基于 HCSIPA 的中英文混合参数合成

徐英进，贾珈，蔡莲红
（清华大学计算机系，普适计算教育部重点实验室，清华信息科学与技术国家实验室（筹）
北京，邮编：100084）

摘　要：基于双语说话人的混合合成可以解决双语合成中音色统一性的问题，但当语言差距较大时，还会存在音色的差距。本文在基于双语说话人的中英文混合合成中，提出了一种新的中英文通用音标符号-HCSIPA，提高了中文和英文的发音单元混合度，减少了语言差距带来的音色差距。同时，本文构造了针对 HCSIPA 的中英文共用问题集，提高了中英文在决策树结构上的区分度。实验结果表明，该系统可以合成出较高质量的中英文语音，而且中英文混合对单种语言合成的质量下降不明显。

关键词：HMM 语音合成，混合语音合成，音标符号，HCSIPA 符号，双语说话人，音色统一

Mandarin-English Mixed TTS Based on HCSIPA

SO YongJin，JIA Jia，CAI Lianhong
（Department of Computer Science & Technology Tsinghua University, Key Laboratory of Pervasive Computing, Ministry of Education, Tsinghua National Laboratory for Information Science and Technology，Beijing， Zip Code：100084）

Abstract: The mixed TTS based on bilingual speaker can solve the problem of timbre unity in the multi-language mixed synthesis, but there is still a difference in timbre when the gap between the languages is very large. This paper proposes a Mandarin-English general phonetic alphabet – HCSIPA in bilingual speaker based Mandarin-English TTS. It can increase the mixing with pronunciation unit of Mandarin and English, to reduce the timbre gap caused by language gap. And to increase the distinguishing between Mandarin and English in the decision tree structure, a Mandarin-English common question set suitable for HCSIPA is proposed in this paper. The experiment result shows a Mandarin-English mixed TTS based on HCSIPA can synthesis the speech of Mandarin and English with a high quality, and the quality decline to single language synthesis caused by mix of Mandarin and English is not obvious.

Keywords: HMM-based TTS, Mixed Speech Synthesis, Phonetic Alphabet, HCSIPA alphabet, bilingual speaker, timbre unity

1　引言

多语言混合合成系统是指用同一个人的声音合成出多语种语音的系统[1]。研究多个语言能共用的语言合成技术可以提高语音合成系统的通用性和扩展性，因此成为了国内外语音合成研究的一个热点。很多研究者针对不同语言的说话人是不同人的情况，提出了跨语言的说话人自适应算法[2-3]。

基于多语言说话人的语言合成是多语言混合合成中最简单、最直接的方法，该方法可以避免不同语言音色不统一的现象[4]。但是，当语言本身之间的差距很大时，即使是同一个说话人，不同语言的音色会存在区别，由此语音的音色特征为语种识别常用的声学特征，如 MFCC, LPC 等等[5]。因此提高音色统一性是多语言混合合成中核心的一个关键问题。多语

言混合合成的另一个重要问题为多语言的混合会导致单语种语音合成的质量下降，因此减少混合代价是多语言混合合成需要解决的关键问题之一。

本文针对以上的两个关键问题，提出了解决方法和策略。在尽量提高中文和英文之间的发音单元混合度的原则上，本文提出了一种新的中英文通用音标符号-HCSIPA。同时，构造了针对 HCSIPA 的中英文共用问题集，提高了中英文在决策树结构上的区分度，以减少中英文混合对单种语言合成的质量下降。

2 基于双语说话人的中英文混合合成框架

图 1 表示中英文混合合成系统的训练阶段流程。为实现中英文混合合成，首先需要构造中英文混合语料库，即要构造中英文混合合成的训练数据。中英文混合语料库的构造通过中英文语音数据的合并和中英文文本标注数据的合并实现。基于双语说话人的混合合成中，可以避免中英文语音数据合并中的音色不相同的问题，因此该合成中核心的研究点为中英文文本标注数据的合并。

中文和英文在发音结构上有较大区别，其发音单元的大小和表示方法也存在较大的不同。例如，英文合成系统一般以英语音素作为建模单元，而中文合成系统的建模单元可以是一个音节、半音节（声母和韵母）或者是音素，而即使中文和英文合成系统的建模单元都是音素，其音素之间的发音方法和表示方法都会不同。由此，将中文音素和英文音素合并为中英文通用音标是本文研究的核心内容之一。

基于中英文通用音标符号合并的中英文混合语料库经过 HMM 建模，根据不同的发音和语境特征，建立不同的 HMM 模型。此时，系统同时拥有纯英文音素的模型、纯中文音素的模型以及中英文共用音素的模型。这些大量 HMM 模型根据语境信息，进行决策树聚类，最终形成决策树结构的中英文混合模型库。在此中英文混合决策树聚类过程中，需要解决的关键问题为聚类使用的问题集构造，也就是中英文共用问题集的构造。合并中文语音合成和英文语音合成的问题集，构造中英文共用问题集是本文研究的第二个核心内容。

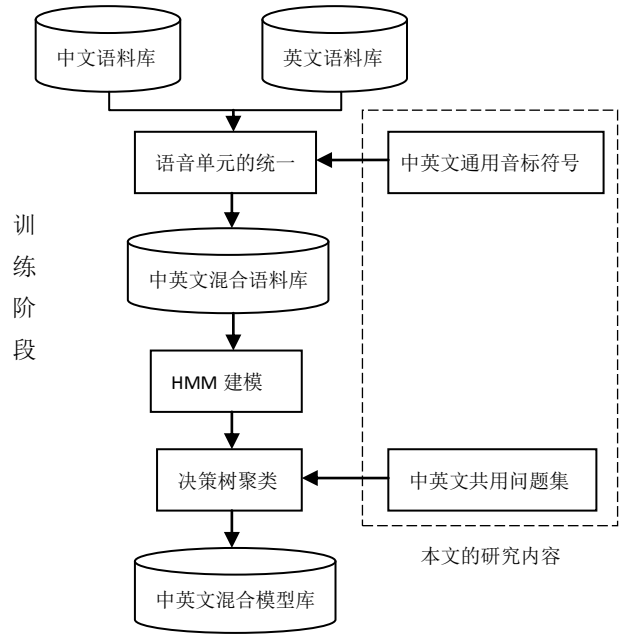

图 1. 中英文混合合成训练阶段流程

3 新的中英文通用音标符号-HCSIPA

随着语言学和语音信号处理技术的不断发展，很多研究人员关注了国际音标的计算机可

读符号化，并提出了各种机读音标符号。1995 年，将原来的 SAMPA (Speech Assessment Methods Phonetic Alphabet)扩展成 X-SAMPA (Extended SAMPA)，统一了各种语言的 SAMPA 表示方法，同时包含了所有国际音标符号[6]。同时，针对具体的单种语言，X-SAMPA 的引申形式也逐渐被提出来了，如英文的 ARPAbet 和中文的 SAMPA-C、SAMPA-SC[7]和 MCIPAbet[8]等等。

本文为了实现中英文混合合成，参考英文的 X-SAMPA 表[6]和中文的 SAMPA-SC 表[7]，构造了一种新的中英文通用音标符号集-HCSIPA。该音标符号的构造原则为：

- 以发音方式和发音部位为构造标准
- 中文和英文拥有尽量多的共同音素
- HTS 可读（只由英文字母组成）

根据以上标准，本文构造的 HCSIPA 包含 67 个音素，由 30 个元音和 37 个辅音组成，表 1 为其中 6 个音素的符号、国际音标、使用语种以及该符号的解释。

Y. Qian 在基于跨语言状态映射的中英文混合合成中，根据国际音标（IPA），构造了一种中英文通用音标符号集 [2]。由于该文很注重音标符号对中文韵律特点的表示，因此中文的大多数音素符号不同于英文音素符号，中文和英文的共同音素较少。而本文为了提高音色的统一性，更注重音素的发音方式和部位，找了更多的中英文共同发音，因此中英文之间的共同音素较多。

HCSIPA	IPA	语种	解释
ay	A	中	汉语的 a ，如“吧，卡”
va	ʌ	英	英语的短 a (strut [str\Vt])
ai	a	中，英	英语 aɪ的 a， 汉语 ai（开，赛）的 a
hx	x	中	汉语的 h（哈，和）
hh	h	英	英语的 house [haUs]
mm	m	中，英	英语的 mouse [maUs]， 汉语的 m（妈，没）

表 1. 中文和英文训练语料的平均音素个数

表 2 表示 HCSIPA 和[2]符号集的一些音素之间的区别，从表可以看出[2]符号集根据不同的韵律，将一个发音分成多个音标，而 HCSIPA 更注重发音本身，将[2]符号集的一个音标按照国际音标细化为两个音素。[2]符号集共有 91 个音标，其中中英文共用的音素只有 4 个元音和 8 个辅音，所占的比例为 14.29%，而在 HCSIPA 中，有 11 个元音和 9 个辅音为中英文共用的音素，所占的比例为 29.85%。

IPA	HCSIPA	[2]的符号集
/ɾ/	rc	/ɾ/
/ɻ/	rr	
/ɛ/	ee	/ɛ˥/ /ɛ˧/ /ɛ˨/
/a/	ai	/a˥/ /a˧/ /a˨/
/æ/	ae	

表 2. HCSIPA 和[2]符号集的比较

4 针对 HCSIPA 的中英文共用问题集

根据通用音标符号-HCSIPA 得到中英文混合语料库之后，该混合语料库通过 HMM 建模和决策树聚类，建立中英文混合模型库。而在此训练过程中需要解决的一个关键问题为构造中

英文共用问题集来实现混合决策树聚类。

中英文共用问题集的设计原则为尽量提高中文和英文在决策树结构上的区分度，更好地体现中文的韵律特点，以减少两种语言混合对单种语言合成的影响。本文根据以上设计原则，合并中文和英文合成的问题集，构造了针对 HCSIPA 的中英文共用问题集。

一般语音合成系统的问题集包含三大类问题：发音类问题、位置类问题和数目类问题。其中，位置问题和数目问题在两种合成系统里基本相同，因此，两种语言的发音类问题合并成为共用问题集设计的核心内容。

本文使用的中英文发音类问题合并方法如下：

- 中文和英文的所有发音均以 HCSIPA 的符号表示
- 语言无关的问题直接合并回答
- 增加可区分语言的问题

在发音类问题中，有关发音方式和发音部位的问题跟语言无关，可以直接合并回答。如问题“当前音素是否擦音？”在中文的回答为“ff, hx, rz, sc, ss, sx, zc, zh”，在英文的回答为“ch, dh, dz, ff, sh, ss, tx, vi, zs, zz”，而合并之后在共用问题集上的回答为“ch, dh, dz, ff, rz, sc, sh, ss, sx, tx, vi, zc, zh, zs, zz”。

另一方面，如果共用问题集的所有问题均为语言无关问题时，聚类之后的决策树在结构上很难区分出中文和英文，该结果的缺点为中英文混合对单语种合成的影响最大。本文为了提高中文和英文在决策树结构上的区分度，本文在共用问题集里增加了一些可区分中文和英文的问题，也就是语言有关问题。例如，问题“当前音素的声调是多少？”在中文中的回答取 0-4 之间的值，而在英文中的回答取 x 值，问题“当前音素的强度(stress)是多少？”的在英文中的回答取 0-2 之间的值，而在中文中的回答取 x 值。

增加可区分问题起到了中英文共用音素在决策树结构上的分离效果，有助于保证中英文混合系统合成单种语言时的语音质量。基于 HCSIPA 的中英文混合语料库通过 HMM 建模和混合决策树聚类，构建了中英文混合模型库。

5 实验及结果分析

本文使用一个双语说话人的 1,400 句中文和 1,400 句英文构建了中英文混合合成系统，并通过主观听测实验评估了混合合成系统的性能以及该系统和单语种系统之间的差距。

母语是中文且有很高英文水平的 9 名人员参加主观听测，而评测语料为训练集外的 10 句中文和 10 句英文。训练和测试的所有语音均为 16KHz 采样率的 wav 格式，HMM 训练和合成通过 HTS-2.1.1 工具和脚本进行了。在混合合成系统性能测试中，使用了 MGC 和 STRAIGHT 两种声学特征，而在混合合成和单语种合成的对比实验中，使用了 STRAIGHT 特征。使用 MGC 特征时，声学特征为 78 维向量：25 维 MFCC，1 维 log F0 以及其一阶和二阶动态特征，而使用 STRAIGHT 特征时，声学特征为 135 维向量：39 维 STRAIGHT-MFCC，1 维 log F0，5 维子带非周期成分以及其一阶和二阶动态特征。

本文的中英文混合合成系统的 MOS 评分结果如图 1，结果表明基于 HCSIPA 的中英文混合合成系统能够合成出质量较高的中文和英文。评分结果中，由于 STRAIGHT 特征在音色和激励参数的分离效果以及特征的总维度上优于 MGC 特征，因此合成语音明显好于 MGC 特征。而中文的评分普遍高于英文的评分，其原因为英文训练语料相比于中文训练语料要短一些，而且测试人员对中文和英文的认知和理解程度还是存在一些差距。

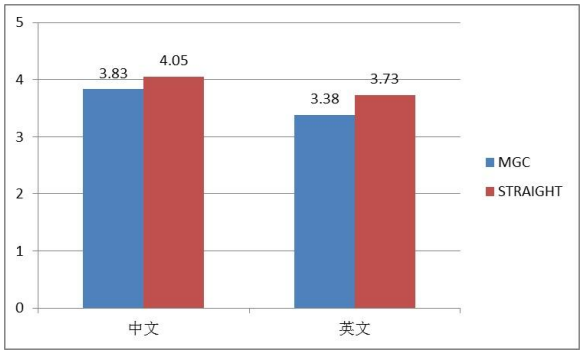

图 2. 中英文混合合成系统的 MOS 评分结果

图 2 表示本文的中英文混合合成系统随着合成文本中不同的中英文混合度，和单语种合成系统进行对比的 DMOS 评分结果。图的横轴为合成文本中包含中英文共用音素的比例，我们选择了分布在 33%-100%之内的 10 句文本进行对比实验，图的纵轴为混合合成系统与单语种合成系统之间的 DMOS 评分。

评分结果表示随着中英文共用音素的涵盖率提高，合成语音的质量会有下降趋势，但即使共用音素占 100%，合成语音的 DMOS 评分也会达到 4.5 分左右，即合成语音的质量下降不明显。该结果说明针对 HCSIPA 的中英文共用问题集可以有效的区分中英文共用音素在混合决策树上的结构，同时说明本文的中英文通用音标符号 HCSIPA 是很合理的，即使中英文共用音标占整个音标集的 29.85%，但这些共用音标对单语种合成质量的影响很小。

在混合合成系统和单语种合成系统的 DMOS 评分结果中，在共用音素涵盖率较小的情况下，英文的质量下降大于中文的质量下降，其原因与图 1 中的中文合成结果优于英文合成结果的原因相同。

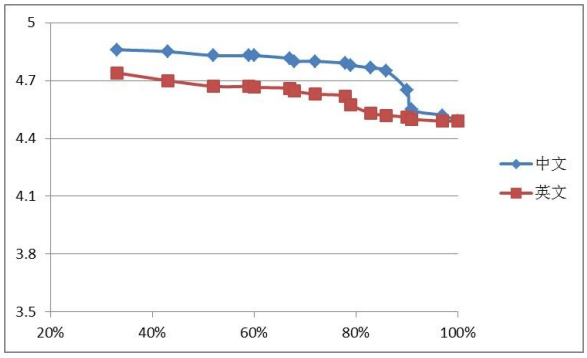

图 3. 混合系统与单语种系统的 DMOS 评分结果

6 结束语

多语言混合合成系统的研究是目前参数语音合成中的重点研究方向之一。本文在基于双语说话人的中英文混合合成中，提出了一种新的中英文通用音标符号-HCSIPA，提高了中文和英文之间的发音单元混合度，减少了语言差距带来的音色差距。同时，本文构造了针对 HCSIPA 的中英文共用问题集，提高了中英文共用音素在决策树结构上的区分度，减少了语言混合对单语种合成的影响。

实验结果表明，基于 HCSIPA 的中英文混合参数合成系统可以合成出音质和自然度较高的中文和英文，合成语音的 MOS 评分平均值为 3.75 分。同时，中英文共用问题集有效地分离了中英文共用音素在中文和英文中的作用，混合合成系统与单语种合成系统的 DMOS 评分平均值为 4.67 分。

7 致谢

文章受国家自然科学基金(61003094, 90820304, 90920302)的资助。

参考文献

[1] C. Traber et al. From multilingual to polyglot speech synthesis. In Proc. of Eurospeech, 1999, pp. 835–838.

[2] Y. Qian, H. Liang, and F.K. Soong. A cross-language state sharing and mapping approach to bilingual (Mandarin-English) TTS. In IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, vol. 17, no. 6, pp. 1231-1239, 2009.

[3] H. Liang et al. A cross-language state mapping approach to bilingual (Mandarin-English) TTS. In Proc. of ICASSP, pp. 4641-4644, 2008.

[4] Y. Zhang and J. Tao. Prosody modification on mixed-language speech synthesis. In Proc. of ISCSLP, pp. 1-4, 2008.

[5] B. Yin, E. Ambikairajah and F. Chen. Combining cepstral and prosodic features in language identification. In Proc. of ICPR, pp. 254-257, 2006.

[6] Zhang Jialu. A SAMPA system for Putonghua(Standard Chinese). In Proc. of Oriental COCOSDA'99，89—92, Taipei, Academia Sinica, 1999

[7] 张家騄. 汉语普通话机读音标 SAMPA-SC. 声学学报. 2009，34(1)：81-86

[8] Y. Zu et al. A Super Phonetic System and Multi-Dialect Chinese Speech Corpus for Speech Recognition. In *Proc.* of ISCSLP, 2006.

基金项目：国家自然科学基金(61003094, 90820304, 90920302）

作者简介：
徐英进，1984，男，博士研究生，主研方向：语音合成和声音转换；
贾珈，讲师、博士；
蔡莲红，教授、博士生导师；