汉语语音合成中说话人自适应的时长优化

徐英进，贾珈，蔡莲红

（清华大学计算机科学与技术系，北京 100084）

摘 要： 在汉语语音合成中，音节内清音和浊音的时长是影响自然度的重要因素、并且与说话人关系较大的个性化特征之一。本文针对基于 HMM 的汉语语音合成说话人自适应，提出了一种清浊音时长优化算法。将原始说话人训练语料的清音在音节中的相对时长特征根据语境特征进行决策树聚类，并进一步使用自适应算法将决策树中的特征值自适应到目标说话人的清音相对时长。在语音合成时，从该决策树得到目标说话人的清音相对时长参考值，合成语音的清浊音时长按照参考值进行调整。实验表明该算法可以提高 HMM 汉语语音合成中说话人自适应的时长预测准确度，有效地提高说话人自适应的相似度和合成语音的自然度。

关键词：汉语语音合成；说话人自适应；时长优化；清浊音；

中图分类号：TN912.33

The duration optimization of speaker adaptation in Mandarin TTS

SO Yongjin, JIA Jia, CAI Lianhong

(Computer Science and Technology Department,

Tsinghua University, Beijing 100084, China)

Abstract： In Mandarin TTS, the duration of unvoiced and voiced phonemes in a syllable is a very important factor related to the naturalness of synthesized speech. It also is a personalized feature has the great relation with the speaker. This paper proposes an unvoiced/voiced duration optimization approach for the speaker adaptation in HMM-based Mandarin TTS. The relative duration of unvoiced part at a syllable in the corpus of source speaker is clustered with context features. This decision tree is adapted by target speaker using the relative duration of unvoiced part in the adaptation data. In synthesis, a reference relative duration of unvoiced part with the target speaker is generated from this decision tree, and the duration of unvoiced part and voiced part in the synthesized speech is adjusted accordingly. Experiments show that this approach can improve the accuracy of duration prediction in the speaker adaptation of HMM-based Mandarin TTS, and it can effectively improve the similarity of speaker adaptation and the naturalness of synthesized speech.

Key words: Mandarin TTS; speaker adaptation; duration optimization; unvoiced/voiced sound

说话人自适应是基于语音模型转换的一种声音转换技术，HMM 语音合成的说话人自适应相对经典的基于声学特征转换的声音转换，需要的目标说话人语音数据量更少，并且在保证很高的声音相似度的同时，也保持了较高的音质和自然度。

Tokuda 在 1998 年提出了基于最大似然线性回归(Maximum Likelihood Linear Regression, MLLR)的说话人自适应算法（Speaker Adaptation）[1]。

[1]**收稿日期：** 2012-10-22

基金项目：国家自然科学基金资助项目(60928005，60931160443)

作者简介：徐英进（1984－），男（朝鲜），朝鲜，博士研究生。

通信作者：蔡莲红，教授，E-mail：clh-dcs@tsinghua.edu.cn

YAMAGISHI 等人在 2003 年提出了平均声音模型(Average Voice Model)的概念以及其模型训练方法和决策树聚类方法[2][3]，并将 MLLR 算法引进到平均声音模型上[4]。Zen 等人在 2004 年提出了 HSMM(Hidden Semi-Markov Model)的概念[5]，并将 MLLR 算法应用到 HSMM 上[4]。在算法本身的性能提高上，Nakano 等人在 2006 年提出了将经典的 MAP(Maximum a Posteriori) [6]算法和 MLLR 算法结合的方法[7][8]。但以往说话人自适应算法的研究都比较注重转换算法，也就是注重提高模型的转换准确率，转换参数一直局限在 MFCC 等频谱参数和基频参数。

考虑到汉语语音中清浊音时序性，在基于 HMM 的汉语合成中，有些研究通过清浊音判决的改进，减少清浊音预测的误差。康世胤等人在音节级汉语语音合成系统中，统计了训练语料的音节内清音相对时长，构造了根据拼音的清音相对时长参考值表，指导合成时的清浊音判决[9][10]。在音素级汉语语音合成系统中，Qian Yao 等人提出了清浊音累积误差(accumulated u/v error)，描述“清音-浊音”双音素序列中清浊音转换点位置误差，并以最小误差状态边界作为最终清浊音的转换边界[11]。清浊音判决的改进可以减少激励和频谱的清浊性不一致的问题，但无法修正清浊音时长的预测误差。同时，由于清浊音判决的改进算法是基于大量统计数据的，难以应用于只有少量目标说话人语料的说话人自适应。

在基于 HMM 的汉语语音合成中，基元（音节、声韵母、音素等）的时长由时长模型预测出来的状态时长序列决定。在音素级的中文语音合成系统中，每个音节被分解成为一个或多个音素（复合韵母将被分解为若干音素），每个音素的时长是独立预测的。由于其时长预测会存在误差，会影响音节内清音和浊音的时长预测。另一方面，音节内的清浊音时长是与说话人相关的个性化声学特征之一。因此，汉语语音合成的说话人自适应中，合成语音的清浊音时长分布会影响说话人自适应的相似度。

本文针对音素级的 HMM 汉语语音合成和说话人自适应，提出了一种音节内清浊音时长的优化算法。首先，从原始说话人的大量训练语料中提取清音在音节中的相对时长特征，并根据语境特征进行了决策树聚类。同时，从少量的自适应语料中，提取清音相对时长特征，进行了对原始说话人清音相对时长决策树的说话人自适应。合成时，从该决策树得到目标说话人的清音相对时长参考值，合成语音的清浊音时长按照参考值进行调整。

1 音节内清音相对时长的个性化分析

在汉语语音中，当音节时长变化时，音节内清音和浊音的变化比例并不相同，清音时长变化和其发音等语境特征和说话人关系较为密切，而浊音时长变化则根据音节和清音时长的变化而定[12][13]。不同说话人表现不同的说话风格、情感表达和语速，相应地其清音时长也会不同。即对于不同的说话人，即使说话的文本内容是相同的，音节内的清浊音时长分布也会不同。

因此，清音对音节的相对时长是表示不同说话人和不同语境下时长变化的重要声学特征。本文选用清音时长与音节时长的比例 r_{uv} 表示清音相对时长：

$$r_{uv} = \frac{t_u}{t_s} \tag{1}$$

式中，t_s 为该音节的音节时长，t_u 为该音节的清音段时长。

为了测量音节内清音相对时长与说话人个性之间的关系，本文进行了不同说话人清音相对时长的对比分析。在分析实验中，采用自然语音与合成语音之间的清音相对时长平均误差，对比了不同说话人对清音相对时长特征的影响。清音相对时长平均误差为同文本的自然语音与合成语音各对应音节的清音相对时长差值的平均值：

$$d = \frac{\sum |r'_{uv} - r_{uv}|}{N} \tag{2}$$

其中，d 为清音相对时长平均误差，N 为总音节数，r'_{uv} 与 r_{uv} 分别为合成语音与录音中对应音节的清

音相对时长。

实验分别使用说话人 A 和说话人 B 的 5,000 句语料合成语音，选取了 50 句说话人 A 的合成语音和 50 句说话人 B 的合成语音，分别计算了清音相对时长平均误差。实验结果如表 1 所示。

表 1 不同说话人的清音相对时长平均误差

	说话人 A	说话人 B
平均误差值	0.10769	0.176687

通过表 1 可以看出，不同说话人对清音相对时长的影响是不同的。结果验证清音相对时长特征是与说话人相关的个性化声学特征。

2 基于清音相对时长自适应的时长优化

2.1 算法框架

本文在汉语语音合成的说话人自适应中，增加了清音相对时长特征的说话人自适应，提高了合成语音的清浊音时长预测准确度和说话人相似度。

添加清音相对时长的说话人自适应后，语音合成流程如图 1 所示。原始说话人 A 的大量语料经过 HMM 建模和决策树聚类，构造 A 的参数化模型库（包含基频、频谱和时长特征的决策树）和清音相对时长模型库（决策树结构）。从目标说话人 B 的少量语音中提取基频、频谱、时长等声学特征和清音相对时长特征，采用其基频、频谱、时长声学特征对 A 的参数化模型库进行自适应，而其清音相对时长特征对 A 的清音相对时长模型库进行说话人自适应。最后，在合成阶段，根据自适应后的清音相对时长决策树调整每个音素模型的状态时长。

本文的清音相对时长采用了单状态 HMM 进行建模，其自适应原理和一般 HMM 语音模型的说话人自适应相同。

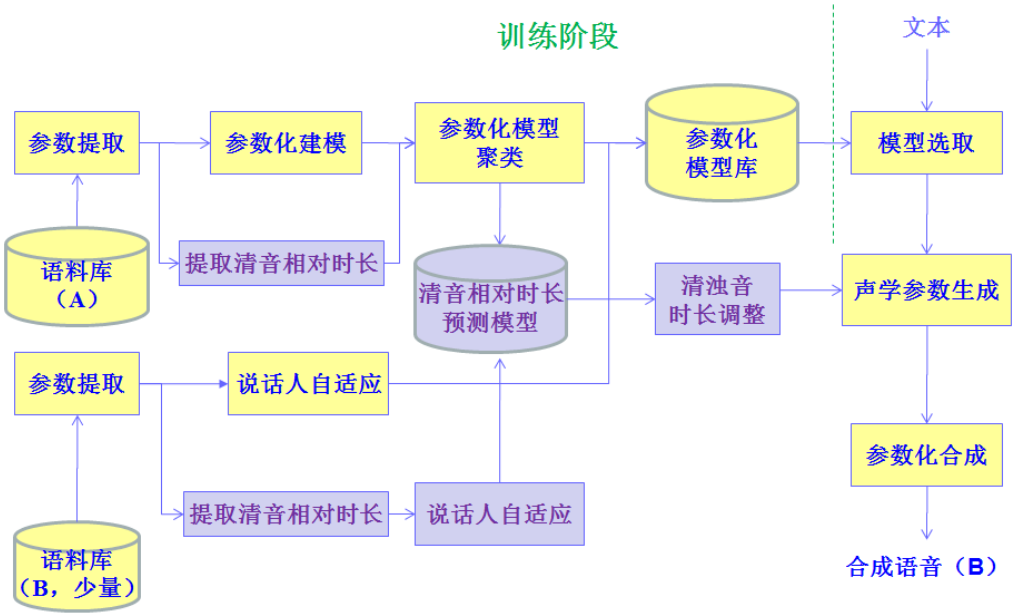

图 1 具有清音相对时长自适应的汉语语音合成流程

2.2 清音相对时长的预测建模

音节中清音时长不仅与其拼音有关，并与其上下文语境有关。如送气清音时长比不送气清音长，孤立音节的清音时长比词语中音节的清音时长长。本文采用决策树建立了基于语境特征的音节内清音相对时长预测模型。

首先，在基于 HMM 的汉语语音合成的训练阶段，根据训练语料的基频标注信息，提取了清音相对时长特征。之后使用上下文语境信息对音节的清音相对时长特征进行决策树聚类，聚类所使用的语境特征包括发音特征、位置特征和长度特征。发音特征包含当前音节和前后音节的拼音以及发音方式

和部位等；位置特征包含当前音节在韵律词、韵律短语和语句中的前向位置和后向位置等；长度特征包含当前韵律词、韵律短语和语句的音节数等。

聚类之后的预测模型为决策树结构，决策树的每个叶节点中包含相同语境的音节。每个叶节点中所有音节的清音相对时长特征被训练得到一个GMM模型，其GMM模型的均值为该叶节点的清音相对时长预测值 $r_{uv}^{'}$ 。叶节点GMM模型的分布服从：

$$p(x|\lambda)=\sum_{k=1}^{K}\omega_k N(x|\mu_k,\Sigma_k) \tag{3}$$

式中，ω_k 为第 k 个高斯分量的权重，$N(x|\mu_k,\Sigma_k)$ 为高斯密度函数。

本文将清音相对时长的 GMM 模型看作为单状态的HMM模型，并直接使用HTS工具进行了训练。

2.3 清音相对时长的说话人自适应和时长优化

在清音相对时长的个性化分析中，已验证了该特征与说话人之间的相关性。为了提高合成语音的说话人相似度和自然度，本文在基频和频谱参数的说话人自适应中，增加了清音相对时长特征的说话人自适应。

在 HMM 语音合成的说话人自适应当中，频谱特征和基频特征采用 HTS-2.1.1 工具中的CSMAPLR(Constrained Structural Maximum A Posteriori Linear Regression)算法，而本文的清音相对时长模型也采用了该算法。

在合成阶段，首先根据 HMM 模型预测得到状态时长序列，即当前音节的清音和浊音的时长。然后根据清音相对时长预测模型预测的清音相对时长参考值，并对音节内清音和浊音的时长进行了调整。

当音节时长 t_s 确定后，通过该音节的清音相对时长参考值 $r_{uv}^{'}$ ，先计算清音段的目标时长为：

$$t_u^{*}=r_{uv}^{'}\cdot t_u \tag{4}$$

式中，t_u 为该音节的清音段时长。根据清音段目标时长，得到浊音段的目标时长为：

$$t_v^{*}=t_s-t_u^{*} \tag{5}$$

确定清音和浊音的目标时长之后，当前音节的所有音素的 HMM 状态时长按照清音和浊音的时长调整比例进行相同比例的调整。

3 实验及结果分析

3.1 实验环境

本文使用一个女性说话人（说话人A）的5,000句训练语料，建立了音素级的HMM中文合成系统。另一个女性说话人（说话人B）的500句语料使用于说话人自适应。

语料库中的训练语音数据保存为16 KHz采样率的wav格式文件。使用HTS-2.1.1工具及脚本建立了两个具有说话人自适应的音素级 HMM 汉语语音合成系统 S_1 与 S_2。在 S_2 系统中增加了清音相对时长的说话人自适应和清浊音时长调整模块，而 S_1 系统中不包含清音相对时长的自适应和调整。

合成系统使用的声学特征为 25 维 MFCC[14]，1维 log F0 及其一阶和二阶动态特征，共78维。每个音素包含5个状态。

本文通过客观实验和主观实验，比较了原始说话人自适应系统 S_1 和清浊音时长优化之后的合成系统 S_2 的合成语音。

3.2 客观实验

本文在目标说话人的训练集内随机选择了50句语音，比较了原始自适应系统 S_1 和具有清音相对时长自适应的自适应系统 S_2 的清音相对时长平均误差值。为了评估目标说话人的自适应数据量对自适应结果的影响，分别选择了50句、100句和500句的数据量，进行频谱、基频、时长特征和清音相对时长特征的说话人自适应。

在中文语音合成的说话人自适应中，随着不同自适应数据量的、清浊音时长优化之前和之后的清音相对时长平均误差结果如图 2 所示。结果表明，经过清浊音时长优化之后，清音相对时长特征的预

测误差明显减少，可以有效地提高说话人自适应的相似度。

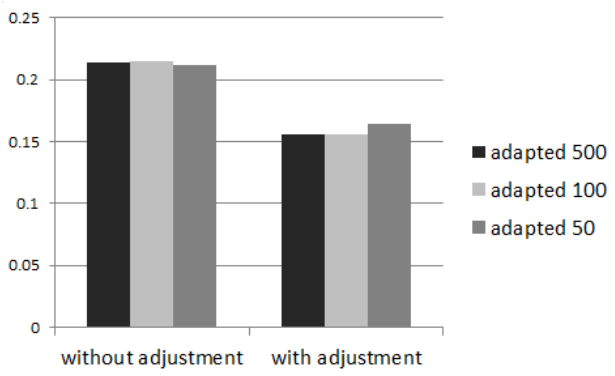

图 2 在说话人自适应中，随着不同自适应数据量的、清浊音时长优化之前和之后的清音相对时长平均误差比较

此外，不同的自适应数据量对说话人自适应结果的影响很小，结果说明清浊音时长优化算法可以使用于少量目标说话人语音的说话人自适应中，实际上，当只有 50 句目标说话人语料时，算法已经达到了很好的自适应效果。

3.3 主观实验

本文采用了 ABX 偏好性选择实验的方法，验证了具有清音相对时长说话人自适应的时长优化算法的有效性，主观听测实验由 9 个母语为汉语的专业人员进行评测。从训练集外选择了 8 句语音，比较了自适应系统 S_1 和 S_2 的合成语音的相似度。频谱、基频、时长以及清音相对时长特征的说话人自适应采用 500 句目标说话人的自适应数据进行了。对每一句测试语音，三个选项中选择一个，三个选项分别为 1- 自适应系统 S_1 更接近于目标说话人，2-自适应系统 S_1 和 S_2 在说话人相似度上无法分辨，3- 自适应系统 S_2 更接近于目标说话人。

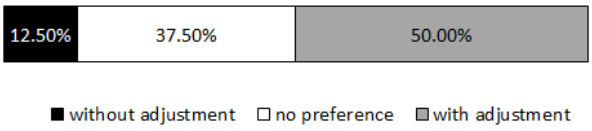

图 3 清浊音时长优化算法的偏好性实验结果

实验中，自适应系统 S_1 和 S_2 的一对测试语音是随机顺序播放的，以保证实验的公正性和客观性。主观测听实验结果如图 3 所示，结果表示增加清音相对时长的说话人自适应，可以有效地提高说话人自适应的相似度，即语音合成中的个性化表现力。

4 结论

音节内的清音相对时长是影响说话人个性和语音自然度的一个重要声学特征。本文针对基于 HMM 的汉语语音合成的说话人自适应，提出了基于清音相对时长自适应的清浊音时长优化算法。在频谱和基频声学特征的说话人自适应的基础上，增加了音节内清音相对时长特征的自适应。合成时，根据目标说话人的清音相对时长参考值，调整合成语音的清音和浊音的时长。

客观实验和主观实验结果表明，经过清浊音时长优化之后，目标说话人的清音相对时长特征预测误差明显减少，可以有效地提高说话人自适应的相似度，即语音合成中的个性化表现力。

参考文献 (References)

[1] M. Tamura, T. Masuko, K. Tokuda, T. Kobayashi, Speaker adaptation for HMM-based speech synthesis system using MLLR, Proc. ESCA/COCOSDA Workshop on Speech Synthesis, pp.273-276, Nov. 1998.

[2] Junichi YAMAGISHI,Masatsune TAMURA,Takashi MASUKO,Keiichi TOKUDA,Takao KOBAYASHI. A Training Method of Average Voice Model for HMM-Based Speech Synthesis. IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences, 2003, Vol.E86-A No.8 pp.1956-1963.

[3] Junichi YAMAGISHI,Masatsune TAMURA,Takashi MASUKO,Keiichi TOKUDA,Takao KOBAYASHI. A Context Clustering Technique for Average Voice Models. IEICE TRANSACTIONS on Information and SystemsVol.E86-DNo.3pp.534-542, 2003.

[4] Junichi YAMAGISHI, Takao KOBAYASHI. Average-Voice-Based Speech Synthesis Using HSMM-Based Speaker Adaptation and Adaptive Training, IEICE TRANSACTIONS on Information and Systems Vol.E90-D No.2 pp.533-543, 2007.

[5] Heiga Zen, Keiichi Tokuda, Takashi Masuko, T. Kobayashi, T. Kitamura, Hidden semi-Markov model based speech synthesis, in Proc. of International Conf. on Spoken Language Processing 2004, vol.II, pp.1397-1400, Oct. 2004.

[6] J. Gauvain and C. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[7] K. Ogata, M. Tachibana, J. Yamagishi, and T. Kobayashi, "Acoustic model training based on linear transformation and MAP modification for HSMM-based speech synthesis," in *Proc. ICSLP'06*, Sep. 2006, pp. 1328–1331.

[8] Y. Nakano, M. Tachibana, J. Yamagishi, and T. Kobayashi, "Constrained structural maximum a posteriori linear regression for average-voice-based speech synthesis," in *Proc. ICSLP'06*, Sep. 2006, pp. 2286–2289.

[9] 康世胤，段全盛，双志伟，等. HMM 语音合成中基频清浊音优化算法研究[C]//康世胤. 全国人机语音通讯学术会议论文集. 兰州：兰州大学出版社，2009. 317-321.

Kang Shiyin, Duan Quansheng, Shuang Zhiwei, et al. The research on voiced/unvoiced decision algorithm for HMM-based speech synthesis[C]//Kang Shiyin. Proc of NCMMSC. Lanzhou: Lanzhou University Press, 2009. 317-321

[10] Kang Shiyin, Shuang Zhiwei, Duan Quansheng, et al. Voiced-Unvoiced decision algorithm for HMM-based speech synthesis[C]//Kang Shiyin. Proceedings of Interspeech. Brighton: UK, 2009. 412-415.

[11] Qian Yao, Frank Soong, Wang Miaomiao, et al. A minimum V-U error approach to F0 generation in HMM-based TTS[C]. Qian Yao. Proceedings of Interspeech. Brighton: UK, 2009. 408-411.

[12] Jia Jia, Xu Jun, Xu Yingjin, et al. A speech modification based singing voice synthesis system[C]//Jia Jia. Proc of NCMMSC. Lanzhou: Lanzhou University Press, 2009. 446-450.

[13] Liu Yuxiang, Jin Zeyu, Jia Jia, et al. An automatic singing evaluation system[J]. Applied Mechanics and Materials, 2011, 128(1): 504-509.

[14] R McAulay, T Quatieri. Speech analysis-synthesis based on asinusoidal representation[J]. IEEE Trans Signal Process, 1986, 34(1): 744-754.