HMM语音合成中基频清浊音优化算法研究*

康世胤[1]，段全盛[1]，双志伟[2]，秦勇[2]，蔡莲红[1]
(1. 清华大学计算机科学与技术系，北京 100084；2. IBM 中国研究中心，北京 100094)

文　摘：本文提出一种用于 HMM 参数化语音合成的针对清浊音优化的基频建模和预测方法。在参数化合成方法中，清浊音预测直接决定激励源的选择，对合成质量有关键影响。针对这一问题，该方法从基频参数提取和预测两个方面同时入手，使用语料标注信息参与基频提取，建立音节清浊音转换时刻的高斯混合模型预测基频，改善清浊音判决质量。合成语音的听测实验表明，该方法与原系统相比，合成音质和韵律都有较大改善，MOS 评分由 3.0 升至 3.5。

关键词：语音合成；隐马尔可夫模型；基频建模；高斯混合模型；清浊音判决
中图分类号：TN912.3

近年来，基于隐马尔可夫模型（HMM）的参数化语音合成技术[1,2]受到了广泛的重视，发展非常迅速。与传统的基于大语料库的单元拼接合成方法相比，HMM 语音合成技术与发音人和语种的相关性小，占用资源少，构建成本低，有利于快速建立语音合成系统；此外，这种技术使用参数化的合成方法，灵活性强，音色和韵律特征都易于调整。

常规 HMM 语音合成技术分为训练和合成两个阶段。在训练阶段，首先提取训练语料的谱参数和基频参数，联合语音参数的动态特征，建立隐马尔可夫模型，并根据语境信息生成模型预测决策树[3]。其中，基频参数包含清浊音变化，不是一般意义下的连续参数，因此使用多空间概率分布 HMM（MSD-HMM）[4]进行建模。MSD-HMM 使用不同的代数空间分别刻画基频的清音和浊音参数。在合成阶段，首先对待合成文本进行文本分析，根据语境信息预测得到模型状态序列，应用语音参数生成算法[5]生成语音参数序列，最后通过参数合成器合成目标语音。其中，基频参数清浊音的判决根据最大似然准则，由 MSD-HMM 中的空间出现概率决定。

与拼接合成方法不同，在参数化合成方法中，基频清浊音预测直接影响到参数化合成器激励源的选择，因此对合成质量起到关键作用，而基频参数建模和预测的精确度受两个环节影响较大。其一，基频参数估计算法准确率有限。这一点在朗读波动较大的训练语料中表现的尤为明显。目前使用的通用基频估计算法仅仅依靠语音数据估计基频参数，未考虑语料标注等先验知识所能提供的信息，因而对部分特定发音单元的估计准确率不足，最终影响到这部分发音单元的合成效果。其二，MSD-HMM 各状态输出概率分布间的独立性与基频变化的连续性之间存在矛盾。由于 HMM 中相邻状态的清浊音判决相互独立，在一个发音单元中，预测出的基频曲线经常分裂为几段。表现最为明显的就是错误的把浊音段预测为清音，导致合成语音沙哑不清晰。由于统计模型精确性的关键作用，其他研究人员在现有 HMM 模型基础上，提出若干改进模型[6,7,8]，但都不能很好的解决清浊音预测的问题。

为此，本文提出一种用于 HMM 语音合成的基频清浊音优化算法。从前述两个环节同时入手，改善清浊音提取和预测效果，进而提高合成语音质量。

文本内容组织如下：第一部分介绍针对基频清浊音建模和预测优化的 HMM 语音合成系统结构；第二部分阐述针对清浊音优化的基频建模和预测算法；第三部分给出一个合成系统的实例，并进行实验评测；第四部分进行总结。

1　语音合成系统结构

本文构建的合成系统基本框架如图 1 所示。系统以音节为基本发音单元建立 HMM 模型，语音参数共 78 维，包括 24 阶 LSF 系数、对数基频值，及其一阶、二阶差分。

在训练阶段，使用有标注指导的基频参数提取算法提取基频，与提取得到的谱参数一起，在标注信息的指导下训练生成 HMM 模型。

*基金项目：国家自然科学基金（60805008, 90820304）；国家 863 课题（2007AA01Z198）；国家 973 课题（2006CB303101）
作者简介：康世胤（1984-），男（汉族），新疆，硕士研究生。
通讯联系人：蔡莲红，教授，E-mail：clh-dcs@mail.tsinghua.edu.cn；双志伟，研究员，E-mail：shuangzw@cn.ibm.com

同时，根据基频参数中的清浊音信息，在切分标注的指导下，计算发音单元清浊音转换时刻，建立发音单元的清浊音转换时刻 GMM 模型。

在合成阶段，对待合成文本进行文本分析，根据语境信息预测得到 HMM 状态序列，使用清浊音转换时刻 GMM 模型进一步预测状态的清浊音属性，应用语音参数生成算法生成语音参数序列，最终通过参数合成器合成高质量的目标语音。

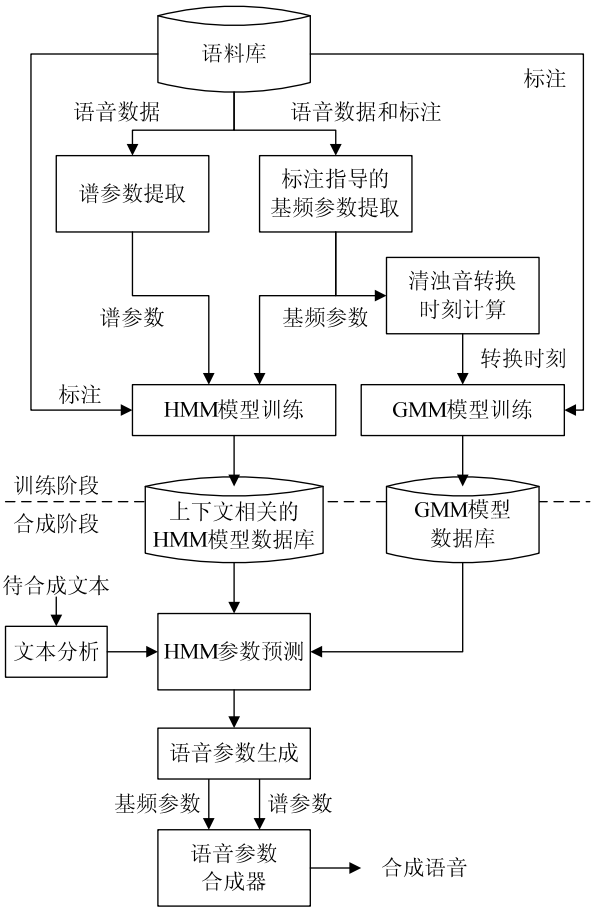

图 1　合成系统框架

2　针对清浊音优化的基频建模和预测算法

语音合成中，基频提取和清浊音界定是两个关键的基础问题。而在参数化合成方法中，基频清浊音的预测更是关系到激励源的选择，其预测准确率对合成音质有很大影响。为了得到更高质量的合成语音，本文同时从基频提取和预测两个方面改进现有方法。

2.1　标注信息指导下的基频提取

在常规 HMM 语音合成技术中，基频提取算法仅仅使用语料库中的语音数据估计基频参数，而由于基频提取算法的限制，经常会出现基频提取不准确甚至清浊音判断出错的问题。若在现有基频提取方法的基础上，考虑基频曲线固有的连续性，并同时考虑标注信息，能够在很大程度上改善基频提取结果。

通过对汉语的音节结构分析，音节总是先出现清音，随后才是浊音；有时也可以不出现清音，只有浊音。具体可以划分为表 1 所示的三种组合方式。这就是说，汉语音节的声韵母标注信息就能决定基频的清浊音界定。

表 1　汉语音节清浊音组合方式

组合方式	举例
清音声母+浊音韵母	“开” (kai)
浊音声母+浊音韵母	“来” (lai)
浊音韵母	“应” (ying)

例如，上声音节“吕（lv3）”在音高最低点处有较大概率提取不到基频，从而被当成清音段处理。若考虑其发音标注，则很容易得到，“吕”是一个全浊音音节。这就为后续对基频曲线进行修补提供了依据。

当语料库中包含音素或半音等较小尺度的切分信息时，每个发音单元中不包含清浊音变化，可以直接根据发音标注确定发音单元的清浊音属性。而对于仅有音节等较大尺度切分信息的语料库，由于一个切分单元内有清浊音转换，单凭现有信息不足以判断语音单元的清浊属性，因此可以使用语音识别工具进行较小单元的自动切分，使用自动切分的结果估计语音的清浊音属性。

本文尝试使用以汉语声韵母为单元的语音识别工具，对仅有音节标注的语料库进行声韵母边界的自动切分，并在此基础上对原始语音的基频参数进行修正。根据音节标注信息和切分信息，确定原始语音的清音段和浊音段划分。由于自动切分中使用了手工标注的音节切分信息做指导，自动切分误差造成的影响可以忽略。实验表明，原始基频提取结果中的错误主要出现在浊音段，因此基频修正主要针对浊音段进行。对于浊音段基频数据，首先使用 5 点中值滤波进行平滑，剔除残留的倍频和半频数据，然后使用分段二次曲线进行数据拟合，并对浊音单元中未提出基频的帧使用插值方法进行修补。

图 2 给出了标注信息指导下的基频提取算法对一个典型的上声音节“点儿”的基频提取结果与常规方法的对比。其中标注信息包括发音标注和音节切分数据；使用自动切分工具，还可以得到声韵母切分数据。在常规方法中，上声音节的基频提取往往比较困难，经常出现基频提取不到的现象。在我们的方法中，利用标注信息和识别工具得到声韵母切分边界，根据发音标注确定浊音段“ianr”，并对

浊音段中的基频间断进行插值和平滑处理，从而得到较好的基频提取结果。

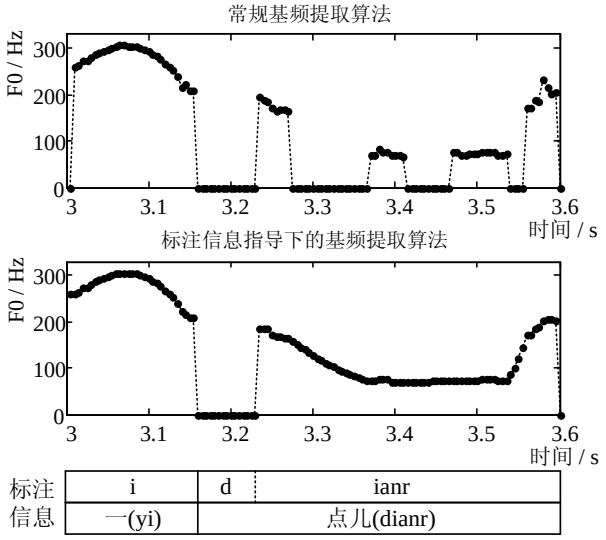

图 2　标注信息指导下的基频提取

2.2　基于GMM的清浊音预测

HMM 语音合成技术中，发音单元的语音参数向量序列是由 HMM 各状态的输出概率分布来描述的。注意到，虽然 HMM 相邻状态间具有相关性，但是各状态输出概率分布则是相互独立的。这种独立性与语音参数向量随时间变化的连续性之间存在矛盾。针对这一问题，文献[3,9]在建立 HMM 模型时中引入语音参数向量的动态特征，结合相应的参数生成算法，保证了浊音段基频曲线变化的连续性。但是，在描述基频参数的 MSD-HMM 中，浊音空间出现概率这一参数独立存在，缺少相应的机制来保证浊音段的连续性。

在这种情况下，即使采用了前述改进的基频提取算法，基频预测环节仍然有可能在真实的清浊音转换点附近出现清浊音状态交错。要解决这一问题，就必须在较大尺度上建立一个约束 HMM 状态间时间连续性的模型。

针对以音节为基本发音单元的 HMM 汉语语音合成系统，本文使用音节的归一化清浊音转换时刻作为特征参数建立 GMM 模型。

进一步，对于任何语种，只要训练 HMM 使用的基本发音单元中，清浊音转换点不多于 1 个，本文提出的方法就具有普适性。

为尽可能保证 HMM 参数生成算法对于基频参数和普参数预测的一致性，本文并不直接使用 GMM 预测音节的清浊音转换时刻，而是从现有 HMM 参数生成算法中取得候选的清浊音转换点，应用相应音节的 GMM，根据最大似然准则选出最优点，进而得到的音节的清浊音预测结果。具体做法如下。

首先使用基频参数计算音节的归一化清浊音转换时刻：

$$x=\frac{s_{unvoiced}}{n_{end}-n_{start}+1} \tag{1}$$

其中 n_{start},n_{end} 分别是音节起始帧和结尾帧；$s_{unvoiced}$ 是当前音节中清音帧数总和。

为每种无调音节的归一化清浊音转换时刻，使用 EM 算法建立 GMM 模型，服从分布：

$$p(x\mid\lambda)=\sum_{k=1}^{K}\omega_k N(x\mid\mu_k,\Sigma_k) \tag{2}$$

其中 ω_k 是第 k 个高斯分量的权重，$N(x\mid\mu_k,\Sigma_k)$ 为高斯密度函数：

$$N(x\mid\mu_k,\Sigma_k)=\frac{1}{\sqrt{2\pi\mid\Sigma_k\mid}}e^{-\frac{1}{2}(x-\mu_k)\Sigma_k^{-1}(x-\mu_k)} \tag{3}$$

对于 n 状态 HMM，在参数生成算法中，首先根据上下文预测得到状态时长划分，记音节起始时刻 t_0，各状态结束时刻依次为 $t_1,t_2,\ldots,t_n$，易知 t_n-t_0 即为预测的音节时长。

记候选清浊音转换时刻为：

$$t_{i_1},t_{i_2},\ldots,t_{i_m},(0\le i_1<i_2<\cdots<i_m\le n) \tag{4}$$

他们满足如下条件：

$$\begin{cases} p_{i_j}(voiced)<threshold_{voiced} \\ p_{i_j+1}(voiced)\ge threshold_{voiced} \end{cases},(1\le j\le m) \tag{5}$$

其中，$p_i(voiced)$ 是状态 i 为浊音的概率，且定义 $p_0(voiced)=0, p_{n+1}(voiced)=1$。

根据最大似然准则，找到

$$J=\arg\max_{1\le j\le m}p(\frac{t_{i_j}-t_0}{t_n-t_0}\mid\lambda) \tag{6}$$

则 t_{i_j} 即为当前音节的最优清浊音转换时刻。

图 3 给出一个基于 GMM 的清浊音转换模型指导音节清浊音预测的例子。在训练阶段，使用音节“hui”的清浊音转换时刻建立 4 高斯 GMM。图中给出了原始训练数据的统计直方图，4 个高斯分量的密度分布和 GMM 的密度分布。在原始方法中，各 HMM 状态（以虚线划分）独立进行清浊音预测，其中第 6 和第 10 个状态被预测为清音，预测基频成为一条间断的曲线，合成结果沙哑不清晰。本文提出的方法以此为基础，首先找出候选清浊音转换时刻，分别是第 4、6 状态的结束时刻，如 GMM 概率密度图中两条竖直虚线所示。根据最大似然准则，应用训练好的 GMM 可以得到，第 4 状态结束时刻是较优的清浊音转换时刻。根据此结果，可以得到连续的基频曲线和清晰自然的合成语音。

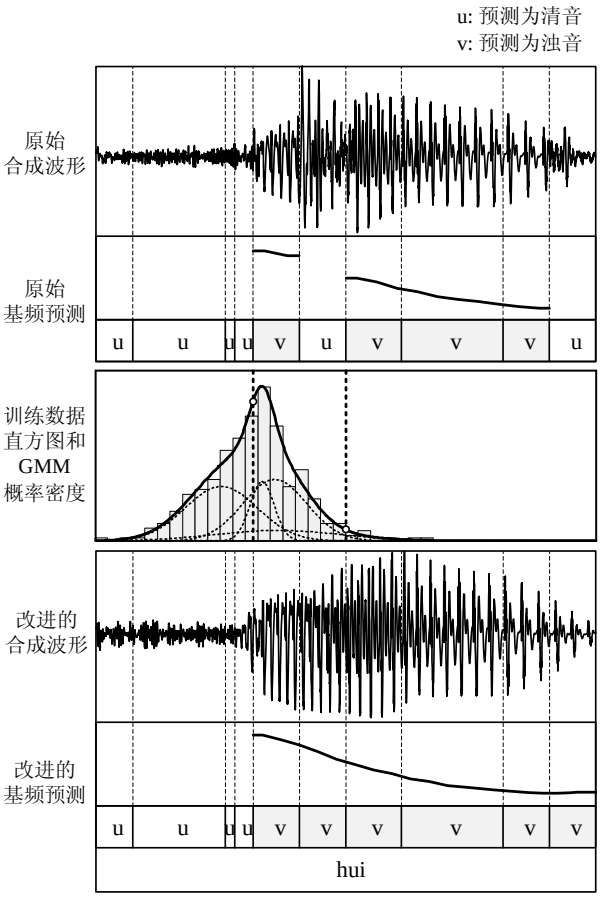

图 3　基于 GMM 的清浊音预测结果

3　实验评估

3.1　构建实验系统

实验采用 1000 句声学覆盖均衡的女声普通话语料作为训练数据，所有数据都经过手工标注。语音数据采样精度 16KHz，量化精度 16 位，非静音段总时长约 1.45 小时，共包含 19712 个音节，。

系统使用的语音参数共 78 维，包括 24 阶 LSF 参数，对数基频值，以及他们各自的一阶、二阶差分。提取语音参数时使用的帧长为 25ms，帧移 5ms。

训练阶段，系统以音节为基本发音单元，使用 10 状态 HMM 进行建模，并使用基于 MDL 准则的决策树聚类算法建立上下文相关的 HMM 模型。决策树聚类使用的问题集所包含的上下文信息见表 2。其中使用到的韵律单元包括音节、韵律词、韵律短语、语调短语和句子等。

合成阶段，首先分析待合成文本得到发音标注和上下文信息，通过决策树选择得到语音参数的概率密度函数序列。其中，谱参数序列可以直接从该概率密度函数序列计算生成，基频参数则先需要通过计算 GMM 似然度判断帧的清浊音，之后对浊音帧利用概率密度函数计算基频参数序列。最后由语音参数合成语音。

表 2　决策树问题集中使用的上下文信息

类别	语境信息
发音信息	当前音节拼音及声调
	当前音节声韵母及其类别
	前一音节拼音及声调
	后一音节拼音及声调
	前音声母及其类别
	后音声母及其类别
韵律信息	韵律单元的正序位置
	韵律单元的倒序位置
	当前韵律单元中子单元个数
	前一韵律单元中子单元个数
	后一韵律单元中子单元个数

3.2　主观评测

我们对本文提出的方法进行了主观评测。9 位有经验的评测者使用平均意见评分（MOS）方法，综合考虑自然度，清晰度和可懂度，对从测试文本中随机选取的 10 个句子的合成语音进行评分。

评测针对如下三种合成语音：1) 原始系统合成语音，2) 使用标注信息指导的基频提取算法的合成语音，3) 使用标注信息指导基频估计算法和基于 GMM 的清浊音预测算法的合成语音。

评测结果如图 4 所示。标注信息指导的基频提取算法将合成语音的 MOS 评分从 3.0±0.1 提升至 3.2±0.1，进一步，使用清浊音预测算法后，合成语音的 MOS 评分提高至 3.5±0.1。从中可以看出，本文提出的基频提取和预测方法较大幅度的提升了合成语音的质量。

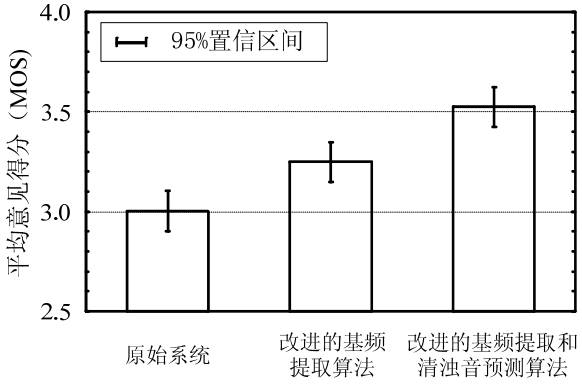

图 4　合成语音的 MOS 评分对比

4　总结

本文提出一种基频清浊音建模和预测优化算法，并应用于 HMM 语音合成。该方法针对参数化语音合成方法中，激励源选择对合成质量影响很大

的特点，对基频清浊音建模和预测进行优化。在建模阶段，通过使用语料标注信息指导基频提取，提高清浊音提取的正确性；在合成阶段，通过应用清浊音转换时刻的高斯混合模型，减少由于 HMM 状态输出概率分布间的独立性造成的清浊音预测错误。主观评测表明，该方法有效的改善了合成语音的音质和韵律。

5 致谢

本文的部分工作是在 IBM 中国研究中心大学联合研究计划的支持下完成的。非常感谢语音组各位老师提供的建议和帮助。

参 考 文 献

[1] Tokuda K, Zen H, Black A W. An HMM-based speech synthesis system applied to English [A]. Proc. of IEEE Workshop on Speech Synthesis [C]. 2002.

[2] LING Zhenhua, WU Yijian, WANG Yuping, et al. USTC system for Blizzard Challenge 2006 an improved HMM-based speech synthesis method [A]. Proc. of ICSLP Satellite Workshop, Blizzard Challenge [C]. 2006.

[3] Yoshimura T, Tokuda K, Masuko T, et al. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis [A]. Proc. of Eurospeech [C]. 1999, vol. 5, 2347-2350.

[4] Tokuda K, Mausko T, Miyazaki N, et al. Multi-space probability distribution HMM [J], IEICE Trans. Inf. & Syst., 2002, vol.E85-D, no.3: 455-464.

[5] Tokuda, K, Yoshimura, T, Masuko T, et al. Speech parameter generation algorithms for HMM-based speech synthesis [A]. Proc. of ICASSP [C], 2000, vol. 3, 1315-1318.

[6] Dines J, Sridharan S. Trainable speech synthesis with trended hidden Markov models [A]. Proc. of ICASSP [C], 2001, 833–837.

[7] Zen H, Tokuda K, Kitamura T. An introduction of trajectory model into HMM-based speech synthesis [A], Proc. of 5th ISCA Speech Synthesis Workshop [C]. 2004.

[8] Toda T, Tokuda K. Speech parameter generation algorithm considering global variance for HMM-based speech synthesis [A]. Proc. of Eurospeech [C]. 2001, 2801–2804.

[9] Masuko T, Tokuda K, Kobayashi T, et al. Speech synthesis from HMMs using dynamic features [A]. Proc. of ICASSP [C]. 1996, 389-392.

A research of F0 extraction and prediction algorithm for HMM-based speech synthesis

KANG Shiyin1, DUAN Quansheng1, SHUANG Zhiwei2, QIN Yong2, CAI Lianhong1

(1. Department of Computer Science & Technology, Tsinghua University, 100084; 2. IBM China Research Lab, Beijing, 100094)

Abstract: This paper proposes a F0 extraction and prediction algorithm for HMM-based speech synthesis, which is optimized for unvoiced/voiced determining. In the conventional method, the unvoiced/voiced determining of the F0 parameters, which directly decides the source type of the excitation in Vocoder Synthesis, has a high error rate, and this leads to a poor quality of the synthesized speech. The novel method, in which the information of the text labels is used to help extracting the F0 and the normalized time of the unvoiced-to-voice switch of the syllables is modeled by GMMs for the F0 prediction, improves the quality of the unvoiced/voiced determining. The result of a perceptual evaluation demonstrates that the proposed method improves the quality of the synthesized speech obviously, and the MOS rise from 3.0 to 3.5 compared with the conventional method.

Key words: speech synthesis; HMM; F0 modeling; GMM