ISSN 1000-0054 清华大学学报（自然科学版）2011 年 第 51 卷 第 9 期 1/33
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2011, Vol. 51, No. 9 1161-1166

面向多口音语音识别的声学模型重构

张 超[1,2]， 刘 轶[1]， 郑 方[1]

（1. 清华信息科学技术国家实验室技术创新与开发部语音和语言技术中心，北京 100084；2. 清华大学 计算机科学与技术系，北京 100094）

摘 要：该文提出了应用声学似然分作为置信度来生成可靠口音相关单元的方法。基于可靠口音相关单元构造声学模型，并通过声学模型重构的方法将它们融合到标准普通话模型中，以改善普通话语音识别器对带多方言口音语音的识别效果。另外，还提出了使用增量式决策树融合及根据支配度选择 Gauss 混合 2 种方法来减少冗余的 Gauss 混合，从而提高了重构后的声学模型的效率。实验表明：该方法在不降低对标准普通话的识别率的前提下，对粤、吴口音的绝对音节错误率分别下降了 9.25%和 9.21%。

关键词：语音识别；多方言口音；可靠口音相关单元；声学模型重构

中图分类号：TN 912.3 文献标志码：A

文章编号：1000-0054(2011)09-1161-06

Acoustic model reconstruction for multi-accent Chinese speech recognition

ZHANG Chao[1,2], LIU Yi[1], ZHENG Thomas Fang[1]

(1. Center for Speech and Language Technologies, Division of Technology Innovation and Development, Tsinghua National Laboratory for Information Science and Technology, Beijing 100084, China;
2. Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract: The acoustic likelihood score is used as a confidence measure to generate reliable accent-specific units and to merge such reliable accent-specific units through acoustic model reconstruction. The decision tree merge and acoustic model reconstruction efficiencies are improved by reducing redundant Gaussian components through an incremental decision tree merge procedure and selection of Gaussian components according to their dominance. Tests on Cantonese and Wu accents show that this approach yields significant 9.25% and 9.21% absolute syllable error rate (SER) reductions without degrading the performance on standard Putonghua.

Key words: speech recognition; multiple accents; reliable accent-specific unit; acoustic model reconstruct

汉语中共有八大方言，即：官话、吴语、湘语、赣语、客家语、闽南语、闽北语以及粤语[1]。其中，官话是与标准普通话最为接近的一种方言，其他各种方言在声学发音以及语言学表现上都与标准普通话有着显著的差异。由于多数普通话使用者把普通话作为第二语言来掌握，他们的普通话发音不可避免地受到其方言母语发音的强烈影响[2]。相关文献指出，80%左右的普通话使用者带有不同程度的方言口音[3]。由前所述，方言口音在汉语语音中是一个严重的问题。实践表明：当说话人带有某种方言口音时，针对标准普通话构造的语音识别器的性能往往会大幅下降。

口音相关单元(accent specific unit)常用来表征口音中的发音变异，并有多种使用方法。其中，用口音相关单元来扩展发音基元是一种被广泛采用的办法[4-5]。然而，引入的扩展发音基元以及带有多种候选发音的多发音字典往往会增加语音识别器的词汇混淆度。Oh 和 Kim 利用发音变异基元在最大后验概率(maximum a posteriori, MAP)和最大似然线性回归(maximum likelihood linear regression)自适应方法中生成额外的回归分类，从而改善了对带口音语音的自适应效果[1,6-8]。这类自适应方法的 1 个主要缺点是会不可逆地改变声学模型的参数，从而导致自适应后的模型不再适用于标准语音或其他口音。针对以上方法的缺点，Liu 和 Fung 提出了使用口音相关单元构造声学模型，并将这些模型通过声学模型重构及状态级发音模型等方式用作隐式模型[2,9]。这种方法不但同时对多种口音有效，而且不降低系统对标准语音的识别效果。不过，这种方法在如何选择恰当的口音相关单元，以及如何减少冗余的 Gauss 混合以保持恰当的模型复杂度方

收稿日期：2011-07-15

基金项目：国家自然科学基金资助项目（60975018）；
教育部新教师基金（20090002120012）

作者简介：张超(1986—)，男(汉)，内蒙古，硕士研究生。

通信作者：刘轶，副研究员，E-mail: eeyliu@tsinghua.edu.cn

面仍然存在着难点。

本文提出了使用声学似然分作为置信度来筛选可靠的口音相关单元的方法。基于筛选出的可靠口音相关单元来构造独立的口音相关声学模型,再利用决策树融合和声学模型重构的方法把这些模型融合到它们对应的标准普通话声学模型中。这样,可以在一个语音识别系统中既同时处理多种方言口音,又不降低系统对标准普通话的识别率。另外,为避免从口音相关模型中引入冗余的 Gauss 混合,与传统方法不同的是,本文提出了增量式决策树融合的方法,并根据发音变异的混淆度从对应的口音相关模型中引入不同数目的 Gauss 混合,从而对不同重要性的发音变异进行合理的区别处理。综合使用上面提出的方法,在不牺牲语音识别系统对标准普通话识别率的前提下,大幅改善了系统对多种方言口音语音的识别率,同时有效地控制了声学模型复杂度。与前期对某种口音发音变异进行分层,并使用自适应或上下文无关的状态级发音模型来优化声学模型的方法相比[1,4],本文构建了上下文相关的发音模型,使得重构后的声学模型既保持对普通话的识别率,又能同时改善对多种方言口音语音的识别率。

1 汉语方言口音

各种汉语方言中的发音往往与普通话中的标准发音有着明显的差异。例如,语言学研究表明:粤语和吴语中分别只有 60%和 70%的发音与普通话的接近[2]。而且粤语和吴语之间的发音差异也非常显著,语言学家甚至把这 2 种方言在音系、词汇以及句法结构上视作 2 种不同的语言。不同汉语方言口音间存在着区别和联系,本文使用粤语和吴语为例来说明这点。

在汉语语音识别系统中,通常使用声韵母作为子词单元(sub-word unit)来构造声学模型。一个声母通常对应于一个音素,而一个韵母则可能是单元音、二合元音或三合元音。标准普通话中共有 22 个声母和 36 个韵母,粤语中共有 20 个声母和 53 个韵母,吴语中则包含了 35 个声母和 41 个韵母[10]。这样,不同方言间的声母存在着显著的差异。例如,与普通话相比,粤语包含一个额外的后鼻音“ng”,并且粤语中“f”和“x”这 2 个声母的发音相同。吴语中的声母分成了清和浊 2 类,而普通话中只有清声母,比如普通话中的清声母“f”在吴语中则是浊声母。对于韵母来说,吴语的韵尾“n”与“ng”在发音上没有区别。另外吴语比普通话具有更多的单元音韵母,例如普通话中的韵母“an”对应了吴语中的/ae/、/oe/、/ie/这 3 个单元音韵母。粤语的韵母也要比普通话的韵母结构更为复杂,相比于普通话仅有 2 种韵尾“n”和“ng”,粤语包含了“m”、“n”、“ng”、“p”、“t”和“k”这 6 种韵尾。

从上面的发音差异可以看出,母语是粤语或吴语的人在拼读某些普通话的声韵母时会遇到困难,从而导致发音变异的出现。本文把母语是某种方言的说话人在拼读普通话时带有的发音变异的共性称为方言口音。带有方言口音的普通话在不同程度上具有它们对应方言的发音的特点。比如,当母语是吴语的说话人发普通话韵母“zh”的音时,他的发音很可能既像“zh”又像“j”。因此,仅使用标准普通话的声韵母无法完全表示带方言口音语音中可能遇到的所有发音。

虽然不同方言口音之间存在着差异,但也存在着相似性。比如,粤语和吴语都没有卷舌音“sh”,却都有“s”。这样,带粤、吴 2 种方言口音的普通话中都存在着“sh”到“s”的发音变异。于是,由于发音变异本身的复杂性,需要一种自动化的方法来找出 2 种不同方言口音间重叠的发音变异,以避免对相似变异的重复建模。同时,一个先后在粤语和吴语地区生活过的说话人可能带有这 2 种方言的混合口音。在极端情况下,混合口音发音变异的复杂性会导致“zh”的发音变异遍及从“zh”到“z”的整个分布[2]。因此,对多方言口音需要进行更加灵活的声学模型建模。

2 可靠口音相关单元的建立

口音相关单元通常被用来表征由口音产生的发音变异。设 B 是变异前的基准发音,S 是变异后的表象发音,一个典型的口音相关单元可以写作“$B\rightarrow S$”[2]。使用规则或者数据驱动的方法,可以从 1 种口音中提取 1 套口音相关单元。因此,通过对多套口音相关单元进行联合建模,可以表征出多口音中发音变异的多样性。

2.1 生成可靠口音相关单元

为了获得口音相关单元,本文使用声韵母级别的人工标注作为基准发音序列;同时使用自由基元识别[12]进行声韵母循环的识别,识别结果作为表象发音序列。使用基于动态规划的对齐工具(dynamic programming alignment tool)来获得基准发音和表象发音间的对齐序列[11],进而得到口音相关单元的

所有可能候选。通常来说，这些候选的数目很多，而其中许多候选可能仅有很少的样本出现在对齐序列中，因此并不适合用来训练口音相关模型。本文丢弃所有出现次数少于 30 次的候选，并把剩余的候选称作初始口音相关单元。初始口音相关单元中包含有真实的发音变异、数据和识别器错误，以及普通话中的固有混淆[11]。本文通过参考普通话的混淆矩阵及语音学知识，从初始口音相关单元中筛选出真正的口音相关单元。部分被筛除的初始相关单元如表 1 所示。

表 1　筛除的部分初始口音相关单元

基准发音	zh	s	q	p	c	b	b	_i	ao	a
表象发音	sh	z	t	t	t	p	d	_v	e	an

2.2　可靠口音相关单元

并非所有口音相关单元都适合用于声学模型重构。本文用图 1 来说明口音相关单元可靠性的重要，其中：C_0 是某一维的均值，$P(C_0)$是相应的输出概率。假设 B 和 S 是不同的发音基元，它们各自的声学模型都只包含一个 Gauss 混合，分别记做 $G_B(\mu_B,\delta_B)$和 $G_S(\mu_S,\delta_S)$，其中 μ 和 δ^2 是 Gauss 混合的均值和方差。X_B 指代 B 的声学样本。易知当 $P(X_B|B)$小于 $P(X_S|S)$时就发生了误识别，X_B 被错分类为 S，所有这些被错分的口音相关单元 $B\rightarrow S$ 的样本都位于图 1 中决策面(图 1 中虚线)的左边。当这些样本都聚集在图 1 中 X_1 附近时，由于 $P(X_1|B)$与 $P(X_1|S)$大小相近，通过对 B 的模型分布进行重构来使 $P(X_1|B)>P(X_1|S)$是合理的。否则，当 $B\rightarrow S$ 的样本集中在 X_2 附近时，由于 $P(X_2|S)$显著大于 $P(X_2|B)$，说明 S 的许多样本也在 X_2 附近，因此对 B 的分布进行变形来使 $P(X_2|B)>P(X_2|S)$，将会导致 S 的这些样本被错分为 B，从而不可行。

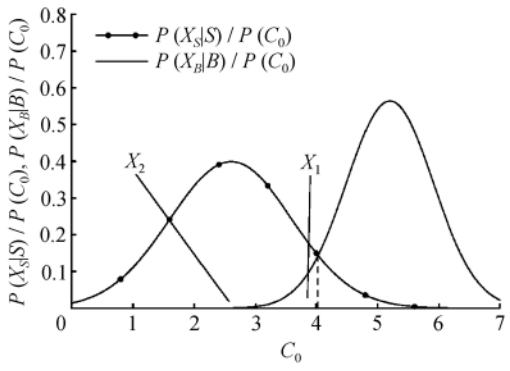

图 1　口音相关单元的可靠性的例子

假设 $Y_1Y_2\cdots Y_M$ 是样本 X 的语音帧序列，则 X 属于基元 B 时的平均声学似然分定义为

$$\mathrm{score}(X,B)=-\frac{1}{M}\sum_{i}^{M}\ln P(Y_i\mid B). \quad (1)$$

于是，用于选择可靠口音相关单元的置信度可以写为

$$\frac{1}{N}\sum_{j}^{N}(\mathrm{score}(X_j,B)-\mathrm{score}(X_j^*,S)). \quad (2)$$

其中：N 是 $B\rightarrow S$ 的样本的数目；score(X,B)和 score(X,S)分别是基准发音 B 和表象发音 S 的声学似然分，这些声学似然分通过对基准发音序列和表象发音序列分别作强制对齐来获得[12]。

表 2 列出了筛选出的部分粤口音的可靠口音相关单元。

表 2　部分粤口音可靠口音相关单元

基准发音	zh	zh	zh	sh	sh	n	d	d	c	ei
表象发音	z	s	j	s	r	l	p	p	z	ui

表 3 列出了实验中得到的部分吴口音的可靠口音相关单元。

表 3　部分吴口音可靠口音相关单元

基准发音	zh	zh	z	x	sh	sh	sh	r	ch	ch
表象发音	z	j	j	q	x	s	f	l	q	c

比较表 2 和 3 可以发现，2 种口音具有一些共同的可靠口音相关单元。但是，由于不同口音中对应于相同发音变异的口音相关单元往往具有不同的变异趋势和声学参数，并不能直接将这样的可靠口音相关单元进行合并处理[2]。

3　声学模型重构

为了对不同口音各自建模，本文针对每种口音对应的口音相关单元分别构造三音子模型。通过声学模型重构可以调整基准发音模型的输出分布的形状，以使基准发音模型的分布边界具有更大的灵活性、鲁棒性，从而覆盖相应的发音变异。

3.1　辅助决策树和声学模型重构

本文使用基于决策树聚类进行状态共享的三音子声学模型。可靠口音相关单元对应的三音子模型的结构与标准的三音子模型类似，只是决策树的中心基元并非普通话声韵母，而是相应的可靠口音相关单元。本文将可靠口音相关单元对应的决策树称作辅助决策树，而将标准普通话模型的决策树称作标准决策树[11]。

使用决策树融合的方法对基准发音的模型进行重构。在使用决策树聚类进行状态共享的三音子声学模型中,对三音子模型的重构等价于对标准决策树和辅助决策树的叶节点进行融合。辅助决策树上的每个叶节点都融合到其基准发音对应状态的标准决策树上距离它"最近"的某个叶节点中[11],如图 2 所示。即基准模型从有关的可靠口音相关模型中"借"Gauss 混合来调整它的输出分布的结构,以达到能够覆盖相应发音变异的目的。传统决策树融合算法详见文[11]。

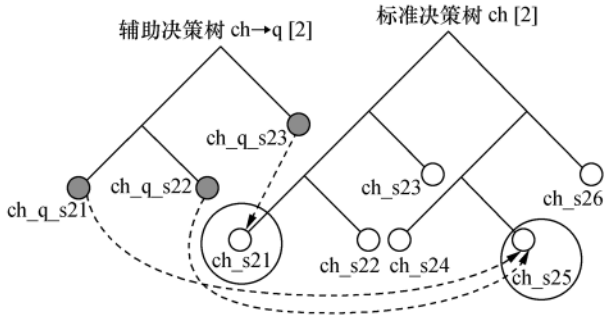

图 2 决策树融合示意图

3.2 增量式决策树融合

传统决策树融合算法中,在所有辅助决策树叶节点与标准决策树叶节点之间的对应关系都被确定下来之前,不将任何一个辅助决策树叶节点真正融合到对应的标准决策树叶节点中。即在计算辅助决策树的每个叶节点与对应的标准决策树的哪个叶节点距离最近的过程中,标准决策树的所有叶节点的输出分布是固定不变的。但是,实际上每融合 1 个辅助决策树叶节点,都会改变标准决策树叶节点的概率密度函数中 Gauss 混合的数目和输出分布的形状。当某个标准决策树叶节点融合了若干辅助决策树叶节点后,这个标准决策树叶节点的输出分布可能已经被变形到足以覆盖某些尚未融合的辅助决策树叶节点对应的发音变异。为减小融合后声学模型的大小,本文提出增量式的决策树融合方法。

在增量式决策树融合算法中,每当一个辅助决策树叶节点与标准决策树叶节点之间的对应关系被确定下来,就马上进行相应的融合以改变标准决策树叶节点的输出分布。也就是说,标准决策树叶节点中概率密度函数在整个融合过程中被"实时"更新。假设辅助决策树的 N 个叶节点被最终融合到标准决策树的某个叶节点中,在前 $i(1\leqslant i\leqslant N)$ 次融合后,该节点的输出概率密度函数 $P_i(x\mid b)$ 可以写为

$$\lambda_i P_{i-1}(x\mid b)+(1-\lambda_i)P(x\mid s_i)P(s_i\mid b). \quad (3)$$

其中

$$\lambda_i=\begin{cases}\dfrac{\lambda}{\lambda+P(s_1\mid b)}, & i=1;\\ \lambda+\sum\limits_{j=1}^{i-1}P(s_j\mid b) & \\ \dfrac{}{\lambda+\sum\limits_{j=1}^{i}P(s_j\mid b)}, & i>1.\end{cases} \quad (4)$$

式(4)中:$P_0(x|b)$为已有的基准发音的模型的输出分布;λ 是线性插值系数的初始值,设置为基准发音被误识别为对应表象发音的概率;$P(s_i|b)$是口音相关模型 s_i 及其对应基准模型 b 之间的混淆度。通过迭代很容易证明:通过式(4)最终得到的输出分布等价于通过原始决策树融合方法得到的输出分布。也就是说,增量式决策树融合算法在不改变原有决策树融合算法融合原理的前提下,给原算法增加了"实时性"。

对决策树融合时辅助决策树叶节点与其最近的标准决策树的叶节点的距离设置一个阈值,当最近距离小于该阈值时,可以认为辅助决策树叶节点的输出分布与对应标准决策树叶节点的输出分布相近,这时的标准决策树叶节点已能够覆盖该辅助决策树叶节点对应的发音变异。于是这样的辅助决策树叶节点不再被融合到任何叶节点中,从而减小了融合后的模型的大小。

3.3 Gauss 混合的支配度

为避免重构后声学模型大小的膨胀,传统的声学模型重构方法丢弃掉了混淆度较低的口音相关单元,即使它们具有足够多的训练样本。然而,根据长尾理论(long tail theory)[13],所有这些被丢弃的低混淆度的口音相关单元的混淆度之和仍然可能占据总混淆度的很大比例。例如,在本文的实验中,韵母"e"共有 24 个可靠口音相关单元,其中 17 个的混淆度均小于 2%,但这 17 个单元共占据了"e"总混淆度中的 41.84%。因此,通过保留这些低混淆度的可靠口音相关单元,可以显著增加重构后模型的对发音变异的覆盖能力,从而带来识别率的提升。

然而,保留所有低混淆度的口音相关单元会造成辅助决策树数目的显著增加。为了使融合后的声学模型仍然具有适当的大小,可以不"借用"辅助决策树叶节点的概率密度函数中的所有 Gauss 混合,而选取具有最大支配度的 Gauss 混合[11],并融合进对应的基准发音模型中。如果状态 S 的概率密度函数仅包含一个 Gauss 混合 $G_S(\boldsymbol{\mu}^S,\boldsymbol{\delta}^S)$时,那么某

Gauss 混合 $G_M(\boldsymbol{\mu}^M,\boldsymbol{\delta}^M)$对该状态的支配度定义为

$$\mathrm{dom}(G_M,S)\left[\sum_{i=1}^{d}\frac{(\mu_i^M-\mu_i^S)^2}{\delta_i^M\delta_i^S}\right]^{\frac{1}{2}}. \qquad (5)$$

其中 d 是概率密度函数的维度。

当 S 的概率密度函数包含多个 Gauss 混合时，使用文[11]中的 Gauss 混合缩减公式将其概率密度函数缩减为单 Gauss 混合的形式。

4　实　验

实验使用国家“八六三”高技术项目四大方言普通话语音语料库[14]以及 SONY 普通话数据库[4]。国家“八六三”高技术项目四大方言普通话语音语料库是目前规模最大、应用最为广泛的中文带口音语音数据库。

所有语音数据均为 16 kHz 采样率和 16 b 精度。数据库及数据集划分如表 4 所示。TrainSetP 以及 TestSetP 是从 SONY 普通话数据库中划分的训练集和测试集。DevSetC 和 TestSetC 分别是从国家“八六三”高技术项目四大方言普通话语音语料库中划分的粤语口音开发集和测试集；DevSetW 和 TestSetW 分别是从国家“八六三”高技术项目四大方言普通话语音语料库中划分的吴语口音开发集和测试集。

表 4　数据集和数据集划分

数据集名称	语音总时长/h	音节总数	说话人数	语句总数
TrainSetP	51.5	340 556	100	25 920
TestSetP	7.2	36 667	32	4 100
DevSetC	6	50 847	20	2 944
TestSetC	6	56 480	20	3 075
DevSetW	6.2	51 559	20	3 156
TestSetW	6.7	55 378	20	3 420

基线系统的声学模型使用 TrainSetP 训练，HMM 模型的拓扑结构为 3 状态，从左到右无跨越式。使用的声学特征为 13MFCC(Mel frequency cepstrum coefficient)、13ΔMFCC、13ΔΔMFCC。使用包括 6 个零声母在内的标准普通话的 28 个声母和 36 个韵母，作为构建上下文无关 HMM 模型的子词单元。使用 HMM 工具包(hidden Markov model toolkit，HTK)基于决策树聚类的状态共享方法构造了 3 000 状态、每状态 12 Gauss 混合的三音子声学模型[12]。在所有实验中使用含有 412 个标准音节的音节字典。

从 DevSetC 和 DevSetW 中分别提取了 221 和 209 个可靠口音相关单元。在此基础上，使用决策树状态共享方法为粤语口音构建了 633 个辅助决策树和 811 个叶节点；为吴语口音构建了 627 个辅助决策树和 645 个叶节点。通过增量式决策树融合算法，每个口音的所有辅助决策树都被融合到 192 个标准决策树共 3 000 共享状态的标准普通话模型中。当可靠口音相关单元的混淆度大于 0.15 时，从其辅助决策树的每个叶节点中选择 2 个支配度最大的 Gauss 混合融合到标准普通话模型中；否则，只选择一个支配度最大的 Gauss 混合。重构后的声学模型包含 37 307 个 Gauss 混合，平均每个状态包含 12.43 个 Gauss 混合。为了更公平地比较方法的有效性，混合 TrainSetP、DevSetC 和 DevSetW 训练了一个具有 3 109 个共享状态以及总计 37 308 个 Gauss 混合的增强型基线系统。所有比较实验的实验结果如表 5 所示。表 5 括号中的数据为相对于基线系统的绝对音节错误率(SER)下降。

表 5　含重构的声学模型的系统的 **SER**

系统名称	SER/%		
	TestSetC	TestSetW	TestSetP
基线系统	57.29	56.52	24.54
增强型基线系统(TrainP，DevSetC，DevSetW)	49.09(−8.20)	48.65(−7.87)	25.32(+0.78)
基线系统+MAP 自适应(DevSetC，DevSetW)	48.60(−8.69)	45.62(−10.90)	30.90(+6.36)
重构的声学模型系统	48.04(−9.25)	47.31(−9.21)	24.05(−0.49)

从表 5 可以看到，只用标准普通话语音构造的语音识别器的识别率明显受到了口音带来的负面影响。与基线系统相比可以发现，重构的声学模型在 TestSetC 和 TestSetW 上分别带来了 9.25% 和 9.21% 的绝对 SER 下降。这说明声学模型重构提高了模型输出分布的边界结构鲁棒性，使其能够同时覆盖多种口音的发音变异[2]。另外，在所有测试集上，重构的声学模型的性能均比通过重训练得到的增强型基线系统的好。这个结果说明：利用可靠口音相关单元对发音变异进行准确的显式建模，其性能比混合所有数据直接训练模型的更好。这一方面是因为对发音变异进行显式建模使得在同样 Gauss 混合数的条件下，重构的声学模型的 Gauss 混合数目针对不同发音变异分布得更加合理；另一

方面是因为在生成可靠口音相关单元构建模型时,使用基于声学似然分的置信度准则对可能增加模型混淆度的候选单元进行了筛除,使得重构的模型的不同输出分布之间的区分度更好。对发音变异的长尾效应的重视也使得重构的模型对多样的口音变异具有更强的覆盖能力。另外值得注意的是,声学模型重构的方法比重新训练新的声学模型的方法的时间开销更小。

从表5还可以发现,混合DevSetC和DevSetW这2种口音的数据进行MAP自适应,可以使模型同时兼具对这2种口音的鲁棒性。表5中MAP自适应系统对粤、吴2种口音的绝对SER分别下降了8.69%和10.90%。然而,这种对多口音的鲁棒性是以牺牲系统对标准普通话的适应性为代价的,自适应后的系统对标准普通话的绝对SER上升了6.36%。与此不同的是,重构的声学模型中同时包含有原有的对应标准普通话的Gauss混合,以及从可靠口音相关模型中"借"来的对应发音变异的Gauss混合。这些"借"来的Gauss混合增加了原模型输出分布结构的灵活性,使之能够覆盖多口音的发音变异。表5中重构的声学模型在粤语口音上的SER下降大于自适应模型的,同时在吴语口音上的SER下降小于自适应模型的,但没有损伤原有对标准普通话的覆盖能力。

综上可见,利用可靠口音相关单元、增量式决策树融合以及选择高支配度的Gauss混合进行声学模型重构的方法,具有在所有测试集上高于混合数据重训练声学模型的系统的性能,并在不降低对普通话识别性能前提下,与自适应后的模型在2种口音上的SER具有可比性。而实现这些性能提升的代价仅是增加了3.6%的Gauss混合数量,合理地保持了重构后声学模型的大小。

5 结 论

本文提出了一种面向多口音语音识别的声学模型重构方法。使用声学似然分作为置信度为每种口音分别生成一套可靠口音相关单元,并把这些单元对应的口音相关模型通过声学模型重构的方法融合到已有的标准普通话模型中。这种方法通过调整标准普通话模型的输出分布的结构,使得输出分布具有足够高的鲁棒性,以覆盖多种口音的发音变异,同时不降低声学模型对普通话的识别性能。为了避免声学模型重构引入冗余的Gauss混合,使用增量式决策树融合以及根据支配度选择Gauss混合的方法改进了原有的声学模型重构方法。实验结果表明:该方法在对标准普通话识别率不下降的前提下,使得系统对带粤、吴2种口音的绝对SER分别显著下降了9.25%和9.21%,而代价仅是增加了3.6%的Gauss混合数目。

参考文献 (References)

[1] Li J, Zheng F, Byrne W, et al. A dialectal Chinese speech recognition framework [J]. *Journal of Computer Science and Technology*, 2006, **21**: 106-115.

[2] Liu Y, Fung P. Multi-accent Chinese speech recognition [C]// Proc of INTERSPEECH 2006. Pittsburg PA, USA: Curran Associates, 2008: 1887-Mon1BuP. 8.

[3] 中国语言文字使用情况调查领导小组办公室. 中国语言文字使用情况 [M]. 北京:语文出版社, 2006.
Leading Group Office of Survey of Language Use in China. Survey of Language Use in China [M]. Beijing, China: Yu Wen Press, 2006. (in Chinese)

[4] Liu L Q, Zheng F, Akabane M. Using a small development data set to build a robust dialectal Chinese speech recognizer [C]//Proc of INTERSPEECH 2007. Antwerp, Belgium: Curran Associates, 2008: 1729-1732.

[5] Riley M., Ljolje A. Automatic generation of detailed pronunciation lexicons [J]. *In Automatic Speech and Speaker Recognition: Advanced Topics*, 1995, **12**: 285-302.

[6] Tomokiyo L M. Recognizing Non-native Speech: Characterizing and Adapting to Non-native Usage in LVCSR [D]. Pittsburg, USA: Carnegie-Mellon University, 2001.

[7] WANG Z, Schultz T, Waibel A. Comparison of acoustic model adaptation techniques on non-native speech [C]//Proc of ICASSP 2003. Hong Kong: IEEE Press, 2003: 540-543.

[8] Oh Y, Kim H. MLLR/MAP adaptation using pronunciation variation for non-native speech recognition [C]//Proc of ASRU 2009. Merano, Italy: IEEE Press, 2009: 216-221.

[9] Saraclar M, Nock H, Khudanpur S. Pronunciation modeling by sharing Gaussian densities across phonetic models [J]. *Computer Speech and Language*, 2000, **14**: 137-160.

[10] 袁家骅. 汉语方言概要(第二版) [M]. 北京:语文出版社, 2001.
YUAN Jiahua. Summary of Chinese Dialects (2nd edition) [M]. Beijing, China: Yu Wen Press, 2001. (in Chinese)

[11] Liu Y. Pronunciation Modeling for Spontaneous Mandarin Speech Recognition [D]. Hong Kong: Hong Kong University of Science and Technology, 2002.

[12] Young S, Evermann G, Gales M, et al. The HTK Book (3.4 edition) [M]. Cambridge, UK: Entropic Cambridge Research Laboratory, 2009.

[13] Chris A. The Long Tail: Why the Future of Business Is Selling Less of More [M]. New York, US: Hyperion, 2006.

[14] ChineseLDC. org. Resource Introduction to RASC863 [EB/OL]. [2009-01-01]. http://www.chineseldc.org/doc/CLDC-SPC-2004-005/intro.htm.