第 36 卷 第 8 期 计 算 机 工 程 2010 年 4 月
Vol.36 No.8 Computer Engineering April 2010
·人工智能及识别技术· 文章编号：1000—3428(2010)08—0170—03 文献标识码：A 中图分类号：TN912.34

母语与非母语语音识别声学建模

曾 定，刘 加
(清华大学电子工程系清华信息科学与技术国家实验室(筹)，北京 100084)

摘 要：为了兼容母语与非母语说话人之间的发音变化，提出一种新的声学模型建模方法。分析中国人受母语影响产生的英语发音变化，利用中国人英语发音数据库自适应得到语音模型，采用声学模型融合技术构建融合 2 种发音规律的识别模型。实验结果证明，中国人英语发音的语音识别率提高了 13.4%，但标准英语的语音识别率仅下降 1.1%。

关键词：语音识别；非母语；模型融合

Native and Non-native Speech Recognition Acoustic Modeling

ZENG Ding, LIU Jia
(Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084)

【Abstract】 In order to tolerance the pronunciation changes between native speakers and non-native speakers, this paper proposes a new modeling method for acoustic model. By analyzing English pronunciation changes caused by Chinese, it uses non-native English pronunciation database to gain the corresponding speech model with adaptive method, and uses acoustic model merging technology to construct a recognition model merged two pronunciation rules. Experimental results show that recognition rate of non-native English increases by 13.4% and recognition rate on native English decreases by 1.1%.

【Key words】 speech recognition; non-native; model merging

1 概述

一个在实验室环境下工作得很好的语音识别系统在实际应用时性能会变得不够稳健，其中一个重要的原因就是发音的母语口音问题。以英语为例，多数中国人的英语发音并不像母语是英语的人那样标准、清晰，而是受其母语的影响很大，带有各种地方口音。目前语音识别技术主要基于统计模式识别理论[1]。因此，用于模型训练的语音数据库对模型性能会产生很大的影响。提高非母语说话人语音识别率最直接的方法是用非母语说话人的语音训练产生识别系统，但存在的问题是难以获得大量非母语说话人的训练语音数据，并且用非母语发音库训练的模型会降低母语发音说话人语音识别率。另一种方法是运用模型自适应方法，例如 MLLR 和 MAP[2]，这种方法同样会在一定程度上提高非母语说话人的语音识别率，但会降低母语为英语的说话人的语音识别率。在实际应用场合中，一个英语识别系统最好能兼顾不同口音的英语发音，表现出稳健性。本文先以母语为英语的大规模语音库为基础，建立标准英语发音的语音识别模型，然后针对小规模中国人英语发音的语音库，考虑中国人受母语影响带来的英语发音变化，训练出能适应中国人英语发音的语音识别模型，在这基础上采用模型融合技术，目标是保证高性能的母语说话人的英语语音识别率，同时提高母语非英语的说话人语音识别性能。

2 非母语声学模型构建

母语是人们进行思考时头脑中下意识反应出来的语音发音，因此，再去学习一种新的语言时，一般很难达到母语级别的掌握程度。母语和非母语说话人之间会因发音变化带来不匹配，这就需要建立新的模型来容忍这种变化。非母语带来的影响可以在非母语和母语中体现出来。文献[3]通过建立法语和英语之间的发音变化对应关系，在英语模型的基础上，利用法语语音数据做自适应，得到一个新的模型，然后把该模型插入英语模型中，从而提高母语是法语的说话人的英语识别率。文献[4]采用类似的方法提高了母语是英语的说话人阿拉伯语的识别率。本文分析了中国人受母语影响带来的英语发音变化，然后利用中国人英语发音数据库训练得到适合中国人的英语发音识别模型。

2.1 中国人的英语发音特点

受母语的影响，语音发音变化常表现出 2 种情况：

(1)音素的变化，如音素的替代、插入、删除错误。对于此类发音变化，一般采用字典自适应的方法进行处理，即在发音字典中加入反映说话人发音特点的实际发音。

(2)声音的变化。某些实际发音的变化由于不具有类似于音素变化中 $A \rightarrow B$ 的显著特征，无法通过发音字典来解决，因此需要改进语音模型，使其能容忍非母语说话人的发音变化。

当前获取发音变化的主要途径有：基于专家知识和基于数据驱动。基于专家知识的方法是依赖专家的先验知识分析

基金项目：国家自然科学基金委员会与微软亚洲研究院联合基金资助项目(60776800)；国家“863”计划基金资助项目(2006AA010101, 2007AA04Z223, 2008AA02Z414)

作者简介：曾 定(1984－)，男，硕士研究生，主研方向：语音识别；刘 加，教授

收稿日期：2009-11-20 **E-mail：**cengd06@mails.tsinghua.edu.cn

母语和目标语言之间的异同以获取音素替代关系。这些知识或者规则是语音专家经过长期观察得到的，因此，受限于专家人数和研究时间，要得到客观准确的语音学知识比较困难，优点是不需要非母语说话人的语音数据。数据驱动可采用模型距离矩阵和统计音素混淆矩阵来获取母语和非母语之间的音素替代关系。

本文采用两者相结合的方法。首先分析受汉语影响造成的普遍性的发音变化，得出中国人经常发错音的音素。通过这些专家知识的分析，得到一个易混淆音素列表。基于数据驱动的方法则利用了音素混淆矩阵[5]。在基于数据驱动的发音字典自适应过程中会产生大量可能的发音变化。发音变化趋势不够集中主要是由基于音素识别的解码过程中的识别器错误以及非母语与母语说话人的发音“偏差”造成的。针对此问题，本文采用基于概率值的剪枝策略，选择最具代表性的发音变化。结合这 2 种方法，分析得到中国人英语发音的易混淆音素对及其混淆概率，如表 1 所示。其中，概率值根据混淆矩阵得到，等于音素 A 误识为音素 B 的个数与 A 的总数之比。从表 1 可以看到，汉语中不存在的音素表现出很高的混淆概率。

表 1　中国人英语发音的易混淆音素对及混淆概率

易混淆音素对		混淆概率	易混淆音素对		混淆概率
/ʒ/	/ʃ/	0.48	/e/	/æ/	0.28
/θ/	/s/	0.35	/æ/	/ai/	0.22
/i/	/i:/	0.34	/au/	/□/	0.20
/v/	/f/	0.33	/u/	/əu/	0.18
/u:/	/u/	0.28	/ŋ/	/n/	0.17
/p/	/pə/	0.22	/b/	/bə/	0.24

2.2　非母语模型构建

母语和非母语说话人之间会由于发音变化带来不匹配，因此，需要建立新的模型来容忍这种变化。由于笔者已采集到约 30 h 的中国人英语发音数据库，因此先利用中国人英语发音数据库来训练简化的非母语模型 M3。在建模过程中，根据表 1 中的混淆音素在训练过程中加入多发音字典，在识别时同样对词条网络采用多发音字典。实验表明该方法能在一定程度上提高中国人英语发音的识别率。同时，利用非母语数据库对标准的英语模型 M1 自适应得到一个模型 M2。模型 M2 与 M1 具有相同的聚类结构，可以很方便地对 2 个模型做融合。

3　模型融合

由于上文建立的模型 M2 和 M3 都会在提高非母语说话人语音识别率的同时牺牲母语语音的识别率，因此本文采取模型融合技术，使得提高中国人英语发音识别率的同时兼顾母语英语的识别率。

3.1　融合模型的选择

在基于音素的英语语音识别中，将来自于母语英语的模型与来自于非母语背景的模型在状态相同的前提下依据一定的准则进行归并，归并后的状态包含来自于母语英语和非母语背景英语的高斯混合，增加标准母语模型的空间覆盖度。对于非母语的模型有 2 个选择：上下文无关模型以及与母语模型具有相同聚类结果的上下文相关模型。由于 M2 与 M1 具有相同的聚类结构，可以很方便地对 2 个模型做融合，且自适应得到的模型对非母语识别性能比上下文无关模型好，因此本文选择自适应得到的非母语模型参与融合。

3.2　模型直接插值法

模型插值法是选择一个插值因子，把英语模型 M1 和自适应得到的模型 M2 在状态级相加。在隐马尔科夫模型(Hidden Markov Model, HMM)中，一个状态的输出概率密度函数是通过多个高斯混合分布来描述的，若用 $\boldsymbol{x}$ 表示输入特征向量，s_i 表示第 i 个状态，则 $p(\boldsymbol{x}\mid s_i)$ 可以表示为

$$p(\boldsymbol{x}\mid s_i)=\sum_{k=1}^{K} w_{ik}N(\mu_{ik};\sigma_{ik}) \tag{1}$$

其中，K 表示状态中包含的高斯混合的数目；w_{ik} 表示第 i 个状态的第 k 个高斯混合的权重。2 个模型归并后第 i 个状态的 $p(\boldsymbol{x}\mid s_i)$ 可表示为

$$p(\boldsymbol{x}\mid s_i)=\lambda p(\boldsymbol{x}\mid s_i^{ne})+(1-\lambda)p(\boldsymbol{x}\mid s_{im}^{nne}) \tag{2}$$

其中，s_i^{ne} 和 s_{im}^{nne} 分别表示来自于 M1 和 M2 的第 i 个状态；λ 表示插值因子，其值依赖具体实验数据，可以通过数据驱动的方法获得。在融合得到的模型中，其高斯分量既保持了母语发音的特点，又加入了非母语发音的特点，从而提高了中国人英语发音的识别率，同时兼顾了母语英语的识别率。

3.3　可选择模型归并

模型直接插值法是直接将 2 个模型合并，没有考虑母语和非母语发音变化，对所有音素同等对待，同时会带来高斯数量的增加。基于以上 2 个因素，由模型插值法衍生出模型选择归并法，其核心是考虑母语和非母语发音变化，参与归并的音素除了与自身相同的音素外，还引入发音变化音素，同时对于与自身相同的音素，在状态归并参数共享的过程中，依据一定的策略将需要参与归并的状态和不需要参与归并的状态(冗余的)区分开，突出两者之间的差异，提高状态归并时的针对性，从而达到降低模型规模的目的。考虑到可选择模型归并是为了增加标准母语模型的空间覆盖度，同时又不引入模型的冗余度，本文采用距离度量的策略实现可选择模型归并。因为在状态归并中希望突出两者之间的差异，所以距离度量策略采用非对称的马氏距离，用非母语语音模型的协方差矩阵计算距离，强调母语和非母语语音状态之间的差异性。马氏距离的定义如下：

$$D(A,B)=\sum_{i=1}^{M} w_{Ai}\left[\sum_{j=1}^{M} w_{Bj}d_{A,B}(i,j)\right] \tag{3}$$

$$d(i,j)=(\mu_i-\mu_j)^{\mathrm{T}}\Sigma_j^{-1}(\mu_i-\mu_j) \tag{4}$$

在式(3)中，A 和 B 分别代表来自于标准英语模型 M1 和自适应得到模型 M2 的状态，它们包含 M 个高斯混合分量。式(4)表示非对称的马氏距离，用来计算任意 2 个高斯分量的距离。

可选择模型归并采用基于距离值的剪枝策略，当 M1 和 M2 的状态距离大于阈值时，需要参与状态归并，否则，不参与状态归并。阈值的大小是归并模型的规模和识别率的一个折中值。选择模型归并后，第 i 个状态的 $p(\boldsymbol{x}\mid s_i)$ 可表示为

$$p(\boldsymbol{x}\mid s_i)=\lambda p(\boldsymbol{x}\mid s_i^{ne})+\sum_{m=1}^{M}(1-\lambda)p(\boldsymbol{x}\mid s_{im}^{nne})p(s_{im}^{nne}\mid s_i^{ne}) \tag{5}$$

式(5)中引进了发音变化概率 $p(s_{im}^{nne}\mid s_i^{ne})$ 和可选择的音素数 M。发音变化概率由提取发音变化时得到的音素混淆矩阵获得。音素数 M 代表参与融合的音素个数，可能的取值是 0, 1, 2。

4　实验

4.1　实验数据

实验中数据库分为英语母语数据库(WSJ)和中国人英语

发音数据库(NOSE)。

(1)英语母语数据库：采用 WSJ1 中的 SI_TR_S 作为训练库，包括 200 人的连续语音，共 61 h。词条测试集为由 WSJ1 测试集 CDTest 得到的 1 400 句短句(每句包含 1 个或 2 个单词)。从测试集中随机进行抽取，得到 3 个 1 000 句短句测试集。

(2)中国人英语发音数据库：包括 100 人(50 男 50 女)的连续语音，每人 150 句句子，大约 30 h。其中，每个人的 140 句用于训练数据库；10 句用于做音素识别测试。另一个测试集为 8 人的命令词短句(每句包含 1 个或 2 个单词)，每人 1 000 句短句。

4.2 模型测试结果及分析

利用训练库 WSJ 得到英语标准模型 M1，用 NOSE 数据库对英语标准模型做自适应(MLLR+MAP)[2]得到模型 M2，单独利用 NOSE 训练库得到模型 M3。将 M1 和 M2 直接插值[3-4]得到模型 M4，λ 取值为 0.7。采用可选择模型归并得到模型 M5，模型归并时选择距离阈值为最大距离的 0.2 倍。实验是针对 1 000 个命令词做识别，这 5 个模型对母语和非母语词条的测试结果如表 2 所示。

表 2 母语和非母语模型识别率对比 (%)

	M1	M2	M3	M4	M5
母语英语	98.65	93.58	87.09	97.61	97.52
中国人英语发音	79.76	95.08	95.90	92.78	92.99

从实验结果可以看到，相对于标准英语语音模型 M1，其他 4 种建模方法都能提高中国人英语发音的识别率，但 M3 是单独由非母语数据库训练得到的，所以，对母语英语的识别率下降很多。模型插值和可选择模型归并都是在标准英语语音模型基础上借助于非母语模型，涵盖更多的发音变化，使得融合后的模型有了更大的声学空间覆盖度，从而可以兼容母语和非母语的识别率。采用模型插值技术对中国人英语发音的识别率提高了 13%，而对英语母语的识别率只下降了 1%。可选择模型归并得到的模型相比模型插值，增加了中国人英语发音的发音变化，因此，模型有了更好的声学空间覆盖度，对中国人英语发音的识别率上升了 0.2%，同时考虑了减少引入模型的冗余度，一定程度上达到了精简模型的目的。

在整个实验过程中，一直以音素识别的混淆矩阵作为数据驱动。在此，分析各个模型对中国人英语发音连续语音进行音素识别的结果。因为整个音素混淆矩阵比较大，且实验过程中只关注那些识别率低的易混淆音素，所以表 3 中只给出 3 个语音模型对易混淆音素的单个音素识别结果。

表 3 易混淆音素的单个音素识别率对比 (%)

音素	M1	M4	M5	音素	M1	M4	M5
/ʒ/	20.0	34.8	43.8	/e/	40.5	53.3	53.2
/θ/	37.3	60.4	58.8	/æ/	58.9	60.4	60.6
/i/	43.9	58.5	61.2	/au/	37.6	50.2	50.6
/v/	41.4	55.4	59.1	/u/	21.3	38.6	47.2
/u:/	58.8	73.7	74.1	/ŋ/	61.1	67.5	68.2

从表 3 可以看出，直接用母语英语训练得到的模型对中国人英语发音进行音素识别的识别率很低，正是这些音素的识别率低才导致短语词条的识别率下降。经过模型融合得到的模型 M4 和 M5 的音素识别率都有很大的提高，由于 M5 进一步考虑了发音变化，因此在易混淆音素上表现出更好的效果。由于表 3 是易混淆音素的识别结果，因此单个音素的识别率都偏低，其中某些错误是声音变化导致的，通过采用多发音字典可以进一步改善其识别率。

4.3 多发音字典与模型融合的结合

在声学层面，非母语说话人与母语说话人相比会出现音素变化和声音变化。模型融合方法解决了声音变化问题。多发音字典可以有效解决确定性的音素变化问题[1]。本文将模型融合与多发音字典相结合得到模型 M6。由实验结果发现，增加多发音字典后，对中国人英语发音的识别率从 92.99%提高到 93.2%，而对母语英语的识别率基本没影响。图 1 为 6 个模型的识别结果。从中可以看出，M6 模型综合考虑了母语英语与中国人英语发音之间的音素变化和声音变化，因此，对中国人英语发音的识别性能最好，且对母语英语保持了较高的识别率。

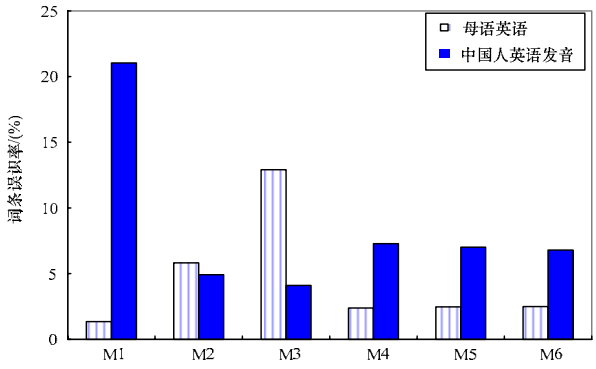

图 1 6 个模型识别错误率比较

5 结束语

为解决非母语说话人识别率低的问题，本文以中国人英语发音为实验背景，利用中国人英语发音数据库分析中国人英语发音的发音变化，构建适应中国人英语发音的语音模型，然后采用模型融合技术提高了中国人英语发音的识别率，同时兼顾了母语英语的识别率，在此基础上考虑多发音字典，将对中国人英语发音的识别率提高到 93.2%，而母语英语的识别率保持在 97.48%，有效扩大了基于命令词的语音识别的应用范围。但非母语的识别率相对于母语的识别率较低。这需要在之后的工作中进一步改进。

参考文献

[1] 刘林泉. 基于小数据量的方言背景普通话语音识别声学建模研究[D]. 北京: 清华大学计算机科学与技术系, 2007.

[2] 丰洪才, 卢正鼎. 基于 MAP 和 MLLR 的综合渐进自适应方法研究[J]. 计算机工程, 2005, 31(5): 4-7.

[3] Bouselmi G, Fohr D, Illina I, et al. Fully Automated Non-native Speech Recognition Using Confusion-based Acoustic Model Integration[C]//Proc. of the 9th European Conference on Speech Communication and Technology. Lisbon, Portugal: [s. n.], 2005.

[4] Tan Tien-Ping, Besacier L. Acoustic Model Interpolation for Non-native Speech Recognition[C]//Proc. of ICASSP'07. Honolulu, USA: [s. n.], 2007.

[5] Decker A M, Lamel L. Pronunciation Variants Across System Configuration, Language and Speaking Style[J]. Speech Communication, 1999, 29(2-4): 83-98.

编辑 张 帆