面向情感语音识别的建模方法研究*)

潘玉春　徐明星　贾培发

(清华大学信息科学技术国家实验室语音技术中心　北京 100084)

摘　要　情感是语音识别研究中一个不可避免的问题，不同的情感对于语音有着不同的影响，这种影响使得中性语音识别系统在实际应用中的识别效果大打折扣。对于类似的影响通常的解决方法有寻找鲁棒特征，特征归一化以及模型调整训练等。本文通过自适应方法，使用少量情感数据，在中性语音模型的基础上自适应得到新的情感语音模型。实验证明，新模型对于情感语音有着更好的整体识别率。

关键词　语音识别，情感语音识别，情感计算，自适应

Research on Modeling Method of Emotional Speech Recognition

PAN Yu-Chun　XU Ming-Xing　JIA Pei-Fa

(Center for Speech Technology, The State Key Laboratory of Intelligence Technology and System, Department of Computer Science and Technology, Tsinghua University, Beijing 100084)

Abstract　As known to all, emotion plays a significant role in speech recognition. The model built on neutral speech degrades dramatically while recognizing speech with emotion. How to deal with emotion issue properly is crucial to achieve good performance in recognition. Most widely used approaches include robust feature extraction, speaker normalization and model tuning/retraining. In the study, a novel method is proposed, that is, adaptation technique is adopted to transform a neutral-based model into emotion-specific one with a small amount of emotion speech. It's shown experimentally that the new model achieves higher accuracy in overall performance.

Keywords　Speech recognition, Emotional speech recognition, Affective computing, Adaptation

1　引言

语音识别的研究工作大约开始于20世纪50年代，标志就是AT&T BELL实验室实现了第一个可识别十个英文数字的语音识别系统——Audry系统[1]。经过半个多世纪，从最初的孤立字(词)、特定人以及小词汇量的语音识别系统到现在的连续语音、非特定人的无穷词汇量语音识别系统，语音识别研究工作在多个方面取得了重大突破和发展。但是，由于研究中所使用的语音大多是在比较理想条件下采集的，而在实际应用中，由于受信道、环境噪音、发音方式、情感状态等因素的影响，语音识别效果有很大的降低，如何解决这些问题成为现在语音研究的热点[2~4]。对于跨信道、去噪、发音方式等问题已经有一定的研究基础，对于情感的研究却比较少。

随着人机交互系统的快速发展，有关"情感"以及"情感计算"的研究越来越受到人们的重视[5,6]，并且已经在人脸表情、姿态分析等方面获得了一定的进展[7]。语音作为人类交流的重要手段，是互相传递信息最直接、最方便且最有效的途径，和人的面部表情一样，语音同样传递着人的情感信息。因此语音识别的最终任务应该是既要识别出文字信息也要识别出语音中携带的情感信息。现阶段对于情感语音的研究无论是国内还是国外都还处于一个起步阶段[6,8]，考虑到情感和态度对语音所引起的变化对语音合成、语音识别、说话人识别的影响较大，语音的情感研究逐渐引起人们的重视[8]。从总体上看，研究工作主要集中在情感分析、情感识别和情感语音合成[9~12]等几个方面，而关于情感语音的识别研究却不多见。

文[2]指出，在背景噪音、传输信道、心理紧张及工作压力和情绪变化等因素中，情感对语音识别率的影响最大。对于这种变异语音识别问题国外在70年代末就开始有了相应的研究，解决方法按照从底层到高层通常可以概括为3类：(1)特征级。这类方法的核心思想在于寻找到既不受各种变异因素影响，又能表示语音内容的鲁棒特征，或者在识别阶段加一个变异规整过程来消除变异情况对语音特征的影响，使得规整后的语音和正常语音特征尽量接近，从而用正常语音训练的识别器仍可以获得很好的识别效果。Hansen等提出了对共振峰带宽和共振峰位置进行补偿的方法[4]；(2)声学模型级。这类方法根据语音变化的特点在特征和模型训练方法上作一些调整，比如Lippman等提出的Multi-style训练法[13]。本文所用的声学模型自适应方法也属于这一类；(3)语言模型级。利用高层知识在语言模型上作的一些调整，T. Athanaselis通过在语言模型中加入情感句子的比例，提高了情感语音的识别率[14]。

对于每种情感语音单独建模可能会大大提高情感语音的识别率，但是这在实际操作中是不实际的，情感语音的采集较中性朗读式语音困难得多，对发音者的要求很高，因此很难在短期内采得大量的数据。本文以几种基本情感为例，研究它们对语音识别的影响以及通过自适应的方法，使用少量情感数据，在中性语音模型的基础上自适应得到情感语音模型，从而提高情感语音的整体识别率。

本文的结构组织如下：第2部分介绍情感数据库以及数据的使用；第3、4部分分别介绍基线系统和自适应系统以及

*)国家自然科学基金重点项目—情感计算理论与方法研究(编号：60433030，分类代码：F020106)。潘玉春　硕士研究生。

对实验结果的分析；最后是研究工作总结和下一步的研究方向。

2 情感语音数据库

2.1 语音数据的类别

情感的分类是一个复杂的问题，有很多学者对此进行了研究[15,16]，但并没有达成一个一致的标准。现阶段对情感语音的研究通常结合实际情况和自己的理解，确定几种基本情感。在我们的研究中，采用了中性(Neutral)、高兴(Happiness)、生气(Anger)、恐惧(Fear)、悲伤(Sadness)等五种情感类型。为了后面行文方便，我们用 N 表示中性，用 H 表示高兴，用 A 表示生气，F 表示恐惧，S 表示悲伤。

2.2 数据库描述

本实验录制了一个 200 词项的情感语音库，所用词表基本覆盖全常用声韵母[17]。在录音中要求发音人用五种基本情感把每个词项都各读一遍。所有发音人员都是从在读大学生中挑选的，具有一定的语言表现力，共 50 人，其中 25 个男生、25 个女生。录音环境是普通的办公房间，录音时比较安静。

随机选出 10 男生和 10 女生的全部语音作为测试语音，剩下 30 人语音作为训练语音。表 1 给出了情感数据库组成情况。

表 1 情感数据库组成

语音 \ 情感	N	F	S	A	H
训练 (15 男+15 女)	6000 句	6000 句	6000 句	6000 句	6000 句
测试 (10 男+10 女)	4000 句	4000 句	4000 句	4000 句	4000 句

3 基线系统

本文采用扩展的上下文相关声韵母进行声学模型建模(Tri-XIF)[17]。考虑到录音所用语料的限制，个别声学基元没有在语料中出现，因此实验中对声学基元集做了一些调整。声学模型为 HMM，每个声学基元有 3 个状态，每个状态用 4 个高斯混合描述。模型训练采用混合分裂的方式进行。通过使用基于决策树的状态共享策略，将任一模型的总状态数控制在 2000 左右。模型所用的特征参数为 39 维 MFCC，包含能量参数，以及一阶和二阶差分。实验中，分别使用中性、恐惧、悲伤、愤怒和高兴的情感语音单独训练五个不同的声学模型，即一个中性语音模型和四个情感语音模型。模型性能用音节识别率进行评价。图 1 给出了中性语音模型和专有情感模型测试结果对比。

从图中我们可以看出：

(1)情感造成的语音变异对语音识别率有很大的影响。

如果使用中性语音模型来测试各种不同情感的语音数据，虽然中性语音可以达到 90.83%的准确率，但是对情感语音，识别准确率都有不同程度的下降。其中，测试生气语音得到的性能最低。这些结果说明，情感对语音识别性能的确有很大影响，中性语音模型对情感语音的识别性能效果很差。由于在实际应用中情感语音是经常碰到的，因此情感语音识别的研究对于实际应用有很大的意义。

(2)情感语音数据训练的声学模型能够有效提高对情感语音的识别率。

与中性语音模型相比，情感语音训练的声学模型对同种情感待测语音的识别率有很大的提高。但是由于对每种情感建模需要采集大量的情感语音数据，并且随着情感类别的增加数据量也急剧增加；同时，根据下文表 2 给出的不同情感声学模型对不同情感语音数据的测试结果，如果可以事先知道待测语音的情感并选用相应的情感声学模型，则对于情感语音的整体识别率为 82.56%，但是相对于中性语音 90.83%的准确率还是有较大的差距。因此，采集大量的情感数据单独训练模型，不仅成本高，而且也不能保证有很好的识别准确率。

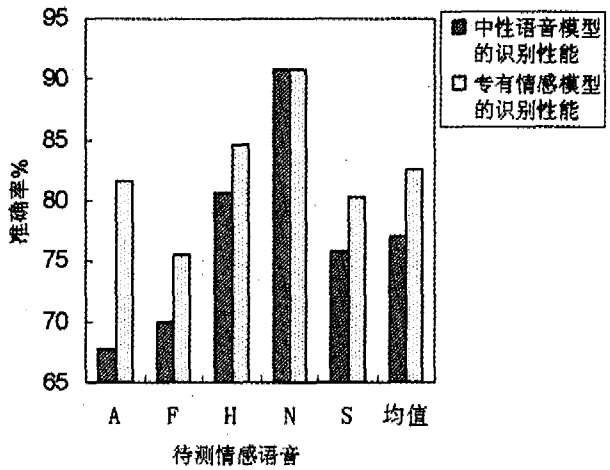

图 1 中性语音模型和专有情感模型测试结果对比

4 情感声学模型的自适应

声学模型自适应技术在语音识别领域得到了广泛的应用[1]。通过使用少量自适应数据对声学模型参数进行调整，识别系统能够更好地匹配由于麦克风、传输信道、环境噪音、说话人等引起的差异[8,18]。为了解决语音情感带来的语音变化，本文将探讨如何将声学模型自适应技术应用到情感语音的识别中来，寻找减小情感对于语音识别影响的途径。

目前，常用的声学自适应技术是模型参数转换自适应法，它主要是针对 HMM 模型的参数作自适应转换，如最大后验概率(MAP)方法和最大似然线性回归(MLLR)方法。考虑到对自适应数据量的需求和自适应速度，本文选用了自适应速度快、自适应数据量需求少的 MLLR 方法[17]。该方法通过一个线性变换把初始模型变化到新的自适应模型上，利用 Baum-Welch 最大似然原理重估出线性变换的转移矩阵。

在以下的声学模型自适应实验中，原始的声学模型是用中性语音数据训练得到的，就是基线系统中的中性语音声学模型，训练过程使用了 6000 句中性语音数据。

4.1 分情感自适应

为了提高中性语音声学模型对情感语音的识别率，我们分别使用少量的情感语音数据，对中性语音声学模型进行了 MLLR 自适应，得到四个情感声学模型。对每种情感，使用的自适应数据量是相同的，都是 1500 句。

下表给出了不同训练方式得到的情感声学模型对各种情感语音数据的交叉测试结果。

从表 2 可以看出，自适应得到的情感声学模型对测试数据的平均识别率要高于只用情感语音训练得到的声学模型的平均识别率，平均误识率下降 8.69%。这是因为自适应声学模型实际涉及到的训练数据并不是仅有那些用于自适应的数据，还有训练中性声学模型的数据。

表 2 情感声学模型对各种情感语音的交叉测试结果

EM / ET	A		F		H		N	S		均值	
	A_mod	A_Adapt	F_mod	F_Adapt	H_mod	H_Adapt	N_mod	S_mod	S_Adapt	X_mod	X_Adapt
A	**81.67**	**74.45**	70.93	68.22	74.96	69.55	67.67	63.47	65.54	71.74	69.09
F	61.24	67.90	**75.45**	**73.18**	70.12	69.22	69.94	72.55	71.69	69.86	70.39
H	71.82	79.66	79.76	81.16	**84.59**	**83.04**	80.66	77.33	79.12	78.83	80.73
N	66.48	83.77	79.64	88.20	81.37	86.92	**90.83**	86.60	89.18	80.98	87.78
S	50.89	68.37	71.99	76.31	67.13	70.91	75.82	**80.24**	**78.04**	69.21	73.89
均值	66.42	74.83	75.55	77.41	75.63	75.93	76.98	76.04	76.71	**74.13**	**76.37**
E RR	25.04		7.61		1.23		—	2.80		8.69	

X_mod 表示单一情感语音模型，X_Adapt 表示自适应情感语音模型，EM 表示不同的情感语音模型；ET 表示不同情感的待测语音。ERR (Error Rate Reduction)为自适应情感语音模型相对于单一情感语音模型对整体语音平均误识率下降的百分比。

另外，无论是只用情感语音训练的专有情感声学模型还是自适应得到的情感声学模型，它们虽然都提高了对于对应情感语音的识别率，但却同时增加了对其他不同情感语音的误识率。可见，不同情感声学模型之间的兼容性比较差，需要根据情感语音的类别选择对应的情感声学模型。

4.2 混合数据自适应

分不同情感对中性声学模型进行自适应，虽然可以提高对相应情感语音的测试性能，但最终并没有提高对全部数据的平均识别率。这是因为不同情感对语音造成的变化是不一样的，只用一种情感语音进行自适应得到的模型，对于其他不同情感类别的语音，不能有效刻画，从而在整体上并不能提高语音识别的效果。

为了使自适应得到的声学模型对不同情感语音都能有效描述，我们把 4 种不同情感的自适应语音数据混合起来，组成新的自适应数据集。表 3 给出了中性语音模型和混合自适应模型测试的结果，以及对于相应的待测情感语音使用相应的自适应模型测试的结果。

表 3 声学模型的性能对比

测试语音的情感种类	A	F	H	N	S	均值
N_mod	67.67	69.94	80.66	90.83	75.82	76.98
X_Adapt	74.45	73.18	83.04	90.83	78.04	79.91
Mix_Adapt	71.24	73.12	83.32	89.45	76.16	78.66

N_mod 表示中性语音模型，Mix_Adapt 表示混合数据自适应模型，X_Adapt 表示对于某种待测情感语音用相应的情感数据自适应模型测试。

由于使用了多种情感的语音，使得不同情感对于语音造成的变化都能在自适应数据中得到体现。从表 3 中可以看出，虽然混合自适应模型对于中性语音的识别性能有一些下降，但对于其他情感语音，识别率都有较明显的提高，而且对全部测试数据的平均识别率相对于中性语音模型也有提高，平均误识率下降了 7.30%。虽然相对于专有情感模型的平均识别率 79.91%还是有一点差距，但是后者需要综合多个情感自适应模型的结果，并且要求待测语音能够进行正确的分类。

总结与展望 本文研究了面向情感语音识别的声学建模方法，探讨了声学模型自适应方法的应用方式。本文实验结果表明：

(1)对于任一种情感语音，虽然用同种情感声学模型测试的结果好于用其他情感声学模型测试的结果，但是跟中性语音模型测中性语音的识别结果相比，还有很大差距。这说明目前在实验中所使用的声学特征参数(MFCC)并不能很好地刻画情感语音，还需要改进；

(2)用少量情感数据对中性语音声学模型进行自适应得到的模型，可以有效地提高对相应的情感语音的识别率。使用混合自适应模型，可以提高对情感语音的整体识别率。

由于情感语音的研究还处于起步阶段，还有很多问题亟待解决。如何利用情感信息，综合不同模型的识别结果以提高整体语音识别结果，或从特征级去除情感对语音的影响，寻找对情感变化更鲁棒的特征，这些都是以后值得深入研究的地方。

参 考 文 献

1 王昱. 语音识别自适应技术的研究与实现:[硕士论文]. 清华大学,2000

2 Hansen J H L, Bou-Ghazale S E. Getting started with SUSAS: A speech under simulated and actual stress database. In: EUROSPEECH-97: Eur Conf Speech Communication Technology. vol 4. Rhodes, Greece, Sept. 1997

3 Bou-Ghazale S E, Hansen J H L. A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech Under Stress. IEEE Transactions on Speech & Audio Processing, 2000

4 Hansen J H L. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. Speech Communication , 1996

5 胡包钢,谭铁牛. 情感计算——计算机科技发展的新课题. 科学时报,2000-03-24(3)

6 Picard R W. Affective Computing. Cambridge, Massachusetts, London, England, The MIT Press, 1998

7 陶建华,谭铁牛. 数字化人类情感——和谐人机交互环境中的情感计算. 微电脑世界,2004(1)

8 韩纪庆,邵艳秋. 基于语音信号的情感处理研究进展. 中国科技论文在线,2005

9 赵力, 钱向民. 语音信号中的情感特征分析和识别的研究. 通信学报,2000

10 Tao Jianhua, Kang Yongguo. Features Importance Analysis for Emotional Speech Classification. In: Proceeding of ACII, 2005

11 New T L. Speech emotion recognition using Hidden Markov models. Speech Communication,2003, 41:603～623

12 Jang D N, Zhang W, et al. Prosody Analysis and Modeling for Emotional Speech Synthesis. ICASSP,2005

13 Lippmann R P , Martin E A , Paul D B. Multi-style training for robust isolated-word speech recognition [A] . In: ICASSP'87 [C] . USA: IEEE Press ,1987. 705～708

14 Athanaselis T. ASR for emotional speech: Clarifying the issues and enhancing performance. Neutral Networks, 2005. 18

15 Jiang Dan-Ning, Cai Lian-Hong. Classifying Emotion in Chinese Speech by Decomposing Prosodic Features. INTERSPEECH 2004 - ICSLP, Oct. 2004

16 Nicholson J, Takahash K, Nakatsu R. Emotion Recognition in Speech Using Neural Networks. Neural Computing and Applications, 2000

17 李净,郑方. 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

18 Leggetter C J, Woodland P C. Speaker Adaptation of HMMs Using Linear Regression:[Technical Report]. CUED/F-INFENG/TR181. Cambridge Univ, Jun. 1994