汉语连续语音识别中不同基元声学模型的复合

张 辉 杜利民
(中国科学院声学研究所语音交互技术实验室 北京 100080)

摘 要 该文研究由不同声学基元训练的声学模型的复合。在汉语连续语音识别中，流行的基元包括上下文相关的声韵母基元和音素基元。实验发现，有些汉语音节在声韵母模型下有更高的识别率，有些音节在音素模型下有更高的识别率。该文提出一种复合这两种声学模型的方法，一方面在识别过程中同时使用两种模型，另一方面在识别过程中避开造成低识别率的模型。实验表明，采用本文的方法后，音节错误率比音素模型和声韵母模型分别下降了 9.60%和 6.10%。

关键词 语音识别，声学模型复合，声学模型选择，错误率

中图分类号：TP391.42 **文献标识码**：A **文章编号**：1009-5896(2006)11-2045-05

Combination of Acoustic Models Trained from Different Unit Sets for Chinese Continuous Speech Recognition

Zhang Hui Du Li-min
(*SITR, Institute of Acoustics, Chinese Academy of Sciences, Beijing* 100080, *China*)

Abstract Combination of acoustic models trained from different unit sets is studied in this paper. For Chinese continuous speech recognition, Prevailing unit sets include context-dependent initial-final unit set and context-dependent phone unit set. Through experiments it is discovered that some Chinese syllables have higher recognition rates under initial-final model while some have higher recognition rates under phone model. In this paper, a method is proposed to combine these two acoustic models. On one hand the two acoustic models can be fully utilized during the recognition process; on the other hand, some models that lead to low recognition rate will not be used. Experiments show that in comparison with initial-final model and phone model, syllable error rate is reduced by 9.60% and 6.10% respectively after using the provided method.

Key words Speech recognition, Acoustic model combination, Acoustic model selection, Error rate

1 引言

近几年来，许多研究者利用多种声学模型或者多种特征矢量相互辅助的方法来提高连续语音识别系统的识别率。Fiscus提出的“ROVER”方法[1]，利用多个识别器的输出构成候选复合网络，再对该候选复合网络进行后处理以提高识别率；这里的一个问题是，在识别过程中没有动态利用各种声学模型的信息。Yan等人提出在识别过程中利用多种特征矢量交叉参考(Cross-reference)，以减少剪枝错误[2]。Nifian等人提出了耦合HMM[3]，在传统语音特征的基础上，加入视觉特征构成耦合HMM。文献[2]和文献[3]在不同角度都在识别过程中动态利用了多种特征矢量，但没有涉及到如何利用多种声学基元。

本文提出一种新的复合方法，在识别过程中充分利用有相同特征矢量(MFCC)但声学基元不同的声学模型。

在汉语连续语音识别中，较为流行的声学基元有两种：上下文相关音素基元[4]和上下文相关声韵母基元[5]。并且，从总体上说，声韵母模型比音素模型有更好的识别性能[5]。但是，这只表明对所有汉语音节的平均效果而言，声韵母模型比音素模型的描述能力更强；而对于具体的音节，情况并不是这样。

本文采用HTK定义的准确率(Acc%)[6]来衡量识别性能。准确率定义如下：

$$\mathrm{Acc}\% = \frac{N - D - S - I}{N} \times 100\% \tag{1}$$

在式(1)中，N，D，S 和 I 分别代表音节出现次数，被删除次数，被替代次数和被插入次数。

表 1 是采用全音节语法网络(每一个音节都可以接任意一个音节，困惑度为 406)时，基线模型识别“863 数据库”训练集中 72 个说话人的一些音节的识别结果。从表 1 可以看到，某些音节如“BEI”和“ZHAO”，两种模型的准确率相当；某些音节如“SHUI”和“BIN”，声韵母模型的准确率远好于音素模型；对于某些音节如“QIONG”和“WEI”，音素模型的准确率远好于声韵母模型。因此，让识别器以一定的方式复合使用两种模型，应该可以提高准确率。

同时还可以发现，用音素模型进行识别时，某些音节如

2005-03-08 收到, 2005-08-15 改回

表 1 一些音节的准确率

Tab.1 Acc% of some syllables

音节名称	组成音节的声韵母基元	组成音节的音素基元	声韵母模型下准确率(Acc%)	音素模型下准确率(Acc%)
A	{_a, a}	{A_A}	–5.33	–148.04
BEI	{b, ei}	{p, e, I}	91.89	90.99
BIN	{b, in}	{p, j, i, n}	72.35	53.25
QIONG	{q, iong}	{ts/_h, j, O, ng}	–22.22	82.22
SHUI	{sh, ui}	{s`, w, e, I}	81.28	70.58
WEI	{_u, uei}	{w, e, I}	61.74	78.87
YUE	{_v, ve}	{H, e_o}	44.82	67.55
ZHAO	{zh, ao}	{ts`, A, U}	81.81	80.21

“A”准确率非常低(准确率为负值是因为插入错误很大)，而在声韵母模型下该音节的准确率要高得多；某些音节如“YUE”，在声韵母模型情况下的准确率较低，而在音素模型中则高得多。因此，如果在识别时，音节“A”只使用声韵母模型，而音节“YUE”只使用音素模型，必然可以提高准确率。

基于上面的分析，本文提出了一种声学模型复合方法。在全音节语法网络利用不同模型分别进行网络扩展(每一个音节都扩展为一个以上 HMM 的串联)后，按照一定规则建立属于不同模型的网络结点之间的连接。另外，本文根据训练数据识别后的统计结果，设计了一个有效的声学模型选择算法，在构造复合的识别网络时，有些音节不使用声韵母模型，有些音节则不使用音素模型。

本文的组织顺序如下：第 2 节介绍如何进行两种声学模型的复合，第 3 节介绍声学模型选择算法，第 4 节给出实验结果以及和其他方法的对比，第 5 节是总结和展望。

2 声学模型复合方法

2.1 匹配结点

基于上下文相关模型的汉语连续语音识别系统中，一个汉语音节在网络扩展时扩展成多个结点的串联。比如采用上下文相关音素模型时，由{ts/、j、ae 和 N}四个音素组成的“JIAN”音节可以扩展成图 1 所示的形式。

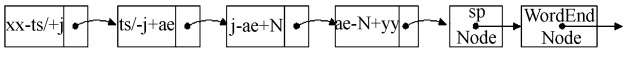

图 1 JIAN 的上下文相关音素模型表示

Fig.1 Context-dependent phone-model for JIAN

其中，“xx”和“yy”是当前音节的上文和下文，起始部分的“xx-ts/+j”结点是音节起始结点，spNode 是短时停顿结点，WordEndNode 是字尾结点；图 1 中所有的 6 个结点都隶属于音节“JIAN”。

在声学模型复合方法中，“匹配结点”是一个重要概念。同一音节在不同声学模型下扩展得到的两个起始结点，如果它们的上文单元(如图 1 中的“xx”)相互对应，那么称这两个结点是匹配结点。

这里以例子解释单元相互对应的概念：以“BAI”音节为例，在声韵母基元下该音节由{b, ai}组成，在音素基元下该音节由{p, a, I}组成。此时，单元“b”和“p”相互对应，单元“ai”和“I”相互对应。

图 2 是音节“BAI”和“JIAN”连接时，连接处的结点示意图。图 2 中“ai-j+ian”和“I-ts/+j”就是匹配结点，因为它们是同一音节“JIAN”音节用不同的声学模型(前者基于声韵母模型，后者基于音素模型)扩展得到的起始结点，并且它们有相互对应的上文单元“ai”和“I”。(图 2 中粗横线的上部基于声韵母模型，下半部分基于音素模型，下同。)

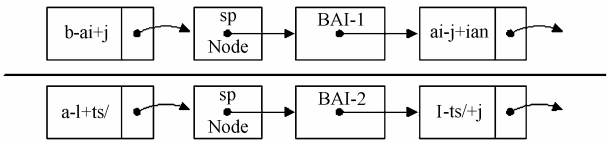

图 2 音节“BAI”和“JIAN”连接处的结点示意图

Fig.2 Link of BAI and JIAN

2.2 声学模型复合的原理

全音节语法网络用声韵母模型和音素模型分别进行独立的网络展开，每一个音节就对应两套串联的网络结点，一套属于声韵母模型，另一套属于音素模型。然后，就进行声学模型的复合：一个字尾结点除了原有的到下一音节起始结点的连接(例如图 2 中“BAI-1”和“ai-j+ian”的连接)外，还连接该起始结点的匹配结点(例如图 2 “BAI-1”还连接“I-ts/+j”)。这样，在识别过程中，当一个传递记号(Token)[7]从某一字尾结点(例如图 2 上半部分中的“BAI-1”结点)发出时，它除了进入原有的下一音节的起始结点(例如图 2 中的“ai-j+ian”)外，还要进入和该起始结点相匹配的结点(例如图 2 中的“I-ts/+j”)。同样地，基于另一个声学模型的字尾结点也会有类似的表现(例如从图 2 下半部分的“BAI-2”发出的传递记号会进 入“ai-j+ian”和“I-ts/+j”结点)。复合后的结点连接关系如图 3 所示。图 3 中，实线为声学模型复合前的结点连接，虚线为复合连接。本文后面图中的实线和虚线意义与此相同。

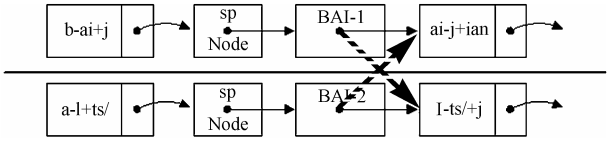

图 3　音节"BAI"和"JIAN"连接处的复合示意

Fig.3 Compound link of BAI and JIAN

上面描述的过程就是声学模型的复合过程。基于动态规划原理的识别算法在搜索网络时，会得到具有最高全局似然分值的最优路径；而该路径包含的所有结点中，一般情况下有些属于声韵母模型，有些属于音素模型。这就相当于在两种声学模型中进行了一定的优化。

通常情况下，全音节语法网络中的音节和音节之间并不直接连接，而是通过过渡音节(也称为 NULL 音节)相连，以减少语法网络的连接数。这种情况下声学模型复合后的结点连接关系如图 4 所示。图 4 中的“ai_nul_j”和“I_nul_ts/”称为过渡结点，这些过渡结点对应语法网络中的同一个过渡音节，它们的不同是因为上下文不同。这种情况下的复合原理和图 3 描述的复合原理是相同的。如无特殊说明，本文后面的原理描述都不考虑过渡结点。

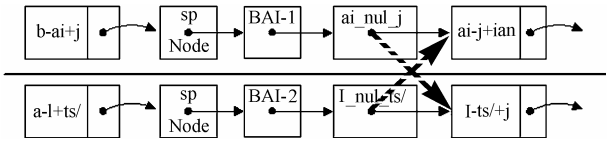

图 4　通过过渡结点建立的复合连接

Fig.4 Compound link built by transitional nodes

3　声学模型选择算法

3.1　音节误识率分析

从表 1 可以看到，音节“A”在两种模型下的准确率都非常低，并且在音素模型下的准确率远远低于声韵母模型下的准确率；音节“YUE”在声韵母模型下的准确率较低，并且远低于音素模型下的准确率。因此，在声学模型复合时，如果音节“A”只考虑声韵母模型，而音节“YUE”只考虑音素模型，那么总体的插入错误必然低于只使用单一声学模型时的插入错误。这就是声学模型复合过程中的选择思想。

下面是音节“BAI”连接“A”和“YUE”时，按照上述思想建立复合连接时的结点连接关系如图 5 所示。图 5 中，基于声韵母模型的“BAI-1”和基于音素模型的“BAI-3” 发出的传递记号都只进入基于声韵母模型的“ai-_a+a”结点，而不进入基于音素模型的“I-A_A+x”结点；类似地，基于声韵母模型的“BAI-2”和基于音素模型的“BAI-4”发出的传递记号都只进入基于音素模型的“I-H+e_o”结点，而不进入基于声韵母模型的“ai-_v+ve”结点。

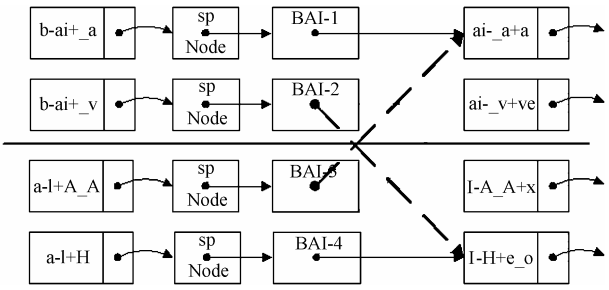

图 5　音节“BAI”连接“A”和“YUE”的复合

Fig.5 BAI links to A and YUE

但并不是每一个音节都只选择使用一种声学模型。从表 1 可以看到，“BEI”和“ZHAO”音节在两种模型下，不论正确率和准确率都相差很小。因此，这样的音节同时使用两种模型。下面是音节“BAI”连接“A”和“ZHAO”时，建立复合连接并应用选择算法后的结点连接关系如图 6 所示。图中，音节“A”只使用声韵母模型，而音节“ZHAO”同时使用两种声学模型。

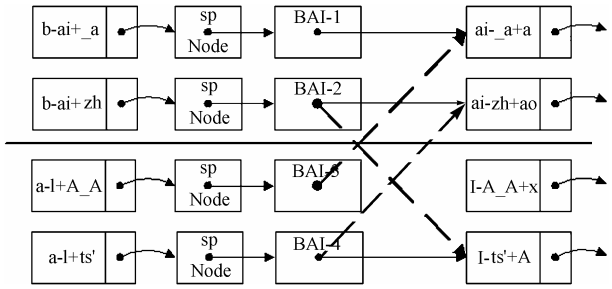

图 6　音节“BAI”连接“A”和“ZHAO”的复合

Fig.6 BAI links to A and ZHAO

3.2　算法描述

根据上面的分析，我们可以根据音节的误识情况决定在声学模型复合时，该音节是否两种模型都使用。这里我们根据训练数据的识别结果来设计算法。记当前音节在声韵母模型和音素模型下的准确率分别是 accIF 和 accPhone，并设置一个准确率阀值 accThresh。算法描述如下：

$$\text{Flag}=\begin{cases}0, & \text{accIF}<\text{accPhone}\text{且}\text{accIF}<\text{accThresh}\\ & \text{且}\text{accPhone}<\text{accThresh}\\ 1, & \text{accIF}\geq\text{accPhone}\text{且}\text{accIF}<\text{accThresh}\\ & \text{且}\text{accPhone}<\text{accThresh}\\ 2, & \text{otherwise}\end{cases}\tag{2}$$

当前音节的 Flag 取 0，1，2 时，分别表示只利用音素模型，只利用声韵母模型和两种模型都使用。

声学模型的复合情况和 accThresh 的值紧密相关。accThresh 的值越大，两种模型都使用的音节就越少，只使用一种模型的音节就越多。极端情况是：当 accThresh 取 100 时，没有音节两种模型都使用；当 accThresh 取负无穷大时，所有音节都使用两种模型。

4　实验结果与分析

4.1　基线模型

数据库：用于训练和测试的语音库分别为 863 汉语语音训练语料库和 863 汉语语音测试语料库。本文使用的训练库包括 75 个女声，75 男声；测试库包括 4 个女声(测试集不包括在训练集内)，所收集语音为 16k 标准汉语连续语音。

声学特征：汉明(Hamming)窗长为 25ms,帧间重叠 15ms，

12 阶美尔倒谱系数加 0 阶倒谱系数，以及各自的一阶、二阶差分共 39 维 MFCC 特征向量。

本文采用HTK v3.0 工具进行模型训练[6]。

上下文相关音素模型：用训练库的全部 150 个说话人的训练数据训练得到，状态输出概率密度函数为 6 混合高斯函数，协方差矩阵为对角阵。

上下文相关声韵母模型：用训练库的 72 个女声的训练数据训练得到，状态输出概率密度函数为单高斯函数，协方差矩阵为对角阵。

4.2 错误率分析

本文选用测试语音库中的 4 个女声为测试集。首先进行单独使用声韵母模型和音素模型的实验，然后取不同的 accThresh 值进行复合程度不同的实验。实验结果如表 2 所示。从表 2 可以看到，随着 accThresh 的不同，系统的识别性能也发生变化。当 accThresh 取 55 时，错误率到达最低值 25.42%，该错误率比声韵母模型和音素模型的错误率分别下降了 9.60%和 6.10%。

4.3 替代错误和插入错误分析

在“引言”中提到，将不同声学模型复合在一起，识别器会自动根据动态规划原理利用较优的模型；并且某些音节有选择性地只利用其中一个声学模型可以减少插入错误。定义替代错误率和插入错误率分别为

$$\mathrm{Sub\%}=\frac{S}{N}\times 100\%,\quad \mathrm{Ins\%}=\frac{I}{N}\times 100\% \tag{3}$$

式(3)中的 N、S 和 I 的意义与式(1)相同。

图 7，图 8 分别显示了各种实验条件下替代错误率和插入错误率的情况。

从图 7 和图 8 可以看到，对于替代错误率，除了“accThresh=100”这个没有任何音节同时使用两种模型的极端情况外，其余复合方法都比单独的声韵母模型和音素模型的替代错误率有一定下降，其中“accThresh＝55”情况下，

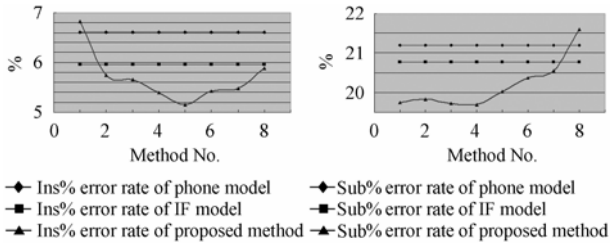

图 7 各种实验条件下复合方法的替代错误率

Fig.7 SUB error rates of proposed method under different conditions

图 8 各种实验条件下复合方法的插入错误率

Fig.8 INS error rates of proposed method under different conditions

替代错误率比单独的声韵母模型和音素模型分别下降了 5.18%和 7.07%。

对于插入错误率，除了“accThresh=负无穷”这个没有进行任何音节选择的极端情况外，其余复合方法都比单独的声韵母模型和音素模型的插入错误率有一定下降，其中“accThresh=60”情况下，插入错误率比单独的声韵母模型和音素模型分别下降了 22.11%和 13.54%。

4.4 与多特征交叉参考的比较

在多特征交叉参考方法[2]中，其中一组识别结果如表 3 所示。表 3 中，“MFCC”列表示选用MFCC为主特征，TLDA和TRAPS为参考特征时系统识别性能；“TLDA”列和“TRAPS”列的意义与此类似。选用不同特征作为主特征时，平均错误率下降了 6.95%。

本文中，当 accThresh 取 55 时，错误率比基线声韵母模型和音素模型的错误率分别下降了 9.60%和 6.10%(平均值为 7.85%)。尽管本文所用的基线系统和多特征交叉参考方法不同，但如果从错误率下降这个指标看，本文的方法取得了比后者略好的性能。

表 2 测试集的识别结果

Tab.2 Recognition results for test sets

复合方法编号	实验条件	同时使用两种模型的音节数	只使用声韵母模型的音节数	只使用音素模型的音节数	准确率(%)	错误率(%)
无	声韵母模型	0	401	0	72.93	27.07
无	音素模型	0	0	401	71.88	28.12
1	accThresh=负无穷	401	0	0	73.11	26.89
2	AccThresh=40	374	16	11	74.14	25.86
3	AccThresh=50	365	19	17	74.29	25.71
4	accThresh=55	352	29	20	74.58	25.42
5	accThresh=60	234	43	24	74.47	25.53
6	accThresh=65	315	56	30	73.81	26.19
7	accThresh=70	291	69	41	73.57	26.43
8	AccThresh=100	0	271	130	72.13	27.87

表 3　多特征交叉参考方法的识别结果

Tab.3 Results for multi-feature reference method

	MFCC	TLDA	TRAPS	平均值
基线模型	27.7%	28.6%	29.9%	28.7%
交叉参考方法	25.8%	26.7%	27.7%	26.7%
错误率下降百分比	6.86%	6.64%	7.36%	6.95%

5　结束语

总体说来，声韵母模型的识别性能高于音素模型。但是，对于特定的音节，音素模型和声韵母模型却各有优劣；本文利用了这个特点，提出了在识别过程中充分利用声韵母模型和音素模型的复合方法。由于该方法一方面使识别器在音素模型和声韵母模型之间做了优化选择，另一方面又使识别器避开了很影响识别性能的干扰模型(具体表现在造成巨大的插入错误)，因此识别性能得到了明显提高。

本文的研究表明，复合与选择思想在不同基元的声学模型结合利用时可以取得一定的效果。在多特征复合和耦合 HMM 的方法中如果利用本文的复合与选择思想，从理论上说，除了可以减少模型的存储空间外，也可以提高系统的识别性能。

参 考 文 献

[1] Fiscus J G. A post-processing system to yield reduced word error rates: Recogniser Output Voting Error Reduction(ROVER). Proceedings of IEEE ASRUWorkshop: Santa Barbara, 1997: 347−352.

[2] Yan Y H, *et al*.. A dynamic cross-reference pruning strategy for multiple feature fusion at decoder run time. In Proc. EUROSPEECH’03 Geneva, 2003.

[3] Nefian A V, Liang L H, Liu X X, Pi X B, Mao C, Murphy K. A coupled HMM for audio-visual speech recognition. International Conference on Acoustics Speech and Signal Processing, Orlando, Florida, May 2002, vol II: 2013−2016.

[4] Ma B, Huo Q. Benchmark results of triphone-based acoustic modeling on HKU96 and HKU99 Putonghua corpora. In Proc. ISCSLP, Beijing, China, 2000: 359−362.

[5] 李净, 郑方, 张继勇, 吴文虎. 汉语连续语音识别中上下文相关的声韵母建模. 清华大学学报(自然科学版), 2004, 44(1): 61−64.

[6] Steve Young S, Kershaw D, Odell J, *et al*.. The HTK Book [EB/OL], http://htk.eng.cam.ac.uk, 2002.

[7] Young S J, Russel N H, Thornton J H S. Token passing:a simple conceptual model for connected speech recognition systems. Technical report, Cambridge University Engineering Department, 1989.

张　辉：　男，1980年生，博士生，研究方向为嵌入式连续语音识别.

杜利民：　男，1957年生，教授，博士生导师，主要从事语音信号处理、语音识别和自然语言理解方面的研究.