基于先验知识的三音子模型聚类结构自适应策略

董 明　　刘润生
(清华大学电子工程系　北京 100084)

摘　要：该文提出了一种基于先验知识的三音子模型聚类结构自适应策略，可以在规模很小的自适应语音库条件下改善三音子声学模型的聚类结构使之更适合应用对象的协同发音特点。以基本声学模型训练过程中的三音子模型聚类结果作为先验知识的聚类中心，依据基本声学模型对自适应语音库的分割，按照最大似然准则迭代地重估新的聚类中心和模型聚类结构。实验表明：基于先验知识的三音子模型聚类结构自适应策略可以在不足两小时的自适应语音库上实现三音子模型聚类结构重估，在针对汉语母语说话人的英语声学模型实验中，该文的模型聚类结构自适应策略可以将系统识别率从 74.59%提高到 83.63%。

关键词：语音识别；三音子模型；模型聚类
中图分类号：TN912.34　　**文献标识码**：A　　**文章编号**：1009-5896(2007)09-2050-04

Transcendental Information Based Triphone Model Tying Structure Adaptation Strategy

Dong Ming　　Liu Run-sheng
(*Department of Electronic Engineering*, *Tsinghua University*, *Beijing* 100084, *China*)

Abstract: A Transcendental Information Based (TIB) triphone model tying structure adaptation strategy is delivered, and this strategy can improve the triphone model tying structure to suit the target co-pronunciation features with small amount of adaptation data. The TIB triphone model tying structure adaptation strategy uses the baseline acoustic model's triphone model tying result as the transcendental model clustering center, with the adaptation data alignment by the baseline acoustic model, re-estimate the TIB triphone model clustering center and model tying structure recursively under maximum likelihood principle. The experiments show that the TIB triphone model tying structure adaptation strategy can improve the triphone model tying structure with only 2 hours' adaptation corpus, and in the experiment of English acoustic model for Chinese speakers, the TIB strategy will increase the recognition accuracy rate from 74.59% to 83.63%.

Key words: Speech recognition;　Triphone model;　Model tying

1　引言

非特定人语音识别系统中，经常会遇到方言母语及外国母语等具有整体性口音差异的应用情况。比如针对四川口音或者针对广东口音的汉语普通话识别系统；或者为了辅助中国的英语学习者练习英语口语的对话系统。这时如使用一般的汉语普通话识别系统或标准英语识别系统，其识别性能将受到严重影响，是当前语音识别技术中亟待解决的问题之一。

考虑到训练语音库采集和标注等的庞大开销以及声学模型建立方法的推广性，一般训练针对某种特定目标应用的声学模型时，都不直接从目标应用说话人采集训练库(约需数十个小时以上的语音)，而只采集一个数小时规模的自适应语音库，再采用 MAP，MLLR 等口音自适应的策略从通用声学模型自适应得到目标应用的声学模型。对于整体性口音差异的情况，这种单纯口音自适应的方法虽然可以取得一定效果，但改进的性能仍不够理想。

单纯的口音自适应效果有限的原因在于，口音自适应只能够改善音子发音上的不一致，但在具有整体性口音差异的问题中，目标群体受到母语的说话习惯等方面的影响非常大，因而目标应用群体与训练通用声学模型的“标准”发音说话人在前后文连读发音特点上也存在很大的差别。所以本文试图同时从协同发音特点的角度来调整目标应用的声学模型，也就是说根据自适应语音库来调整表征协同发音特性的三音子声学模型聚类结构来进一步改善这种情况下识别系统的性能，这一声学模型调整的完整过程如图 1 所表示。

本文针对小规模自适应语音库条件下改善三音子模型聚类结构的问题，提出基于先验知识的三音子模型聚类结构自适应策略，可以实现在不足两小时的自适应语音库上调整

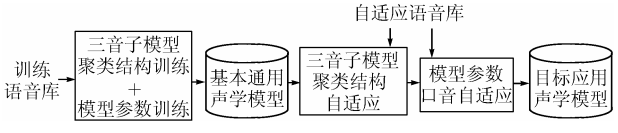

图 1 加入“三音子声学模型聚类结构自适应”的声学模型训练过程

2006-02-27 收到，2006-07-27 改回

三音子模型聚类结构，使之更适合目标应用群体协同发音特点的目标。

2　声学模型训练中的三音子模型聚类问题和常用策略

基于三音子(Tri-phone)子词划分的隐含马尔可夫模型(Hidden Markov Model，HMM)是非特定人语音识别非常主流的方法[1]。三音子子词模型更全面地考虑到一个音子发音同时受其前后两个音子的影响而产生连读差异，用于中文或英文等语言都具有比较优良的性能，然而三音子模型的数量巨大(实际中存在的三音子模型数量，英文系统超过 5 万个，中文有声调系统要超过 20 万个[1, 2])，要直接估计全体三音子声学模型所需的训练库规模十分庞大，所以三音子声学模型的训练过程中模型聚类是一个必需的步骤。

一个单音子(Mono-phone)模型发展而成的全部三音子模型，其实前后文协同发音的差异并不必要被刻画到完全互相各异的程度，比如t-ou+n和b-ou+n的ou(音标/∂u/)其发音并没有很大的区别。三音子声学模型聚类就是通过一定的聚类策略把协同发音比较近似的三音子模型合并到同一个看待，共享训练数据得到模型参数，以实现三音子声学模型的训练并达到比较稳健的声学模型参数估计。常被使用的模型聚类策略，主要是数据驱动的自底向上模型状态聚类策略[3]和基于决策树的自顶向下模型聚类策略[4]。

数据驱动的模型状态聚类策略基于距离测度，对每个单音子模型发展出来的各个三音子模型，计算相对应位置的模型状态间的距离，把距离较近(相似度较大)的模型状态互相合并在一起成为一个模型状态。经常使用的距离测度包括类马氏距离、散度距离等等。

决策树的模型聚类主要依据语音学的知识，以每个单音子模型的全部三音子模型作为根节点向下发展出一棵决策树，常常采用二叉树自顶向下地生长。从根节点开始在每一个非叶子节点上通过回答一个语音学问题(如: 该音子前面的音子是爆破音么？或者，该音子后面的音子是鼻音么？)，按照当前节点的所有音子模型“是”或“否”的不同回答把它们进一步分开到两个子节点上。通过分裂前后的模型对于训练数据的似然度分数增量选择当前节点的语音学提问，或者在似然度分数增量不足时停止该节点的继续分裂，最后不再继续分裂的各个叶子节点上的音子模型被聚类在一起看待。

3　基于先验知识的三音子模型聚类结构自适应策略

声学模型训练中对三音子模型进行聚类的方法中，无论是自底向上的数据驱动聚类，还是自顶向下的决策树模型聚类，都要求一个前提即预先训练出一套全体三音子模型的参数，通过该参数进行似然度分数或者距离测度的计算，实现模型状态的分割或合并。但是一般进行模型自适应调整的语音库，其规模往往只有训练语音库规模的几十分之一，这样小规模的语音库因为大量的三音子模型观察矢量不足而根本无法训练出来一套全体三音子声学模型参数，也就更无从谈起进行模型聚类的工作。

如果我们通过对已训练好的全体三音子模型，应用自适应语音库经过 MLLR 或 MAP 口音自适应的方法得到自适应的三音子模型，这样做的问题在于，新得到的三音子声学模型里会保留相当多的原始模型特性，基于这样的全体三音子模型来进行模型聚类不能达到按照自适应库的协同发音特点调整三音子模型聚类结构的目标。

如何针对小规模的自适应语音库改善三音子模型聚类的结构，进一步地分析，事实上通过原始声学模型训练过程中的三音子模型聚类我们已经得到了一个针对训练集说话人的三音子模型聚类结果，包括各个聚类中心和每个三音子模型归属于哪个类别的聚类结构，我们可以合理地假设针对目标应用说话人的三音子模型聚类结构与该聚类结构具有相当的相似性。同时自适应语音库的数据规模虽无法对全体三音子模型重估参数，要达到重估已聚类的三音子声学模型参数还是容易的。因此以原始声学模型的三音子模型聚类结果作为先验知识的聚类中心，小规模自适应语音库就可以达到对三音子模型聚类的声学模型进行重估，提出基于先验知识(Transcendental Information Based，TIB)的三音子模型聚类结构自适应策略。

TIB 三音子模型聚类结构自适应策略根据原始声学模型的聚类结构产生初始的聚类中心种子，从 N=2 类开始重估每个单音子模型分成 N 个聚类的模型结构，N 值逐一递增。依据原始声学模型对自适应语音库的切分，将各个三音子模型按照其语音观察数据最大似然地归入 N 分类当中迭代重估得到新的聚类中心和聚类结构。通过对声学模型和训练数据的似然度分数设定增量阈值，决定每个单音子模型分成 N 个聚类数目的继续增加或停止，TIB 三音子模型聚类结构自适应策略的简要流程如图 2。

(1) 先验种子序列　在模型聚类后的原始三音子声学模型中，一个单音子模型(序号 p，$p=1,\cdots,P$)发展的全部三音子模型被聚类到 N_p 类，按照每一类所包含的三音子模型

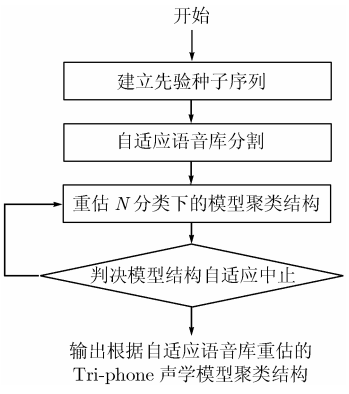

图 2　TIB 三音子模型聚类结构自适应策略的流程

数量从大到小排序，将每一类对应的声学模型即类中心记作 $\lambda_{01},\lambda_{02},\cdots,\lambda_{0N_p}$。该 N_p 个原始声学模型的类中心作为重估第 p 个单音子模型新的聚类结构的先验聚类中心种子序列。

(2) 自适应语音库分割　根据原始声学模型，求取自适应语音库在三音子模型状态级别上的分割信息，得到了全体三音子模型分别对应的语音观察矢量集合，第 (p,t) 个三音子模型对应的语音观察矢量集合记为 $X_{p,t}$ ($p=1,\cdots,P,\ t=1,\cdots,T_p$)。

(3) 重估 N 分类下的模型聚类结构　将一个单音子模型发展的全部三音子模型划分成 N 个类，重新估计模型聚类结构。N 值从 2 开始，即一个单音子模型至少被划分成 2 个类别，随着 N 值逐一递增，三音子模型聚类结构也随之精细。

步骤 1 选取类中心的初值:

$$\lambda_1=\lambda_{01},\ \lambda_2=\lambda_{02},\cdots,\lambda_N=\lambda_{0N}\ (N\le N_p) \tag{1}$$

步骤 2　计算第 (p,t) 个三音子模型($t=1,\cdots,T_p$)对应的语音观察矢量集合，对 N 个类中心的对数似然分数：$\log P\left(X_{p,t}\mid\lambda_n\right)$，按照最大似然准则选择第 (p,t) 个三音子模型所归属类别：

$$\begin{aligned} n&=\arg\max_n \log P\left(X_{p,t}\mid\lambda_n\right)\\ &=\arg\max_n \sum_{x\in X_{p,t}}\log P\left(x\mid\lambda_n\right)\end{aligned} \tag{2}$$

步骤 3　第 p 个单音子模型发展的全部三音子模型归类完毕后，对 N 分类 $\lambda_1,\lambda_2,\cdots,\lambda_N$ 的每一类，记录当前的三音子模型分类归属；并依照其所包含的三音子模型对应的语音观察矢量，重新训练出该类的声学模型即类中心来。

步骤 4　如果对比上一次的三音子模型归类，发生分类改变的三音子模型数量低于一定阈值，或者迭代达到最大容许次数，则本次 N 分类下的模型结构重估中止；否则跳转到步骤 2。

观察数据不足(Un-Seen)的三音子模型处理：不是每个三音子模型都在自适应语音库中出现并有足够观察，对于越小规模的自适应库这种情况更是普遍，本文基于如下的流程对自适应库中未观察到的三音子模型处理。若该模型在原始模型聚类结构中属于 $\lambda_{01},\lambda_{02},\cdots,\lambda_{0N}$，选择其在原始模型的对应类归入；若该模型在 $\lambda_{0(N+1)},\lambda_{0(N+2)},\cdots,\lambda_{0N_p}$ 中，求其所在原始类中心到 $\lambda_1,\lambda_2,\cdots,\lambda_N$ 分别的散度距离，选择距离最小的类别归入。

(4) 增加 N 分类数目并判决模型结构自适应中止　令 N=N+1，重复重估 N+1 分类下的模型聚类结构。

对于一个单音子模型发展的全部三音子模型聚类到 N 类和 N+1 类情况下的两个声学模型，分别计算模型对于训练数据的对数似然分数。设定分类增加的对数似然度阈值，当对数似然分数增量不足时停止该单音子模型的继续扩展。

另外，N 分类的数目不超过 N_p，即先验种子的数量。

最后全部的单音子模型停止扩展，TIB 模型聚类结构自适应过程结束。

4　实验结果和分析

TIB三音子模型聚类结构自适应策略的实验基于针对汉语母语说话人的英语整词识别系统进行。汉语母语说话人的英语识别问题，其应用背景是为中国英语学习者的口语学习机，通过交互式对话的互动听说，为英语学习者营造出一个人与学习机的双向英语交流环境，从而彻底改变现有学习机的单方向授课的教学模式[5]。

英语母语声学模型的训练库采用从美国购买的包含 100 名男性和 100 名女性、总语音时间为 73h 的 WSJ1 连续语音库，训练过程中三音子模型聚类使用决策树结合散度距离的数据驱动聚类策略。英语母语声学模型经过一个本课题组从 38 名男性和 40 名女性汉语母语说话人采集的总语音时间为 1.7h 的 L2_Adapt 连续语音调整库，进行自适应调整得到汉语母语说话人的英语声学模型。实验的两套声学模型差别在于前者直接使用英语母语库训练得到的原始模型聚类结构，后者的三音子模型聚类结构以 TIB 策略针对汉语母语调整库进行了模型聚类结构自适应，两套声学模型均以 MLLR+MAP 的方法用汉语母语调整库 L2_Adapt 进行了口音自适应。

识别测试对于本课题组从 16 名男性和 16 名女性汉语母语说话人采集的 350 个英语整词库 L2_Word _Te。WSJ1，L2_Adapt 和 L2_Word_Te 语音库都是普通办公室环境 16kHz 采样 16bit 量化，说话方式是自然语音。语音特征参数是从 39 维的 MFCC 参数经过特征选择得到 30 维，声学模型是 50 个有声音子的三音子子词模型，TIB 三音子模型聚类结构自适应策略作用在 HMM 模型状态级进行，具体的实验结果如表 1 所示。

表 1　TIB 三音子模型聚类结构自适应策略针对汉语母语说话人的英语整词识别实验结果(测试集共 350 词)

使用的声学模型	包含的 HMM 模型状态数(个)	识别正确率(%)
原始模型聚类结构(无口音自适应)(直接使用英语母语说话人模型)	1612	57.64
原始模型聚类结构+口音自适应	1612	74.59
TIB 模型聚类结构自适应+口音自适应	1137	83.63

通过实验结果可以看到，单纯通过口音自适应的方法虽然对系统性能可以有相当改善，但进一步通过 TIB 三音子模型聚类结构自适应策略调整后的声学模型则具有更加优越的识别性能，相对于原始聚类结构的模型其误识率下降达到了 35.6%，这主要源于重估后的声学模型结构更符合汉语母

语说话人讲英语时的协同发音特点。

另外，以 TIB 策略进行模型聚类结构自适应后的声学模型比原始声学模型 HMM 模型状态个数少一些，这一点应该主要因为英语对于汉语母语说话人来说是外语而非母语，因此汉语母语说话人的英语连读发音一致性不足，相对聚类粗糙的模型会更适合于这种一致性不足的特点。TIB 策略本身并不具有如此大规模压缩声学模型规模的作用。

5　结束语

本文提出基于先验知识的三音子模型聚类结构自适应策略，来解决小规模自适应语音库条件下改善三音子模型聚类结构的问题，达到使声学模型更适合目标应用群体的协同发音特点的目标。以原始声学模型中三音子模型聚类的类中心作为 TIB 模型结构自适应的先验种子序列，依据基本声学模型对自适应语音库的分割信息，按照最大似然准则迭代重估新的聚类中心和模型聚类结构，设定声学模型对训练数据的似然度增量阈值控制 TIB 三音子模型聚类结构自适应的中止。

最后的实验数据，TIB 三音子模型聚类结构自适应策略针对汉语母语说话人的英语整词识别实验，系统识别率从 74.59%提高到 83.63%，相对误识率降低了 35.6%。

参 考 文 献

[1] Lee K F and Hon H W. Speaker-independent phone recognition using hidden Markov models. *IEEE Trans on ASSP*, 1989, 37(11): 1641–1648.

[2] Chang E, Shi Y, Zhou J L, and Huang C. Speech lab in a box: A Mandarin speech toolbox to jumpstart speech related research. Eurospeech 2001, Aalborg, Denmark, 2001.

[3] Young S and Evermann G, *et al.*. The HTK Book (for HTK Version 3.2). Cambridge University Engineering Department, 2002.

[4] Lazarides A, Normandin Y, and Kuhn R. Improving decision trees for acoustic modeling. Proceedings of ICSLP'96. Philadelphia, 1996: 1053–1056.

[5] Liang W Q, Liu J, and Liu R S. An automatic pronunciation quality assessing algorithm for computer assisted language learning. *Chinese Journal of Electronics*, 2005, 14(4):639–643.

董　明：男，1978 年生，助理研究员，研究方向为语音识别及英语发音评测.

刘润生：男，1933 年生，教授、博士生导师，研究方向为嵌入式语音处理技术及语音处理专用芯片.