一种适合HMM汉语语音合成的建模单元挑选算法*

段全盛[1]，康世胤[1]，双志伟[2]，吴志勇[1]，蔡莲红[1]，秦勇[2]
(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：本文比较把不同声学单位作为建模单元时 HMM 汉语语音合成引擎的合成音质，分析建模单元对 HMM 语音合成的影响，并提出一种可变建模单元的 HMM 语音合成方法。考虑语料库的音段切分和 HMM 建模特点，汉语可以选用音节和声韵母两种单元进行 HMM 建模。本文分析音节和声韵母做建模单元的优缺点，通过比较实验验证了建模单元长度及相同模式分类下样本数目对 HMM 语音合成效果有重要的影响。最后本文提出基于样本数的建模单元挑选方法，并使用不同的建模单元进行 HMM 语音合成。主观评测实验表明本文提出的改进方法有效地提高了合成音质。

关键词：语音合成；隐马尔可夫模型；建模单元；音节；声韵母
中图分类号：TN912.3

近年来，基于隐马尔可夫模型（HMM）的语音合成技术[1]因为其架构成本低、适用于嵌入式系统、与语种相关性小等特点受到了广泛的重视。

传统的 HMM 语音合成方法分为训练和合成两个阶段。在训练阶段，提取训练语料的声学参数进行 HMM 建模[2]，并对所得模型进行决策树聚类 [3]。在合成阶段，首先经过文本分析得到待合成文本的语境信息，并使用这些语境信息在训练所得决策树中预测模型状态序列，最后解码 HMM 模型生成参数[4]并通过参数合成器进行语音合成。

汉语（普通话）以音节为发音单元，音节由声韵母相拼地构成。而 HMM 语音合成中建模单元的选择需要综合考虑多方面因素：建模单元切分的准确性、建模单元使用 HMM 模型描述的精度以及对 HMM 模型聚类时样本数目是否合适等。目前基于 HMM 的汉语语音合成大多以声韵母为建模单位[5]。

本文对使用音节和声韵母作为建模单元的 HMM 汉语语音合成系统进行了比较，分析了建模单元的长度及样本数目对 HMM 语音合成效果的影响。通过比较实验验证了音节和声韵母作为建模单元时各自的优缺点，并在此基础上提出了可变建模单元的 HMM 语音合成方法，通过分析待合成语段对应训练集中样本数进行挑选建模单元的种类，提高了合成音质。

文本内容组织如下：第一部分介绍 HMM 汉语语音合成系统结构；第二部分比较使用音节或声韵母作为建模单元对 HMM 语音合成的影响，提出可变建模单元的 HMM 语音合成方法；第三部分阐述基于样本数的建模单元挑选方法；第四部分给出一个合成系统的实例，并进行实验评测；第五部分进行总结。

1. 语音合成系统结构

本文构建的合成系统包括训练及合成两个模块（如图 1 所示）。在训练阶段，首先根据切分信息提取建模单元的基频和频谱参数，之后在标注信息的指导下训练生成 HMM 模型，并用决策树对模型进行聚类。为了准备不同建模单元的训练结果，本文分别进行音节和声韵母为建模单元的 HMM 训练。同时，统计标注信息中各种语境信息的取值概率以及训练集内所有样本联合出现概率分布曲线。

在合成阶段，通过文本分析预测待合成语句的语境信息，根据语境信息进行建模单元选择：计算待合成语境信息的联合输出概率,利用该概率在训练集内样本联合出现概率分布曲线中进行比较。若该概率在训练集中相对较小，该标注语段对应样本数较少，采用声韵母作为建模单元进行合成；若该概率在训练集中相对较大，认为该标注对应的语段对应样本数较多，采用音节为建模单元进行 HMM 语音合成。

选择好建模单元后，在相应的训练结果中进行决策树查找，预测 HMM 模型序列。最后应用语音参数生成算法生成语音参数序列，通过参数合成器合成高质量的目标语音。

*基金项目：国家自然科学基金（60805008, 60433030）；国家 863 课题（2007AA01Z198）；国家 973 课题（2006CB303101）；教育部博士点基金（200800031015）
作者简介：段全盛（1985-），男（汉族），北京，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

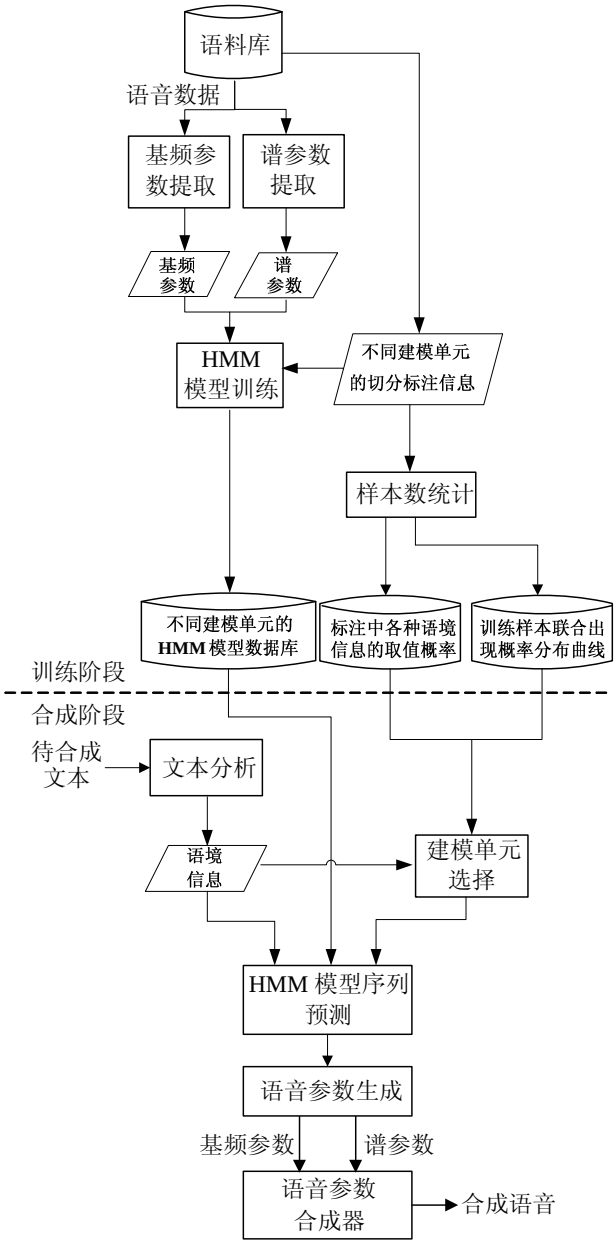

图 1　合成系统框架

2. 可变建模单元的HMM语音合成

2.1 不同建模单元对HMM语音合成的影响

汉语以音节为发音单元，而汉语的音节总是由声母加上韵母或者只由韵母构成。因此在汉语语料库中，音节和声韵母为单位的切分信息较为准确[6]，一般使用这两个声学单位作为建模单元进行 HMM 汉语语音合成。

HMM 语音合成的训练阶段包括 HMM 建模和模型聚类两个重要模块。建模单元对 HMM 语音合成的影响也主要体现在这两个模块之中。

HMM 建模一般采用自左向右的状态序列来描述建模单位[7][8]。在这个状态序列中，状态数固定，每个状态都包含两种概率模型：转移到下一个状态的转移概率和该状态的输出概率。即 HMM 模型采用若干个状态表示整个建模单元，每个状态根据输出概率选择该状态对应的语音参数。因此在状态数固定的前提下，长度较短的建模单元更易于 HMM 模型对其进行精确地刻画。在汉语语音合成中，以声韵母作为建模单元时，每个 HMM 模型输出的参数也更贴近于自然语音参数。

但与此同时，HMM 模型只包含建模单元内部信息，并不包含不同建模单元之间的信息。虽然 HMM 语音合成系统使用了上下文相关的语境信息和动态特性来描述建模单元之间的关系。但总体来讲，建模单元之间的语音特征描述并不像建模单元内部一样有效。针对汉语同音节内部的声韵母关系密切的特点，音节为建模单元的语音合成更能准确地刻画音节内部的韵律结构，提高整个合成语句的自然度和连贯性。

HMM 语音合成中，对训练得到的 HMM 模型使用基于 MDL 准则的决策树聚类[9]。在聚类过程中，不断选取使决策树熵减最多的问题和待分裂节点进行分裂，直到满足终止条件结束节点分裂。因此，聚类结束后得到的决策树中，叶节点有以下特点：或者节点内部样本距离足够小，即相似程度很高；或者节点内部样本数目较少；或者节点内样本分布分散但没有合适的问题可以进行熵减足够大的分裂。提高节点内部相似单元的样本数目有助于改善决策树聚类效果，进而提高预测 HMM 模型的准确性。在使用相同汉语语料库的前提下，以声韵母为单位进行切分显然可以得到更多的同类型单元，有助于提高预测 HMM 模型的准确性，改善合成音质。

综上所述，以音节为建模单位的 HMM 语音合成更适合刻画音节和句子的韵律结构，帮助提高合成语句的连贯性。而以声韵母为建模单位时，HMM 模型可以更加细致地描述语音参数，并提供更大样本集供聚类所用（见表 1）。

表 1 建模单元在 HMM 合成中优缺点

建模单元	优点	缺点
音节	韵律自然，语音连贯	单元长度长，样本数较少
声韵母	建模精确，样本集大	音节及整句的韵律不够连贯

2.2 可变建模单元的HMM语音合成方法

在训练语料库较小的情况下，改变建模单元长度对 HMM 语音合成音质造成的影响十分明显。以声韵母为建模单位可以使得合成语音的清晰度更高，而以音节为建模单位可以使合成语音有更好的自然度。

本文架构了 HMM 汉语语音合成系统（系统设

计见 4.1)，在训练集为 1000 句话的情况下分别使用音节和声韵母作建模单元进行语音合成，得到两种建模单元对应的合成语音。

本文设计比较实验从清晰度和自然度两个方面对两种合成语音进行比较。实验随机选取了训练集内未出现的 8 句话，分别使用两种不同建模单元进行语音合成，并由合成语音以随机顺序组成 8 组待评测数据。请 10 名评测者在每一组语音数据中分别从清晰度和自然度两个方面进行评测，并在每组语音数据中选择清晰度高的合成语音及自然度高的合成语音。

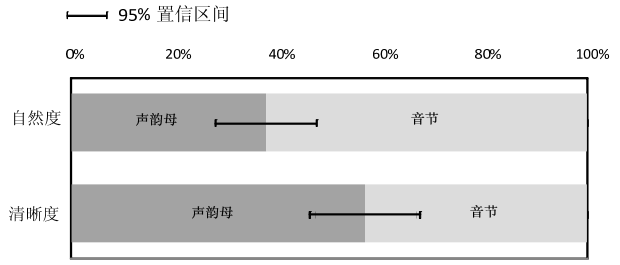

图 2 使用不同建模单元合成语音的自然度清晰度比较

实验结果如图 2 所示，1000 句训练集下以声韵母为建模单元的语音合成有较高的清晰度，而以音节为建模单元的语音合成结果有较高的自然度。

实验说明在小训练集下，声韵母为建模单元可以对声学单元进行更精确地建模和聚类，提高合成语音的清晰度，但因为建模单元较短整体自然度欠佳；音节为建模单元益于描述整个语句的韵律结构，提高合成语音的自然度，但因为样本数较少合成音段的清晰度较低。为了结合 2 种建模单元的优点，使合成语音即清晰又连贯，本文提出一种可变建模单元的 HMM 语音合成方法（见图 3 所示）。

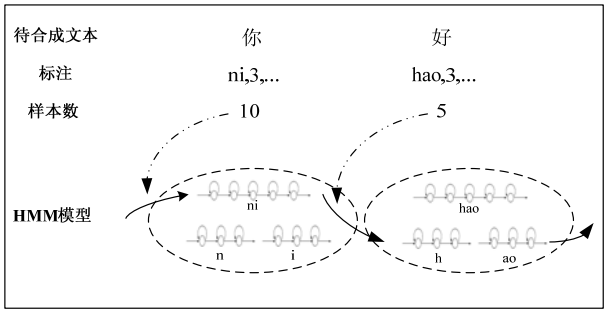

图 3 可变建模单元的 HMM 语音合成方法

本方法使用文本分析预测待合成语段的语境信息，并利用语境信息使用基于样本数的建模单元挑选算法（详见 3.1）判断应采用的建模单元类型。对于训练样本数较少的待合成语段使用短建模单元（声韵母）进行 HMM 建模，从而提高合成语段的清晰度；而对于样本数足够的待合成语段采用较长建模单元（音节）进行 HMM 建模，在保持较高清晰度的前提下提高整体合成语句的自然度。

3. 基于样本数的建模单元挑选方法

3.1 方法说明

训练阶段，本文实现的 HMM 汉语语音合成系统利用多种语境信息描述声学单元的上下文信息。系统利用切分信息提取所有训练语句中包含的声学单元作为训练样本，并同时分析每个声学单元的标注信息 T。若 T 由 m 种语境信息组成，则有 $T=\{T_1,T_2,\cdots,T_k\cdots,T_m\}$。

本方法在训练阶段统计所有样本的标注信息，并为标注中每种语境信息 k 统计其取值 T_k 为不同境况的概率：

$$P_{T_k=i}=\frac{n_{T_k=i}}{N} \tag{1}$$

其中，$n_{T_k=i}$ 是满足 k 类别语境信息的取值 T_k 等于 i 的训练集内样本数目； N 是训练集内样本数总和。显然， $P_{T_k=i}$ 满足：

$$\sum_{\forall i}P_{T_k=i}=1 \tag{2}$$

同时，对训练集中所有样本 t 计算其标注信息联合出现概率：

$$P_t=P_{T_1=t_1}\times P_{T_2=t_2}\times\cdots\times P_{T_k=t_k}\times\cdots\times P_{T_m=t_m} \tag{3}$$

实验对训练集内所有样本的标注信息联合出现概率进行统计，并根据该概率大小进行排序。统计联合出现概率 P_t 及联合出现概率小于 P_t 的样本数占总样本数的比例，得到训练集内所有样本的联合出现概率分布曲线(如图 4 所示)。

合成阶段，预测待合成语句中声学单元的标注信息 $t'=\{t'_1,t'_2,\cdots,t'_m\}$，利用训练阶段统计的各语境信息取值概率 $P_{T_k=i}$ 计算标注 t' 的联合出现概率：

$$P_{t'}=P_{T_1=t'_1}\times P_{T_2=t'_2}\times\cdots\times P_{T_k=t'_k}\times\cdots\times P_{T_m=t'_m} \tag{4}$$

利用 $P_{t'}$ 在训练集内样本联合出现概率分布曲线中进行比较，若集内出现概率小于 $P_{t'}$ 的样本比例小于某阈值 C，认为标注为 t' 的语段对应样本数较少，采用较短建模单元（声韵母）进行合成；若集内出现概率小于 $P_{t'}$ 的样本比例大于等于某阈值 C，认为标注为 t' 的语段对应样本数较多，采用较长建模单元（音节）进行 HMM 语音合成。

3.2 阈值C的设定

本文以采用 1000 句女生普通话语料作为训练数据为例，讲解阈值 C 的设定过程。首先以音节为建模单元进行 HMM 训练，训练过程中统计音节级标注包含的各语境信息取值概率 $P_{T_k=i}$ 和训练集内所有样本的联合出现概率分布曲线（见图 4 所示）。

其中曲线中点的横坐标为联合出现概率的 log 值，纵坐标为联合出现概率小于该点对应概率的样本占样本集的比重。

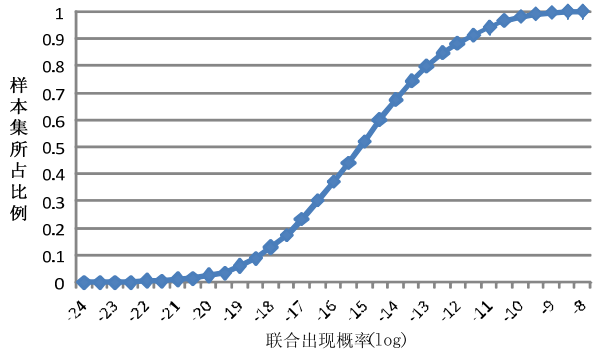

图 4 以 1000 句话为训练集时的联合概率分布曲线

之后对以音节为建模单元的 HMM 汉语语音合成系统的合成结果进行评测，挑选其中合成音质较差的音节，提取这些音节的标注信息。利用训练阶段统计的各语境信息取值概率 $P_{T_k=i}$ 计算所提取标注信息的联合出现概率。对这些音节对应标注的联合出现概率排序并设定差音处理率 r，按出现概率从高到低的顺序在音质较差的音节集 A 中按比例 r 选择待处理音节集 A_r。本实验分别设定 r 值为 90%、85%和 80%，在错音集 A 中选择对应比例的待处理音节集合 A_r，记录 A_r 内部最高标注联合出现概率 p_{A_r}。最后利用 p_{A_r} 在训练得到的联合出现概率分布曲线中查找标注联合出现概率小于 p_{A_r} 的训练集内样本数目（见表 2）。

表 2 音质较差语段联合出现概率分析表

差音处理率 r	待处理差音集最高联合出现概率 p_{A_r} (log)	联合出现概率小于 p_{A_r} 的集内样本覆盖率
90%	-13.8	70.4%
85%	-13.9	69.0%
80%	-14.3	63.6%

由实验数据可知，训练集大小是 1000 句话时，在较差音质的合成语段中，有 90%语段对应的标注联合出现概率 p_{A_r} 较低：在训练集内有只有 70.4%样本的标注联合分布概率低于 p_{A_r}。即如果我们想针对 90%合成音质较差的语段进行处理，至少需要规定阈值 C 为 70.4%。

4. 实验评估

4.1 架构实验系统

本文分别使用音节和声韵母作为建模单元进行 HMM 语音合成，并设计实验对不同方法下 HMM 语音合成的音质进行比较，验证不同建模单元对合成结果的影响。

实验采用声学覆盖均衡的女声普通话语料作为训练数据。为了比较语料库大小对合成结果的影响，分别从语料库中提取 1000 句、2000 句和 3000 句话作为训练集。实验采用手工切分的方式提取音节数据，采取自动切分方法提取声韵母边界点。语音数据采样精度 16KHz，量化精度 16 位。

系统以帧长为 25ms，帧移 5ms 为单位提取语音参数。语音参数共 78 维，包括 24 阶 MGC 参数，对数基频值，以及它们各自的一阶、二阶差分。

训练阶段，系统对音节和声韵母分别使用 10 状态和 5 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。由于不同的建模单元对应不同的标注信息，相应地需要不同的决策树问题集来描述。实验针对不同建模单元分别设计了决策树聚类所用问题集中包含的上下文信息。

本文采用发音特征和韵律特征来描述音节的上下文信息。具体特征信息的设计见表 3。其中声韵母类别包括声母发音方法、声母发音部位、韵头发音方法、韵尾发音方法等。韵律特征描述句子韵律结构，设计韵律单元有：音节、韵律词、韵律短语、语调短语和句子等。

声韵母为建模单元时，决策树问题集涉及到的上下文信息包括音节为单位的所有信息，并在其基础上进行了扩充：在韵律特征方面添加了声韵母级别韵律单元的位置信息和子单元数目。

表 3 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
	前音/后音拼音及其声调
	当前音节声韵母及其类别
	前音/后音声韵母及其类别
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

合成阶段，首先分析待合成文本得到音节和声韵母级别的发音标注和上下文信息，分别通过 2 种建模单元的决策树选择各自的 HMM 模型，得到语音参数的概率密度函数序列。最后计算基频参数和谱参数序列，由语音参数合成语音。

4.2 音节和声韵母作韵律单元比较实验

本文请 13 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 5 个句子的合成语音进

行评分。

评测针对如下七种语音：语料库中原始语音，音节作为建模单元使用 1000/2000/3000 句话作为训练集的合成语音，声韵母作为建模单元使用 1000/3000/5000 句话作为训练集的合成语音。

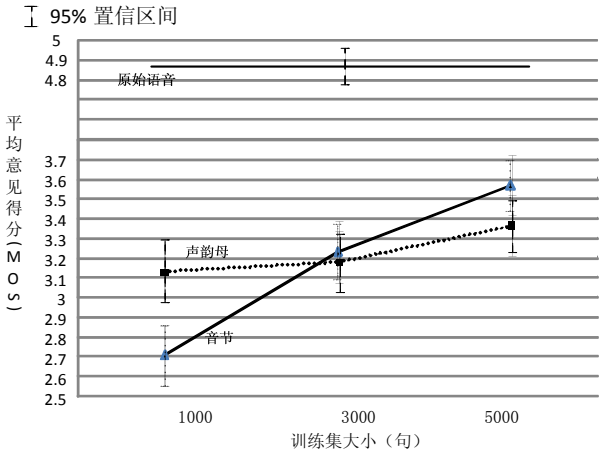

图 5 不同建模单元随训练集规模变化 MOS 评分对比

评测结果如图 5 所示，原始语音 MOS 评分为 4.87 ±0.10；以音节为建模单元时，在 1000/3000/5000 句话训练集下合成语音的 MOS 评分分别为 2.71±0.19/ 3.23±0.15/ 3.57±0.17；以声韵母为建模单元时，在 1000/3000/5000 句话训练集下合成语音的 MOS 评分分别为 3.14±0.20/ 3.18±0.19/ 3.36±0.15。

从图 5 中可以看出，当训练集增大时，两种建模单元下的语音合成系统 MOS 评分都随之提高。而在训练集较小时，以音节为建模单元的合成音质 MOS 评分低于声韵母为单元时的 MOS 评分。但随着样本集的增大，音节为单元时的 MOS 评分增长速度快于声韵母为单元时的评分。最后，在大样本集训练的情况下，音节为建模单元的合成系统得到了更高的 MOS 评分。

实验结果说明建模单元长度对 HMM 语音合成音质有很大的影响。以声韵母为建模单位的 HMM 语音合成在小样本集下可以对声学单元进行更精确地建模和聚类，让合成语音表现出更好的音质。而音节为单元时，合成语音有更自然韵律结构，在样本集大的情况下听起来更连贯。

4.3 可变建模单元HMM语音合成方法评测实验

本实验采用 1000 句女声普通话作为训练语料。请 13 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 5 个句子的合成语音进行评分。

评测针对如下四种语音：语料库中原始语音，音节作为建模单元合成语音，声韵母作为建模单元合成语音，基于样本数挑选建模单元的合成语音。其中基于样本数挑选建模单元方法的差音语段处理率设为 90%，其对应的比重判断阈值 C 为 70.4%。

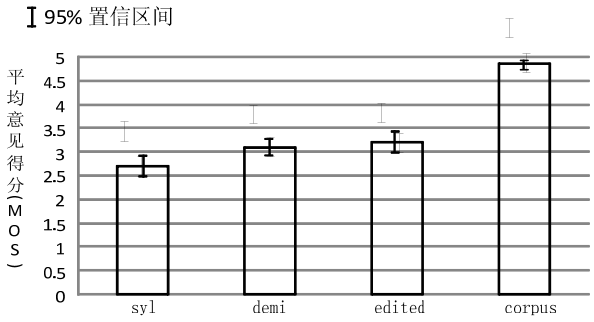

图 6 基于样本数的建模单元挑选实验 MOS 评分对比

评测结果如图 6 所示。原始语音 MOS 评分为 4.87±0.10；以音节为建模单元时，MOS 评分为 2.71±0.19；以声韵母为建模单元时，MOS 评分为 3.14±0.20；使用基于样本数的建模单元挑选方法后，MOS 评分提高至 3.20±0.24。统计发现，5 个评测语句中有 62%的发音单元使用声韵母作为建模单元 38%的发音单元使用音节作为建模单元。

5. 总结

本文对训练过程中 HMM 建模和模型聚类等模块算法进行研究，从多角度分析不同建模单元对 HMM 语音合成的影响。针对 HMM 汉语语音合成，本文分析了音节和声韵母做建模单元的优缺点，并通过比较实验验证了建模单元长度及样本集的规模对 HMM 语音合成效果有重要的影响。最后本文提出基于样本数的建模单元挑选方法，为每一节待合成语段挑选最合适的建模单元，并使用挑选出的建模单元进行 HMM 语音合成。主观评测实验表明本文提出的改进方法有效地提高了合成音质。

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

[2] K. Tokuda, T. Mausko, N. Miyazaki, et al. Hidden Markov models based on multi-space probability for pitch pattern modeling, Proc. of ICASSP, pp. 229-232, Arizona, 1999.

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

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

[5] 吴义坚，王仁华，基于HMM的可训练中文语音合成，中文信息学报，

2005.
[6] H. Kawai, and T. Toda, An Evaluation of Automatic Phone Segmentation for Concatenative Speech Synthesis, Proc. IEEE ICASSP2004, 2004.
[7] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, The HTK Book. Entropic Ltd., Cambridge, 1999.
[8] T. Hain, Hidden Model Sequence Models for Automatic Speech Recognition, PhD dissertation, University of Cambridge, 2001.
[9] K. Shinoda and T.Watanabe, “MDL-based context-dependent subword modeling for speech recognition,” J. Acoust. Soc. Jpn.(E), vol. 21, pp. 79–86, 2000.

A modeling-unit selection algorithm approach to HMM-based speech synthesis on Chinese

DUAN Quansheng1, KANG Shiyin1, SHUANG Zhiwei2, WU Zhiyong1, QIN Yong2, CAI Lianhong1
(1. Department of Computer Science & Technology, Tsinghua University, 100084; 2. IBM China Research Lab, Beijing, 100094;)

Abstract: HMM-based speech synthesis system achieves diverse quality when using different modeling-units. This paper analyzes the effects when using various kinds of modeling-units, such as syllable, initial and final. Therefore a modeling-unit selection algorithm is proposed. Given the accurate segmentation of corpus in Chinese language and the modeling features of Hidden Markov Model, syllable and initial/final are often used in HMM-based speech synthesis on Chinese. This paper examines the advantages and disadvantages when using syllable or initial/final as modeling-unit, verifies that the length and amount of modeling-units relate to the synthesized sound quality. At last, a modeling-unit selection algorithm is put forward for HMM-based speech synthesis on Chinese. The result of a perceptual evaluation demonstrates that the proposed method can significantly improve the naturalness of synthesized speech.

Key words: speech synthesis; HMM; modeling-unit; syllable; initial; final