为 2500 句的发音语料，其全音节覆盖率为 94.2%，能覆盖绝大多数的合法音节。在置信度为 98% 时，傣语语料库与发音文本的相似度为 0.4。

3.2 录音

发音文本挑选完后，我们需要请专业的播音员对语料进行录音。本项目选请的发音人为西双版纳广播电视台的专职傣语播音员，具有播音副高级职称。发音方式按文稿朗读，新闻播音风格，语句之间有明显的停顿。录音设备与环境为广播电视台录音室，采用数字录音。最终获得语料集 2504 句，约 28.2 小时的语音数据（16kHz 采样，16bit 量化）。

3.3 文本归一化

文本归一化，又称为文本正则化，是指将文本中非标准词（如数字、特殊符号等）进行消歧和标准化转换。它主要包括非标准词识别、歧义判断、消歧处理、非标准词转换为标准词四个处理步骤。

傣语中的非标准词主要是纯数字。对于纯数字而言，大部分情况下读为数值，少部分情况下需读为数码。这里重点讲述傣语纯数字的消歧与标准化处理过程。傣语数字有三种形式：阿拉伯数字、傣语数字、傣语数字读音。其三者之间的对应关系如表 1 所示。

表 1　傣语部分数字的对应关系表

阿拉伯数字	傣语数字	傣语读音
0	๐	သုင်
1	๑	ၵဵဵ
2	๒	သွင်
3	๓	သၢမ်
…	…	…

对于傣语数字，其数码与数值的归一化过程是不相同的。若是按数码来读，则将阿拉伯数字转为傣语发音即可。若按数值来读，除了将其转为傣语读音，还需添加数值权重，如：သိဝ်（十）、ရွဲ（百）。这里以数字“234”为例来阐述其数值与数码的归一化流程。如图 3 所示。

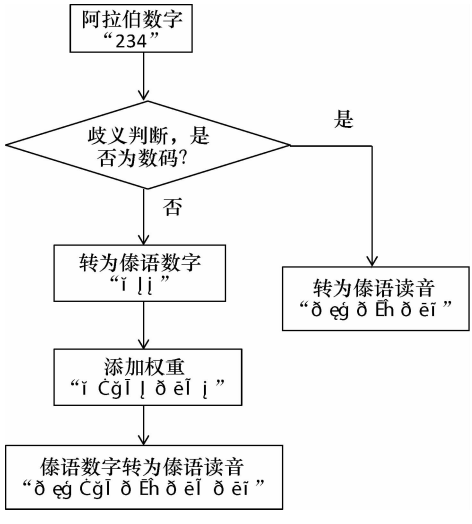

图 3　数字“234”的归一化流程图

3.4 分词

在语音合成系统中，分词的准确与否直接影响整个系统合成语音的自然度。目前，实现自动分词的算法主要包括两类：第一类是基于规则的分词方法；第二类是基于统计模型的分词方法［4］。在本次实

验中，我们分别采用了正向最大匹配（FMM）算法、基于机器学习模型的朴素贝叶斯、决策树、条件随机场方法来实现傣语分词。

FMM 算法是基于词表的算法。傣语词由单音节词和多音节词构成。我们对包含傣语常用词条的词表（含词条 16857）进行统计发现：傣语中单音节词所占比例为 21.03%，双音节词所占比例为 49.30%，三音节词所占比例为 18.66%，四音节词所占比例为 6.65%，其他词汇大部分为汉语借词。因此，傣语分词的最长词汇的长度可设置为 4。FMM 分词算法流程见图 4。

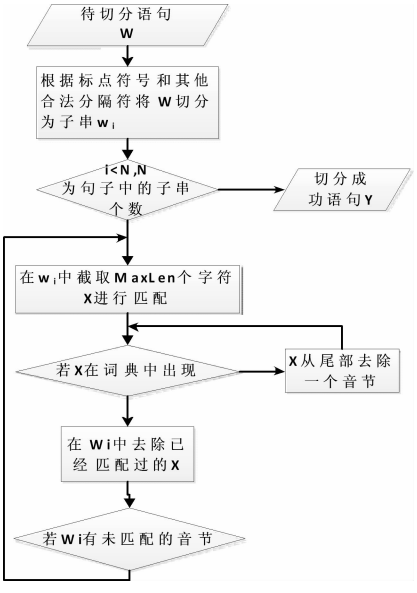

图 4　FMM 分词算法流程

基于机器学习模型的分词算法不需要借助词典。它主要是提取傣语字符属性及上下文相关信息等特征参数，通过机器学习算法进行模型训练与生成来实现分词[5]。此次实验，我们采用了朴素贝叶斯（NaiveBayes）、C4.5 决策树、条件随机场（CRFs）三种分词方法。其中朴素贝叶斯和 C4.5 决策树是利用 Weka 平台进行傣语分词，而条件随机场是用 CRF + + 工具包来实现的。基于机器学习模型实现分词的框图见图 5。

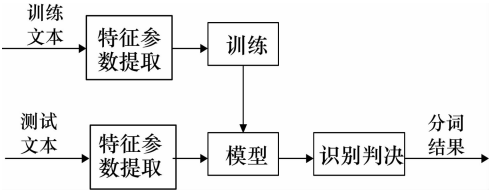

图 5　基于机器学习模型实现分词的框图

将文献［6］提供的傣语语料库用作测试集，对傣语进行分词，可得到以上算法的傣语自动分词评测结果，见表 2。

表 2　　傣语自动分词的评测结果

分词算法	正确率	召回率
FMM	93.8%	91.2%
NaiveBayes	61.0%	99.6%
C4.5	74.5%	76.1%
CRFs	93.6%	93.4%

由以上结果看出，FMM 算法的分词准确率最高，而 NaiveBayes 算法的召回率最高。而从整体分词效果来看，FMM 与 CRFs 算法的整体指标要好于 NaiveBayes 和 C4.5 算法。

3.4 文本罗马化与标注

傣语的音节由声母、韵母、声调组成。声母共有 42 个。韵母共 91 个，包括 18 个单元音，7 个复元音，66 个复合韵母。傣语中有两个显性调符ᦷ和ᦸ和一个隐性调符（即不标调号）。在进行文本分析时，需要将傣语声韵母转化为拉丁字母形式，简称为文本罗马化。我们采用音译和转录相结合的方法，参考并结合傣语声韵母国际音标和某些声韵母的发音特点制定出一套罗马化方案。

在傣语词典中，有大量的傣语词汇是从汉语音译过来的，对这类词，简称为汉语借词。而汉语借词的读音，从音素层来说，有些是按照汉语来发音的。经过统计分析，我们共找出 11 个这样的汉语音素，包括声母 4 个：j、k、r、x；韵母 7 个：ei、ian、iang、iao、uan（声母为 g，h 除外）、ia、ie。我们将这类汉语声韵母称为

"特殊声韵母"。因此我们还需要增加这些汉语音素的罗马化方案。表3展示了部分傣文与汉语音素的罗马化方案。

表3 部分傣文与汉语借词罗马化方案

傣文	国际音标	罗马化	汉语	国际音标	罗马化
[illegible]	/p/	p	sh	/ʂh/	Csh
[illegible]	/t/	t	x	/ɕ/	Cx
[illegible]	/k/	k	z	/tsh/	Cz
[illegible]	/kw/	kv	ch	/tʂh/	Cch
[illegible]	/ph/	ph	j	/tɕ/	Cj
[illegible]	/o/	o	ei	/ei/	Cei
[illegible]	/ɔ/	oa	uo	/uo/	Cuo
[illegible]	/ɯ/	ua	ou	/ou/	Cou
…	…	…	…	…	…

傣文的罗马化实现，主要分为三个步骤：汉语借词词典构建、汉语借词识别与罗马化、傣语罗马化。其流程图如图6所示。

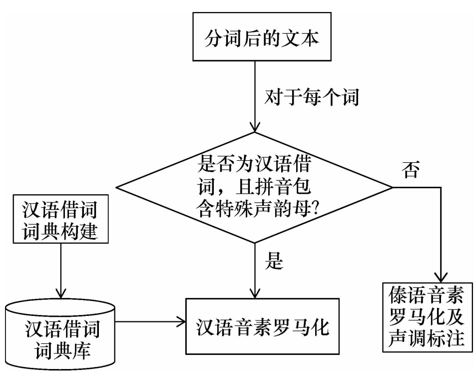

图6 傣语罗马化流程

以下是某一段傣文及其罗马化的结果：

原始傣文：

[illegible]

罗马化结果：

tseau5 vCei4 fu2 su5 Cji2 tseau5 tsaang4 tau5

对文本进行罗马化之后，需要对文本进行标注。基于音频文件的时间信息与罗马化文本，生成带时间信息的单音子monolab文件和不带时间信息的三音子fulllab文件。本实验的文本标注主要利用HMM对齐技术，通过HTK工具HVite实现自动切分，并通过Praat软件进行手工韵律调整。

3.5 上下文属性及问题集

在模型训练之前，需要对上下文属性和问题集进行设计。针对傣语具有声韵调的结构，我们采用上下文相关三音子模型从发音信息和韵律层级信息对上下文属性信息进行描述。以下是本文设计的部分问题集。

表4 部分问题集方案

问题集	说明
QS L-stop	前接爆破音声母
QS L-cNasal	前接鼻音声母
QS C-vRounded	当前为圆唇韵母
QS C-cUnvoiced	当前为清音声母
QS R-cFricative	后接擦音声母
QS R-vFront	后接前位韵母
QSL_ Tone	前接单元声调
QS C_ POS	当前单元词性
…….	………

3.6 HMM训练与合成

数据准备好后，我们就可以把训练数据送到HTS平台中训练。我们选取的平台为HTS-2.0-STRAIGHT。在本文实验中，我们选取声母和带声调的韵母为合成基本单元，对于句子中出现的静音段采用sp建模。提取训练语音数据的帧移为5ms，最终的频谱和基频特征不仅包含静态参数，

还包含一阶和二阶的时域差分参数，以更好地描述语音的动态特征。对于频谱和基频特征，使用的 HMM 为 5 状态从左至右无空跳的模型结构，对于音素时长采用单高斯分布表示。在各特征对应的上下文相关统计模型的训练过程中，我们使用决策树进行模型聚类以解决数据稀疏的问题。

在本文实验中，选用的实验运行平台为 Cygwin。我们采用 800 句标注好的语料训练 HMM，构建傣语语音合成器。在合成过程中，挑选 100 句语料作为待合成语句，其中训练集内 50 个语句，训练集外 50 个语句，具体实验参数见表 5。

表 5　　实验平台参数

参数名称	数据
HTS 版本	2.0
实验运行平台	Cygwin
输入训练数据 wave 文件	800 个
问题集文件	1 个
monolab/fulllab 标注文件	各 800 个
待合成语句	100
输出合成语音 wave 文件	100

4. 实验结果与分析

在以上实验平台下进行数据训练，并对输入的 100 个语句进行合成。在图 7、图 8 中，我们给出了某一句集内测试语句的原始的与合成的语谱图。

待合成语句：

ᦎᧂ ᦔᦱ ᦎᦸᧆ ᦂᦸᦱ ᦔᦸᦱ ᦎᦸᧂ ᦉᦹᧃ ᦠᦴ ᦉᧄ

罗马化结果：

tang4 pha1 tet8 koa6 phoa2 toang4 sean4 hau4 sam6。

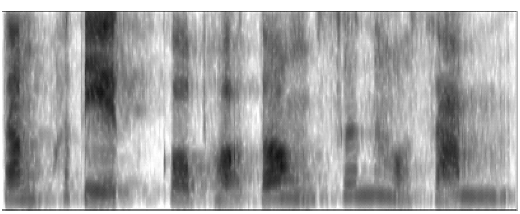

图 7　原始语谱图

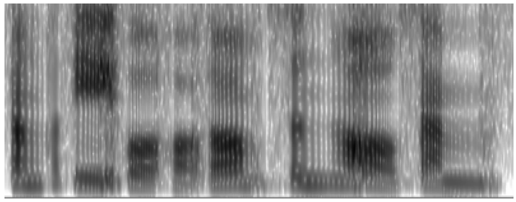

图 8　合成语谱图

由以上两句傣语的语谱图看出，两句语音的共振峰基本一致，而其韵律信息和基频信息并不一致。为了进一步评测合成的语句的可懂度和自然度评测。我们请了三位母语为傣语的学生对实验合成出的 100 个句子进行可懂度与自然度的测试。测试采用随机播放的形式。

测试者听懂每句话的意思及各个音节，则该句子的可懂度为 100%。测试结果显示，合成出的语音，不管是在集内还是在集外的语句，其可懂度都达到了听懂的要求。

自然度测试，采用主观评测方法进行打分，将每个测试者的评分均值作为最后得分。三名评测者分别对原始语音、合成语音进行了打分，其评测结果见表 6 所示。

表 6　　MOS 评测结果

	集内语句			集外语句		
原始语音	4.52	4.68	4.3	4.74	4.62	4.87
合成语音	2.53	2.78	3.20	2.23	2.58	2.67

从上述的评测结果可以看出，合成出的语音可以接受的可懂度，而自然度还处于“可以接受”和“比较自然”之间，需要进一步的研究，以提高合成语音的自然度，使其能达到实际应用的水平。

5. 结论

本文基于 HMM 语音合成框架，设计并实现了傣语的语音合成系统。本文重点阐述了整个语音合成框架的实现过程，主要包括：语料收集与挑选，挑选的语料其

音节覆盖率达到 94.2%；发音语料的录音；文本归一化；自动分词，采用 FMM 算法，分词正确率达到 93.8%；文本罗马化处理；文本自动音节切分和韵律标注；同时依据傣语的语言特点，设计了上下文属性和问题集。最后在 Cygwin 平台下，通过 HTS 合成工具，实现 HMM 训练和合成语音。

经过人工评测，合成出的语音，可懂度达 100%。而在语言自然度上还有待提高。在后续工作中我们可以对文本标注文件进行修改和调整，使得训练的文本标注更加准确，并对音频文件的边界噪声进行处理，以合成出更高质量的语音。

6. 致谢

本文研究得到了国家自然科学基金资助（项目编号：61262068）。傣语播音员刀江萍老师为本文研究录制了语音库，玉罕甩、玉喃哈等同学为本文提供了傣语支持。

参考文献

[1] Kawahara, H., Ikuyo Masuda-Katsuse, Alain de Cheveign. 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. *J. Speech Commun.* 27, 187 – 207.

[2] 玉康、张秋生、岩温龙（2006）《西双版纳傣语基础教程》。昆明：云南民族出版社。

[3] 王仁华、戴礼荣、胡郁、凌震华（2008）基于声学统计建模的新一代语音合成技术。《中国科学技术大学学报》第 7 期，734—925 页。

[4] 付敏（2010）《一个改进的中文分词算法及其在 Lucene 中的应用》，武汉：华中科技大学。

[5] 张燕平、张铃（2012）《机器学习理论与算法》。北京：科学出版社。

[6] 高廷丽、陶建华、戴红亮、李雅（2013）傣文自动分词系统的设计与实现。《中文信息学报》第 6 期，187—191 页。

方　媛　云南大学信息学院，硕士研究生，主要研究领域：语音识别、合成与理解。
E – mail：lena_ guoguo@ 163. com

杨　鉴　云南大学信息学院，博士，教授，主要研究领域：语音识别、合成与理解。
E-mail：jianyang@ ynu. edu. cn

陈志琼　云南大学信息学院，硕士研究生，主要研究领域：语音识别、合成与理解。
E – mail：410228873@ qq. com

王　昱　云南大学信息学院，硕士研究生，主要研究领域：语音识别、合成与理解。
E – mail：cduestcwy1989@ 126. com