变异语音处理的研究进展

张 磊,韩纪庆,王承发
(哈尔滨工业大学计算机科学与工程系,黑龙江哈尔滨 150001)

摘 要: 本文讨论了变异语音处理技术及其研究进展,分析了变异情况对语音识别性能产生的影响,综述了变异语音分类和变异语音识别方法,探讨了变异语音处理研究中存在的问题及未来的研究重点.

关键词: 变异语音; 语音分析; 语音分类; 顽健语音识别
中图分类号: TN912.3 **文献标识码**: A **文章编号**: 0372-2112 (2003) 03-0411-08

Research Progress of Stressed Speech Processing

ZHANG Lei, HAN Ji-qing, WANG Cheng-fa
(Department of Computer Science and Engineering, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China)

Abstract: This paper firstly discusses the technique of stressed speech processing and its recent progress, and analyses the reasons for the degradation of recognition of speech under stress. Then the classification and the recognition methods of stressed speech are reviewed. Finally, the unsolved problems and the prospect in this field are pointed out.

Key words: stressed speech; speech analysis; speech classification; robust speech recognition

1 引言

语音识别研究已经取得了很大进展. 随着语音识别系统向实用化方向发展,影响语音识别算法性能的环境因素就显得日益突出. 要使语音识别系统真正实用化,就必须克服环境因素的影响.

影响语音识别的因素很多,主要包括背景噪声、传输通道变化、心理紧张及工作压力和情绪变化等所产生的发音变异,这些因素共同构成了影响识别系统顽健性(Robustness)的诸要素. 正是由于这些干扰因素的影响,语音识别系统离真正实用化还有一段距离.

对背景噪声和传输通道变化影响进行补偿的语音识别方法已经被广泛研究,但对变异语音识别的研究工作目前相对来说进行得还较少. 发音变异涉及范围较广泛,包括连续语音中上下文不同产生语音变化,不同人之间发音变化及同一个人在不同环境下发音变化等. 本文着重讨论当周围环境或话者自身条件发生异常变化时,话者产生的语音变异. 图 1 左半部分给出了引起语音变异的主要因素[1],这些变异情况可能发生在行驶的汽车中、直升飞机中、战斗机或噪音较大的工厂环境. 在这些应用中,由于使用者大都专注于某项工作,语音识别只是辅助于其它工作的次要工作,这时由于工作压力的存在,对话者发音将会有较大的影响. 同时,由于有背景噪声存在,话者会试图调整发音方式以达到更清晰的表达效果,这就是著名的 Lombard 效应[2]. Lombard 效应影响的大小依赖于背景噪声的强弱和类型. 在很多情况下,话者也会由于一些情绪干扰而产生发音方式的变化,如在愤怒、悲伤、高兴、害怕时. 此外,话者说话方式的快、慢等也会在一定程度上影响语音识别系统性能. 当话者身体受到一些物理冲击时,发音也会产生一定程度的变异. 根据话者不同部位受到影响程度的不同,可以将语音变异分为物理层变异、生理层变异、感知层变异、心理层变异等[3].

在大多数情况下,人类听觉系统能够在发音有变异情况下正确地分辨出语音信息内容,并且可以捕获到额外的反映心理紧张和情绪变化方面信息,通常语音识别算法并不能做到这一点,因而导致识别性能下降. 背景噪声和通道畸变的影响,一般可以认为是均匀地作用在整个发音之上,而发音变异是在原有正常发音基础上,某些音素或音素的某些地方发生畸变. 由于在一个发音中各个音素所受到的影响不相同,很难简单地用模型进行刻画. 同时相同的外界影响因素对不同人的影响效果也是不尽相同的,这一切都增加了变异语音识别的难度.

尽管变异语音处理难度很大,但研究者仍然做了大量的工作,取得了相应的研究成果. 本文将从变异语音分析、变异语音分类以及变异语音识别等方面综述变异语音处理的研究进展,并展望其未来的发展方向.

2 国内外研究现状

早在 1911 年,E Lombard 就发现了话者在背景噪声下会

收稿日期:2001-08-10;修回日期:2002-08-06
基金项目:国家自然科学基金(No. 60085001);教育部留学回国人员科研启动基金

努力调整自己的发音方式,以提高说话的清晰度,这就是后来被研究者称作的 Lombard 效应现象. 在此之后,美国海军航空科研中心研究分析变异条件下人的生理和心理受到的影响,如对 G-Force 下飞机驾驶员的跟踪能力、血流反映模型、心理反应及人的承受能力的分析. 但直到七十年代末才开始有人系统地对变异语音展开相应的研究. 美国空军以 Wright 航空实验室和 Armstrong 航空医学研究室为中心,开展关于战斗机中语音识别研究工作,其中 Armstrong 航空医学研究室和美国麻省理工学院的 Lincoln 实验室还分别参与了美国国防部高级研究计划局 DARPA 顽健语音识别项目. 另外,CMU 大学、BBN 实验室及 Texas Instruments 等也都开展了顽健语音识别的研究. 最初的研究方向只是停留在对各种变异情况下语音的可懂度、各种情况下的变异语音分析,以及发音变异对语音识别系统性能的影响问题上. 研究发现:在变异情况下,语音的可懂度降低[4],并且在 G-Force 变异中,头罩对识别性能的影响很大[5],在其它变异中,供氧系统和呼吸噪音也是影响系统性能的主要原因[6]. 对变异语音数据分析发现,通过对语音的持续时间、强度、声门参数的分析可以测出变异是否存在[7].

近年来,对变异语音处理的研究范围主要集中在分类和识别上. 并且在分类问题中,出现了由传统的线性分类特征向非线性分类特征转化的趋势. 鉴于变异语音数据的采集比较困难,也有人提出关于利用正常语音合成变异语音方法,这是一个很有发展的研究方向,它对增加变异语音训练数据,完善变异语音数据库有很大帮助. 另外,日本 ATR 的 MIC 实验室[8]和 Keio 大学[9]对语音中的各种情感进行了相应的研究和分析. 有关情感的研究问题,已经有学者开始将图像和语音等多通道信息结合起来判断情感[10].

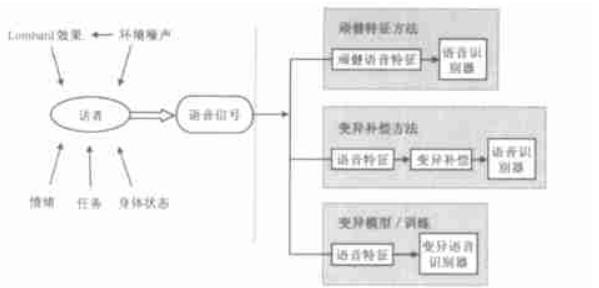

图 1 变异影响因素及识别方法[1]

随着研究内容、范围的拓宽,所涉及到的语音变异类型也在逐步扩大. 早期的研究多是关于 Lombard 变异,近期所研究语音变异的类型包括说话风格变异和单重、双重工作压力变异,从模拟的情感变异到实际的 G-Force 变异等;并且逐渐开始建立起比较规范的语料库. 最早的变异语音语料库是 Texas Instruments 收集的‘模拟变异语料库’[11],它要求话者模拟不同的说话风格进行发音,变异类型包括快(fast)、大声(loud)、轻柔(soft)、大喊(shout)和 Lombard 情况. 其中 Lombard 变异是在 95dB 的彩色噪声中获得;词表包含了包括单音素、多音素、易混词在内的 105 个词. 麻省理工学院 Lincoln 实验室建立了 35 词 11 种风格的语料库,包括正常语音、生气(anger)、大声(loud)、轻柔(soft)、慢(slow)、快(fast)、清晰(clear)、疑问(question)、50 %工作压力变异、70 %工作压力变异、Lombard 变异等[20]. 美国早期关于变异语音的研究都是在这两个数据库的基础上进行的. 在此基础上,美国 Duke 大学建立了一些在实际环境和模拟环境下采集的变异语音数据库 SUSAS(Speech Under Simulated and Actual Stress)[12]. 除了包含前述说话风格变异外,SUSAS 语料库还包含了实际 G-Force 变异和恐惧变异风格. NATO 组织中[3]“语音及语言技术”研究小组在 94 年开始一个“变异语音”项目,建立一个联合语料库,在一个广泛的国际团体中共享这个语料库并交换实验结果. 这个联合语料库包括 SUSAS 数据库及反映紧急情况下变异的 DLP 数据库和在战斗机中录制的 SUSC-0/1 数据库. 另外,英国的爱丁堡大学在加拿大录制的 DCIEM 语料库,它主要包括在睡眠被剥夺和药物作用下的语音变异数据[13]. 国内在变异语音语料库采集方面所做的工作相对较少,哈工大语音处理研究室采集了在模拟航天飞行器中,从 1 个重力加速度到 6 个重力加速度 G-Force 变异语音数据,以及在飞机起降时录制的 Lombard 变异语音数据[14].

近年来,变异条件下顽健语音识别已经得到了广泛重视. 在 ICASSP、EUROSPEECH 等重要国际会议论文集中,经常有这方面的研究论文. Speech Communication 杂志曾在 1996 年出过一个关于变异语音研究的专刊,ICASSP '99 也曾专门组织了一个关于变异语音研究的专题. 国外许多科研机构都成立了专门进行变异语音顽健性研究的小组. 国内近些年来很多研究单位,如中科院自动化所、清华大学、中科院声学所、北方交大、南京大学、哈工大等单位都对噪声和通道畸变影响下的顽健语音识别方法做过很好的研究工作,东南大学对语音信号中的情感的分析和分类也作了相应的工作[15],清华大学还对汉语语音识别中口音问题进行了研究,他们提出一种基于多模式及集成判决的稳健电话语音识别算法[16],能较好的处理包括口音变化等不同情况的发音. 中科院声学所曾对五种不同的发音方式(大声、正常、小声、加快和放慢)对音域和基频的影响作了相应的研究[17],哈工大语音处理研究室从‘九五’末期,就开始了变异语音处理研究工作[14,18,19],利用 TEO 基频特征在对 G-Force 变异的分类中取得了较好的效果[14],并尝试将调整的 Mel 频带特征用到 G-Force 变异语音识别中,使识别性能有了一定的提高[20].

从整体上看,变异语音处理研究工作还处于探索阶段. 关于变异语音分类,目前还集中在二选一的分类上,即对正常情况和一种变异情况进行分类,并且分类正确率一般小于 90 %. 在变异语音识别方面,识别率大多仅在 70 %到 80 %左右. 在下节中我们将具体讨论变异语音处理研究中的方法,并对这些方法进行评述.

3 变异语音处理方法综述

近些年来研究者们开展了许多变异语音处理研究工作,提出了一些处理方法,并取得了一些成就. 总体看来,对变异语音处理研究涉及三个方面的工作:变异语音分析,变异语音分类和变异语音识别.

3.1　变异语音分析

变异语音分析是变异语音研究工作的基础.根据语音产生模型可以知道,在变异情况下不仅声源受到了影响,声道的调音动作同样也受到了影响;因而对变异语音分析主要是从两个方面来考虑:其一是从声源激励角度考虑在变异情况下声源激励所受到的影响.体现这方面变化的典型特征是声门脉冲形状、基频、持续时间等特征;其二是从声道角度考虑变异时声道变化对语音产生的影响.典型反映声道变化的特征有声道截面系数、声道频谱系数、共振峰位置、共振峰宽度及倒谱系数的低阶部分.其中,在发音过程中声道各处截面积取决于舌、唇、颌以及小舌的位置,各部分声道截面积的不同决定了共振峰特性的不同.变异情况对声源激励和声道特性的影响最终引起识别时特征参数的变化,使得这些特征分布特性发生了改变,从而导致训练模型不能很好刻画测试环境中的特征参数分布,最终影响了系统识别效果.为了揭开变异情况下语音产生的奥秘,很多学者对语音产生的五个特征,即声源频率响应、基频、元音持续时间、强度、声道频率响应及共振峰进行了研究.

在变异条件下产生语音时,语音一些特性如基频、频谱斜率、平均共振峰位置、平均共振峰宽度、词或音素的持续时间等都会发生变化,并且这些特征在不同变异条件下变化也不尽相同,其中基频和共振峰是研究最多的特征.Cummings 等人[21]对声源激励受变异情况的影响进行了分析,从声门上升斜率、下降斜率、闭合期、关闭时间、张开时间以及声门处于峰值的持续时间等六个声门参数的角度分析,发现每种变异情况下的声门的这些参数都不相同.Williams 和 Steven 对飞行员在飞行作战遇到问题时语音数据的研究表明[22],在变异情况下基频值会增加,并且正常情况下变化比较平滑的基频轨迹,在变异情况下会变化比较剧烈.他们在后来实验中[23],通过对情感变化的数据分析得出不同情感信息对基频影响不同的结论,如悲伤时,基频值变低,基频轨迹变得平坦;生气时基频值提高,基频的变化范围扩大.文献[24]中,对基频、幅度峰值、词的持续时间三个参数作相应的分析表明,当任务的难度增加时,基频和幅度峰值增加,而词的持续时间减少.张家[illegible][12]对五种不同的发音方式(大声、正常、小声、加快和放慢)的研究结果表明,说话声级的提高导致基频升高、音域扩大,其中说话速度的变化对基频的影响较小,只是速度加快时,音域下限略有上升.Pisoni[25]发现在 Lombard 效应下,幅值、持续时间、基频均发生变化,同时辅音的频谱能量向高频带处偏移.文献[26]的研究表明,在 Lombard 效应下,对大多数音素而言,平均共振峰宽度下降,元音共振峰位置会提升;对于多数音素,共振峰的幅值增加,第一共振峰位置也会后移,这些变化会导致频谱斜率的增加.对于大声情况,其变异和 Lombard 情况类似.对于缓慢说话风格变异,其持续时间变长是最显著的特征[27],并且元音部分持续时间的变化要比其它部分变化明显.在文献[28]中 Hansen 等人从声道形状的变化、声管截面积系数变化、Mel 自相关系数的偏移等方面分析了变异情况对语音产生过程的影响.通过对选定语音每帧数据声道形状变化的研究表明,在正常情况下,声道形状的最大变化处发生在咽喉处,而在生气情况下,最大变化位置转移到舌的边缘及背部和唇等处.这表明在变异情况下,声道的调音运动受到很大的干扰.同样通过对声管截面积系数的研究也发现,变异时声道截面积系数变化和正常语音截面积系数变化也不相同.这种变异情况下调音运动变化的不同导致语音特征参数变化不同,这种差别也可以用 Mel 自相关系数变化来表示.通过对比正常情况和变异情况下,Mel 自相关系数随着帧数增加的变化趋势可以发现,在变异条件下,该系数的变化相对比较缓慢,并且出现了双峰特性.

以上这些研究都表明,正常情况下产生的语音数据和变异情况下产生的语音数据相比,各种参数都有所变化,这些参数的变化反映出了声门激励和声道在变异情况下所受到的影响,这些影响最终使得识别参数发生了变化,从而降低了系统整体的识别性能.

3.2　变异语音分类

从上节对变异语音分析的情况看,在变异条件下,语音产生过程中一些参数特征会发生变化.对这些变化参数定量分析,可以将正常语音和变异语音区分开来,或者可以用来说明语音发生变异的程度,这就是最基本的变异语音分类思想.按照所采用分类特征的不同,可以将变异语音分类方法归纳为以下几类:

3.2.1　基于线性特征的分类

它是利用在传统线性语音产生模型基础上得到的一些语音特征,如基频、共振峰、音素或词的持续时间、强度、声门激励源特性等来进行分类.其中基频和共振峰等特征的提取需要顽健的、不受变异情况影响的算法;持续时间特征一般选用元音部分的持续时间表示;强度特征用每帧信号平均幅度的平方根来表示;并用频谱斜率来代表声源特性,以及用前两个共振峰的位置来表示声道的频谱信息等.分类时,可以用其中一种特征,也可以将几种特征结合起来进行分类.分类器可以是基于隐马尔可夫模型 HMM(Hidden Markov Model)分类器、人工神经网络 ANN(Artificial Neural Network)分类器或贝叶斯分类器.文献[7]中基于这些线性特征采用贝叶斯分类器对正常、生气、大声、Lombard 等情况下变异语音进行分类,实验结果表明,基频是分类效果最好的线性特征,而持续时间、共振峰等不适合单独作为分类特征,但可以和其它特征结合到一起来进行分类.

3.2.2　基于非线性特征分类

非线性特征认为在语音产生过程中会产生涡流.Teager[29]等人在语音和听觉实验中,演示了声道中的气流是时而分离,时而附着在声道壁上的.Teager 认为涡流区域也产生语音,并对语音信号有调制作用.这种由涡流运动产生的语音是非线性的,Teager 等人用 TEO(Teager energy operator)算子来表示这种涡流对单个共振峰能量的调制.此后 Kaiser[30]给出了 TEO 的离散表示形式:

$$\Phi[x(n)] = x^2(n) - x(n+1)x(n-1) \tag{1}$$

其中 $\Phi[\cdot]$ 为 TEO 算子,$x(n)$ 是采样后的语音样本.

由于涡流区域也会产生语音,并且对语音信号有调制作用,Maragos 和 Kaiser[31]等人在此基础上进一步利用 TEO 算子

将语音信号分解成调频分量和调幅分量,即语音信号可以看成是在一定载波频率上经过频率调制和幅值调制作用后的共同结果.这部分内容在文献[32]中有详细的介绍.在此基础上,Zhou 和 Hansen 等人推导出一系列非线性分类特征,包括可以很好的反映语音信号产生时,激励的瞬时变化特性的规正后的 TEO 自相关包络特征(TEO-AUTO-ENV[33~35]);体现语音信号 FM 分量变化的 TEO-FM-VAR[34,35]特征;可以更好反映变异情况下基频特性的 TEO-PITCH[34]特征;以及按照人耳临界频带来组织滤波器的基于临界频带的 TEO 自相关包络特征(TEO-CB-AUTO-ENV[33,35]).对于正常、大声、Lombard 等模拟变异语音库,以及在过山车中录制的恐惧变异语音数据,非特定人孤立词的分类实验结果表明,非线性特征分类的效果优于线性特征.

3.2.3 基于子带分析特征的分类

基于子带分析特征方法的基本思想是,当语音产生变异时,它的频谱能量会在各频带上发生转移,比如在大声和 Lombard 情况下,语音能量通常向低频段转移,而低频段对于人耳来说比较敏感,因而人耳可以很容易感觉到变异情况的出现.最初,子带划分是通过构造相应滤波器组来完成的.Tin Lay New[36]等人用 Mel 滤波器得到规正后的 Mel 子带能量作为分类特征取得了较好的效果.自从小波理论出现之后,可以基于小波包构造一组和人耳的临界频带类似的滤波器组来分析这种转移现象,并基于此分析提取出了一些新特征来进行分类.最早由 Erzin[37]提出的子带能量系数及其倒谱形式,之后 Sarikaya[38,39]在这个基础上利用小波包理论导出了子带能量自相关系数和基于子带的倒谱参数的自相关系数,但这两个特征对于正常语音和清晰、快、慢等风格的分类效果不理想,从而影响了整体平均分类结果.近年来 Hansen[40]通过对子带能量的对数进行小波变换得到一组新的特征,并且后来将这种特征用于变异语音识别方面.

3.2.4 基于 MFCC 特征的分类

Hansen[28]等人对基于 MFCC 以及相应的扩展形式用变异语音分类进行了尝试,并提出两个新的分类特征:Mel 自相关系数 AC-Mel 和 Mel 交叉相关系数 XC-Mel,分别表示相同特征维之间的相关性和不同特征维之间的相关性.其中 Mel 自相关系数可以反映频带之间的相对能量,以及由于变异而产生的频谱斜率在帧与帧之间的相对变化;而交叉自相关系数,提供了精细的频谱结构和粗略频谱结构之间的相对变化的一种数量上的测度.另外,对于声源激励的特征,Cummings[21,41,42]提出了利用声源激励脉冲的张开点、闭合点、张开时间以及闭合时间等一系列特征进行分类的方法,也获得了较好的分类效果.

表 1 总结了有关变异语音分类的一些研究结果.尽管由于词表不同、变异类型不同,以及话者的区别,不能简单地评定这些结果的好坏,但该表还是在一定程度上反映了各种方法的性能.上述方法中有一些实验结果由于缺乏可比性,未在表中列出.

3.3 变异语音识别

一个语音识别系统是否能对各种环境变化都具有顽健性,是该系统能否实际应用的关键.一般情况下,由于训练环境和测试环境不匹配,使得系统在实际应用中识别性能较差;而顽健变异语音识别就是解决在各种变异情况下如何改进识别性能的问题.前面图 1 的右半部分,将各种变异情况下的识别方法归纳为三类:顽健特征提取、变异规正补偿、识别模型调整.下面分别对这些方法进行讨论.

表 1 各种变异语音分类方法的实验结果

研究者与研究单位	基本情况	分类特征	平均结果
G Zhou J H L Hansen[33~35] Duke 大学顽健语音识别研究室	变异类型:生气、大声、Lombard、害怕 HMM 分类器/6 孤立词分类	TEO-FM-VAR TEO-AUTO-ENV TEO-CB-AUTO-ENV TEO-PITCH	70.5 % 79.4 % 92.9 % 80.0 %
G Zhou J H L Hansen J F Kaiser[33] Duke 大学顽健语音识别研究室	变异类型:生气、大声、Lombard 贝叶斯分类器/33 孤立词分类	持续时间 强度 基频 声门源特征 第一共振峰位置 第二共振峰位置	60.7 % 73.5 % 86.7 % 70.8 % 56.3 % 45.3 %
R Sarikaya, J N Gowday[39] Clemson 大学数字语音处理实验室	变异类型:生气、清晰、中等任务 C50、高难度任务 C70、快、大声、Lombard 效应、疑问、大喊、轻柔 HMM 分类器/5 孤立词分类	MFCC MFCC 自相关系数 子带能量系数 基于子带的倒谱系数 子带能量自相关系数 子带能量倒谱自相关系数	44.3 % 45.4 % 47.0 % 59.1 % 52.5 % 48.4 %
J H L Hansen, B D Womack[28] Duke 大学顽健语音识别研究室	变异类型:生气、清晰、中等任务 C50、高难度任务 C70、块、大声 Lombard 效应、疑问、大喊、轻柔神经网络分类器/35 孤立词分类	MFCC MFCC 一阶差分 MFCC 二阶差分 MFCC 自相关系数	33.1 % 24.7 % 47.3 % 32.6 %
Tin Lay New[36] 新加坡国立大学电子工程系	变异类型:生气、不喜欢、害怕、高兴、悲伤、惊奇 HMM 分类器/连续语音分类	规正后的 Mel 子带能量	66.1 %
马永林[14] 哈工大语音处理研究室	变异类型:G-force 应力下变异 HMM 和贝叶斯分类器/10 词孤立词	TEO 基频/HMM TEO 基频/贝叶斯分类器	91.3 % 82.7 %

3.3.1 顽健特征提取

这类方法核心是提取一种对各种变异情况都顽健的特征,使得该特征对这些变异情况都不敏感.采用了这样的特征,则用正常语音训练的识别器可以对变异语音获得较好的识别效果.语音的感知过程与人类听觉系统具有频谱分析功能是密切相关的.近年来,Mel 倒谱系数在语音识别上取得了巨大成就,也是因为它综合考虑了人耳的临界频带分析能力.Hansen 等人通过对口音[43]以及变异情况下语音数据[1,44]的

研究发现,正常情况下,语音数据各频带对识别结果的影响不同于变异情况下各频带的影响.人的内耳相当于一个频谱分析仪器,对于正常语音,它的敏感区域在第一共振峰附近,所以 Mel 频带的划分加重了第一共振峰附近的权值.在变异情况下人耳的敏感区域偏移到第二共振峰附近.根据实验结果,他们对频带进行了重新划分,提出了一个修正的 Mel 尺度和一个指数-对数尺度.

Mel 频域是一个由线性频域到仿人耳频率解析域的映射;而修正后的频带划分 M-mfcc 和指数-对数频带划分 Exp-Log,则是在其基础上通过对中间段频率加强而得出的映射关系.经过重新划分频带后的语音信号,可以降低对各种变异情况的敏感度,从而提高在变异条件下的语音识别性能.

另外,文献[45]中 Hansen 等利用 LSP 参数良好的插值性能,用变异语音信号直接估计出正常的语音信号,但这里只是对元音部分进行规正,辅音部分保持不变.文献[46]中张磊等根据对 G-Force 语音数据分析,发现不同的特征维对变异具有不同的敏感程度,因此提出一种加权的 MFCC 特征,也取得了较好的识别结果.

3.3.2　变异规正补偿方法

这种方法是在识别阶段加一个变异规正过程来消除变异情况对语音特征的影响,使得规正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[47,48].从正常语音和变异语音中得到共振峰位置和带宽补偿系数的补偿因子,进而对变异语音用相应的补偿因子进行调整.该方法改进了各种变异情况下语音识别的性能,然而这种补偿方法要求具有音素边界的知识和变异类型的先验知识,并且计算比较复杂. Chen[11]提出了通过对用正常语音训练得到的词模型在倒谱域上进行线性变换来代替多重训练,并基于此对变异语音进行补偿的方法.该方法首先分析了轻柔、大声、大喊、Lombard 等情况下发音变异在倒谱域上对语音的影响,发现这种影响随着特征维数的变化可以用一个指数函数形式来刻画,他将这种影响称为变异影响因子,并假设它在一个词内不发生变化,可以简单的用正常语音特征的均值和变异语音特征的均值相减来表示变异补偿因子. Chen 用 HMM 模型节点的相对自环时间,对状态均值进行加权来表示未受干扰之前的语音信号,该方法取得了较好的效果.但在该方法中假设变异的影响是平均作用在整个语音之上,因此它在变异语音中的各个音素上都采用了相同的补偿量.而 Hansen 等通过逐帧跟踪 MFCC 前 10 个系数发现[26],在变异情况下,一个词的各部分受到的影响并不相同,因而上述方法存在着不足.基于此,Hansen 等提出一个声源产生器框架[49],它假设语音的产生过程可以用一系列的调音运动来描述,经过这些调音运动,声道达到预期的形状,从而发出特定语音信号.将这些调音运动用一些声源产生器来表示,其中每一个声源产生器可能是一个单音素、双音素或一些临时的过渡部分,一般可以将其分为元音部分/辅音部分/转移部分或有声/转移/无声语音类.在正常情况下,语音的产生可以看成是在 F 维特征空间下,从一个声源产生器到另一个声源产生器的一系列运动,这些运动轨迹是一个合理的路径.而在变异情况下,声音的激励源部分及声道的调音动作都会发生干扰,这些干扰使正常时的合理路径发生了偏移,这些偏移可以用该声源产生器在 F 维特征空间上的调整来补偿.他们首先将这种方法应用到 MFCC 系数的补偿方面[26,49,50],这里虽然也是假设变异补偿因子服从指数分布,但将一个词用几个声源产生器表示,根据各个声源产生器部分受到的影响不同,形成互不相同的补偿因子,并且在这种方法中,由于同时有变异语音数据和正常语音数据,可以直接用变异语音和正常语音对应的声源产生器中特征均值的差或商来估计变异补偿因子,因此不需要从变异语音的 HMM 模型中估计正常语音均值,这样大大简化了计算复杂性.并且对于同一个词可以根据一定原则将其划分成不同的声源产生器,每个声源产生器的补偿因子各不相同.声源产生器理论可以用于语音产生过程中的很多特征领域,如文献[50]中,Hansen 等用这个理论对前 8 个 MFCC 系数进行补偿;文献[51]中提出的方法是对前 4 个共振峰的位置和宽度进行补偿等.文献[47]是针对每个音素进行补偿的;对于每个音素,通过正常语音和变异语音,可以得到它们比值的修正因子.在识别阶段用这个因子对变异语音进行修正,使得它和正常语音接近,从而达到提高识别性能的目的.声源产生器也被用于一些能量、基音轨迹等方面的补偿[51].

此外,Stanton[52]发现,变异情况下的语音在 0～500Hz 和 4～8kHz 之间的能量存在比较严重的偏移,导致测试数据和模板之间距离加大.他提出一种特定斜率加权(Slope-Dependent Weighting)方法来减少由于能量偏移带来的频谱距离.该方法主要通过对不同频率的频谱距离加不同的权值来减弱上述频段的作用.

3.3.3　调整模型的方法

这种方法一般是在训练阶段对所用特征或模型加以改进以达到较高的识别性能.最早是 Lippman 等提出的多重风格(Multi-style)训练法[53],该方法通过采用多种模拟说话速度和环境变异情况下的语音数据来混合训练识别模型,以改进识别系统的性能.近年来的研究表明[54],在特定人的识别系统中,这种方法可以在一定程度上提高识别系统性能,但对非特定人的识别系统,系统性能反而会下降.这可能是由于在非特定人系统中,用有限的训练数据表示各种变异情况,使得数据分布过于分散;并且这种方法要求话者产生其在紧张情况下的模拟语音,而这种模拟语音往往不能充分反映该话者在实际紧张情形下的发音变异.另外,变异语音数据采集比较困难,想用多重风格训练方法概括所有变异情况不太现实. Bou-Ghazale 和 Hansen[27,55]等利用声源产生器对正常语音在参数级别上加以调整使其转化为变异语音,从而加大了训练样本数,提高了系统识别性能.这种方法假设变异语音每个声源产生器的持续时间服从正态分布,用变异语音可以估计出持续时间的分布参数.利用得到的概率密度函数随机产生变异情况下每个声源产生器的持续时间,然后对正常语音按照这个持续时间进行规正.在此基础上,用变异情况和正常情况下,每个声源产生器产生的倒谱特征平均差值作为一个干扰因

子,并对正常语音在特征参数上进行扰动,使正常语音特征接近于变异语音特征,这样可以在一定程度上加大 HMM 训练数据的规模,更好地反映出变异语音参数变化的特性,从而提高系统识别性能.除了在参数级别上对正常语音进行调整加大模型训练数据外,Hansen 等人也尝试了利用正常语音来合成变异语音的方法[56~58]以充实变异语音数据量.这种方法也可以解决由于变异语音数据量不充分而引起的识别性能下降的问题.由于前述方法是在参数级别上将正常语音转化为变异语音,无法确定转化后的参数是否能真正表示变异语音的特性.而利用基于词的 HMM 模型来合成变异语音方法,可以首先通过主观听觉测试去除那些合成效果较差的语音,然后通过变异语音分类器进一步确定合成的语音是否和变异语音接近.这样,可以充分保证合成后的语音质量,因此要比单纯的参数级别上变化的方法效果好些.

最早的补偿方法大多是假设词或音素内,补偿因子保持不变.在利用声源产生器的补偿方法中,也只是将一个词或音素分成几个声源产生器,每部分补偿因子有所不同.但有研究表明[59],在一句话甚至是一个词中,HMM 模型各状态所处的变异风格也可能有所不同.基于此,Womack 提出多通道 HMM 模型(图 2 所示),将不同变异风格结合到一个模型中,其中平面上的每个状态表示不同变异情况,如生气、大声、清晰以及 Lombard 变异等.这种方法将 HMM 模型扩展到多维,每一维可以是一种变异风格,并且每一维的状态之间可以相互转移,在同一状态及相邻状态的不同维之间也可以相互转移,因此这种方法可以在状态级别上刻画变异风格的变化.除了对模型本身的改进之外,在初始模型训练中,采用了对相应帧能量进行加权来减少低能量语音帧对参数估计的影响,并进一步用选择训练方法[60]消除奇异点对训练参数的影响.这种算法优点是可以将变异分类和变异语音识别结合到一起,并可同时对几种变异风格语音进行识别,而不需要对参数再作修改.实验结果表明这种方法有效地提高了系统性能.但不难看出,这种方法的计算量较大,训练时比较复杂.

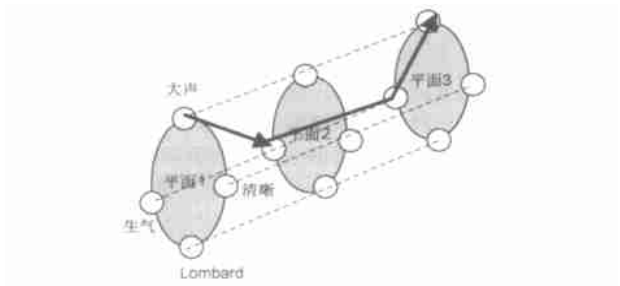

图 2 N-通道 HMM 模型[59]

表 2 概括了一些变异语音识别方法的性能.上述的方法中有一些是针对音素的实验结果,由于缺乏对比性,因此未在表中列出.

4 存在的问题及未来研究重点

尽管已经有很多学者对变异语音处理进行了研究,并提出了相应的识别算法,但到目前为止,本领域的研究还处于初级阶段,许多问题还没有很好地解决,这些问题包括:

(1) 仍没有一个统一的可以完整描述变异情况对语音影响的数学模型;

(2) 对于变异语音识别的研究,基本上还停留在对单一影响因素的研究上,而在实际情况中,往往是多种变异情况混合出现;

(3) 变异情况下顽健语音识别方法的识别率离实际的要

表 2 各种变异语音识别方法的实验结果

研究者与研究单位	基本情况	识别特征或改进模型	平均结果
J H L Hansen, M A Clements[47] Duke 大学顽健语音识别研究室	变异类型:生气、清晰、中等任务 C50、高难度任务 C70、快、大声、Lombard 效应、疑问、大喊、轻柔 HMM 识别器/20 词孤立词识别	基本系统 共振峰宽度补偿 共振峰位置补偿 共振峰位置+宽度补偿	60.3 % 63.5 % 58.8 % 64.6 %
Y Chen[11] MIT Lincoln 实验室	变异类型:轻柔、大喊、快、Lombard、大声 HMM 识别器/105 词孤立词识别	基本系统 多重风格 假设驱动	74.1 % 89.5 % 91.0 %
S Bou-Ghazale[1] Duke 大学顽健语音识别研究室	变异类型:生气、大声、Lombard 效应、HMM 识别器/30 词孤立词识别	mfcc/fft mr-mfcc/fft Exp-log/fft mfcc/lp mr-mfcc/lp Exp-log/lp	56.7 % 60.6 % 63.8 % 64.8 % 65.6 % 66.7 %
B D Womack[59] Duke 大学顽健语音识别研究室	变异类型:生气、清晰、Lombard 效应 N 通道 HMM 识别器/35 词孤立词识别	单通道 HMM 多通道 HMM	63.7 % 93.8 %
S Bou-Ghazale[56] Duke 大学顽健语音识别研究室	变异类型:大声、生气、Lombard 效应 HMM 识别器/29 词孤立词识别	正常模型 原始语音训练模型 合成语音模型	83.9 % 89.5 % 86.1 %
S Bou-Ghazale[27] Duke 大学顽健语音识别研究室	变异类型:大喊、大声、Lombard 效应 HMM 识别器/35 词的孤立词识别	正常模型 利用持续时间和 MFCC 系数规正后的训练模型	57.3 % 72.7 %
马永林[20] 哈工大语音处理研究室	变异类型:G-force 应力变异 HMM 识别器/10 词的孤立词识别	mr-mfcc/fft Exp-log/fft	63.7 % 68.6 %
张磊[46] 哈工大语音处理研究室	变异类型:G-force 应力变异 HMM 识别器/15 词的孤立词识别	加权 MFCC 特征	89.9 %

求还有很大距离;

(4)对变异语音自适应方法的研究还很少;

(5)对于提出的各种方法,缺乏一个统一的评定框架.

根据对变异语音处理研究的现状分析,我们认为如下方面可能成为未来变异语音研究主要方向:变异语音产生的基础理论研究,包括对变异语音数学建模、顽健变异特征的提取等;混合变异情况下的具体识别方法研究;对变异情况自适应研究等.

另外,变异语音合成不仅可以有效地解决变异语音数据不充足的问题,并且对于语音合成的自然度的提高也会有帮助,因此也是一个很有前途的研究方向.

5　结论

变异语音处理的难点在于变异情况对语音的影响不仅受话者不同、词不同的影响,甚至在同一词的内部,不同部分受到的影响也不相同;并且在实际应用中,往往是几种变异情况同时存在.目前提出的方法虽然在一定程度上改进了语音识别的性能,但离真正的要求还相差较远.语音识别系统要想在未来信息化社会中适合各种应用场合,就必须解决发音变异情况下顽健语音识别问题.这一问题的解决无疑对紧急情况下的电话语音呼救,行驶的车辆中、战斗机座舱中、以及未来航天飞机和数字化战场上的可穿戴计算机等情况下的声控指挥应用具有重要意义.同时,这方面的研究若能取得突破,对于我们这样的人口大国,在解决不同地域的人讲普通话时,由于口音不同所引起的变异也将会有一定的参考价值.

参考文献:

[1] Bou-Ghazale S, Hansen J H L. A comparative study of traditional and newly proposed feature for recognition of speech under stress [J]. IEEE trans on Speech and Audio Processing, 2000, 8(4): 429 - 442.

[2] Lombard E, Le Signe de l'Elevation de la Voix. Ann [J]. Maladies Oreille, Larynx. Nez, Pharynx, 1911, 37: 101 - 119.

[3] Steeneken H J M, Hansen J H L. Speech under stress conditions: overview of the effect on speech production and on system performance [A]. Proc ICASSP [C]. USA: IEEE Press, 1999: 2079 - 2082.

[4] Murry T, Nelson E J, Swenson E W. Speech Intelligibility During Exercise at Normal and Increased Atmospheric Pressures [R]. AD-A749321, 1972.

[5] Montague H. Voice Control Systems for Airborne Environments [R]. AD-A0432526, 1977.

[6] Glenn J W, Gordon R N, Moschetti G. Voice Initiated Cockpit Control and Interrogation (VICCI) System Test for Environmental Factors [R]. AD-A727574, 1971.

[7] Busch A C, Eldredge D. Duration and Intensity of Vocalic Elements as Physical Correlates of Acoustic Stress [R]. AD-A647507, 1966.

[8] Nicholson J, Takahashi K, Nakatsu R. Emotion recognition in speech using neural networks [A]. ICONIP '99 [C]. Australia: Springer Verlag, 1999, 2. 495 - 501.

[9] Moriyama T, Ozawa S. Emotion recognition and synthesis system on speech [A]. IEEE International Conference on Multimedia Computing and Systems [C]. Italy: IEEE Press, 1999. 840 - 844.

[10] Chen L S, Huang T S, Miyasato T, Nakatsu R. Multimodal human emotion/expression recognition [A]. Proceedings. Third IEEE International Conference on Automatic Face and Gesture Recognition [C]. Japan: IEEE Press, 1998. 366 - 371.

[11] Chen Y. Cepstral domain talker stress compensation for robust speech recognition[J]. IEEE Trans on Acoustics, Speech and Signal Processing, 1988, 36(4): 433 - 439.

[12] Hansen J H L, Bou-Ghazale S. Getting started with SUSAS: a speech under simulated and actual stress database [A]. EUROSPEECH '97 [C]. Greece: Patras Press, 1997. 1743 - 1746.

[13] Bard E G, Sotillo C, Anderson A H, Taylor M M. The DCIEM map task corpus: spontaneous dialogue under sleep deprivation and drug treatment [A]. ICSLP 96[C]. USA: IEEE Press, 1996. 1958 - 1961.

[14] 马永林,韩纪庆,张磊,等.应力影响下的变异语音分类[A].863计划智能计算机主题学术会议论文集[C].2001. 374 - 378.

[15] 赵力,钱向民,邹采荣,吴镇扬.语音信号中的情感识别研究[J].软件学报,2001,12(7):1050 - 1055.

[16] 潘胜昔,刘加,江金涛,等.基于多模式及集成判决的稳健电话语音识别算法研究[A].王承发,张凯.第五届全国人机语音通讯学术会议论文集[C].1998. 154 - 159.

[17] 张家[illegible].超音段特征间的相互作用[J].声学学报,1993,18(4): 263 - 271.

[18] 吕成国,张磊,韩纪庆,等. G-Stress 和 Lombard 效应作用下的变异语音语谱图[J].高技术通讯增刊,2000:223 - 226.

[19] 韩纪庆,张磊,王承发.心理紧张情况下的 Robust 语音识别方法[J].计算机科学,2000,27(9):44 - 46.

[20] 马永林.应力影响情况下的 Robust 变异语音识别方法[D].哈尔滨:哈尔滨工业大学工学,2001.

[21] Cumming K E, Clements M A. Application of the analysis of glottal excitation of stressed speech to speaking style modification [A]. ICASSP '93 [C]. USA: IEEE Press, 1993. 207 - 210.

[22] Williams C E, Stevens K N. On determining the emotional state of pilots during flight: an exploratory study [J]. Aerospace Medicine, 1969, 40: 1369 - 1372.

[23] Williams C E, Stevens K N. Emotions and speech: some acoustic correlates [J]. the Journal of the Acoustical Society of America, 1972, 52(4): 1238 - 1250.

[24] Griffin G R, Williams C E. Effects of Different Levels of Task Complexity on Three Vocal Measures [R]. AD-A2017911, 1987.

[25] Pisoni D B, Bernacki R H, Nusbaum H C, Yuchtman M. Some acoustic-phonetic correlates of speech produced in noise [A]. ICASSP '85 [C]. USA: IEEE Press, 1985. 1581 - 1585.

[26] Hansen J H L. Morphological constrained feature enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect [J]. IEEE Trans on Speech and Audio Processing, 1994, 2(4): 598 - 614.

[27] Hansen J H L, Bou-Ghazale S. Robust speech recognition training via duration and spectral-based stress token generation [J]. IEEE Trans on Speech and Audio Processing, 1995, 3(5): 415 - 421.

[28] Hansen J H L, Womack B D. Feature analysis and neural network based classification of speech under stress [J]. IEEE Trans on Speech and Audio Processing, 1996, 4(4): 307 - 313.

[29] Teager H M, Teager SM. Some observation on oral air flow during phonation [J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980, 28(5): 599 - 601.

[30] Kaiser J F. On a simple algorithm to calculate the 'energy' of a signal [A]. ICASSP '90 [C]. USA: IEEE Press, 1990. 381 - 384.

[31] Maragos P, Kaiser J F, Quatieri T F. Energy separation in signal modulation with application to speech analysis [J]. IEEE Transaction Signal Processing, 1993, 41: 3024 - 3051.

[32] 张磊, 韩纪庆, 王承发. 声道的调频-调幅模型及其在语音分析中的应用 [J]. 计算机研究与发展, 2002, 39(6): 689 - 695.

[33] Zhou G, Hansen J H L, Kaiser J F. Linear and nonlinear feature speech analysis for stress classification [A]. ICSLP '98 [C]. Australia: ASSTA Publication, 1998. 840 - 843.

[34] Zhou G, Hansen J H L, Kaiser J F. Classification of speech under stress based on features drived from the nonlinear teager energy operator [A]. ICASSP '98 [C]. USA: IEEE Press, 1998. 549 - 552.

[35] Zhou G, Hansen J H L, Kaiser J F. Nonlinear feature based classification of speech under stress [J]. IEEE Trans on Speech and Audio Processing, 2001, 9(3): 201 - 206.

[36] Tin L N, Foo SW, De Silva L C. Speech based emotion classification [A]. Electrical and Electronic Technology [C]. Proceedings of IEEE Region 10 International Conference on TENCON, Singapore: IEEE Press, 2001. 297 - 301.

[37] Erzin E, Cetin A E, Yardimci Y. Subband analysis for robust speech recognition in the presence of car noise [A]. ICASSP '95 [C]. USA: IEEE Press, 1995. 417 - 420.

[38] Sarikaya R, Gowday J N. Wavelet based analysis of speech under stress [A]. Southeast Con '97 Engineering new Century [C]. USA: IEEE Press, 1997. 92 - 96.

[39] Sarikaya R, Gowday J N. Subband based classification of speech under stress [A]. ICASSP '98 [C]. vol. 1, 1998. 569 - 572.

[40] Sarikaya R, Hansen J H L. High resolution speech feature parametrization for monophone-based stressed speech recognition [J]. IEEE Signal Processing Letters, 2000, 7(7): 182 - 185.

[41] Cumming K E, Clements M A. Analysis of glottal waveforms across stress styles [A]. ICASSP '90 [C]. USA: IEEE Press, 1990. 369 - 372.

[42] Cumming K E, Clements M A. Improvements to and applications of stressed speech using glottal waveforms [A]. ICASSP '92 [C]. USA: IEEE Press, 1992. 25 - 28.

[43] Arslan L M, Hansen J H L. Frequency characteristics of foreign accented speech [A]. ICASSP '97 [C]. Germany: IEEE Press, 1997. 1123 - 1126.

[44] Bour-Ghazale S, Hansen J H L. Speech feature modeling for robust stressed speech recognition [A]. Proceedings of ICSLP98 [C]. Australia: ASSTA Publication, 1998. 918 - 921.

[45] Pellom B L, Hansen J H L. Spectral normalization employing HMM of line spectrum pair frequencies[A]. ICASSP '97 [C]. Germany: IEEE Press, 1997. 943 - 946.

[46] 张磊, 韩纪庆, 等. 基于特征加权的应力影响下顽健语音识别方法[J]. 中文信息学报, 2001, 16(1): 7 - 12.

[47] Hansen J H L, Clements M A. Stress compensation and noise reduction algorithms for robust speech recognition [A]. ICASSP '89 [C]. Scotland IEEE Press, 1989. 266 - 269.

[48] Hansen J H L. Analysis and compensation of speech under stress and noise for environment robustness in speech recognition [J]. Speech Communication, 1996, 20: 151 - 173.

[49] Hansen J H L. Adaptive source generator compensation and enhancement for speech recognition in noisy stressful environment [A]. ICASSP '93 [C]. USA: IEEE Press, 1993. 95 - 98.

[50] Hansen J H L, Cairns D A. ICARUS: source generator based real-time recognition of speech in noisy stressful and lombard environment [J]. Speech Communication, 1995, 16(4): 391 - 422.

[51] Hansen J H L, Clements M A. Source generator equalization and enhancement of spectral properties for robust speech recognition in noise and stress [J]. IEEE Trans on Speech and Audio Processing, 1995, 3(5): 407 - 415.

[52] Stanton B J. Robust Recognition of Loud and Lombard Speech in the Fighter Cockpit Environment [R]. AD-A1970938, 1988.

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

[54] Hansen J H L, Womack B D. Classification of speech under stress using target driven features [J]. Speech Communication, 1996, 20: 131 - 150.

[55] Bour-Ghazale S, Hansen J H L. Duration and spectral based stress token generation for HMM speech recognition under stress [A]. Proceedings of ICASSP94 [C]. Australia: IEEE Press, 1994. 413 - 416.

[56] Bour-Ghazale S, Hansen J H L. HMM-based stressed speech modeling with application to improved synthesis and recognition of isolated speech under stress [J]. IEEE Trans on Speech and Audio Processing, 1998, 6(3): 201 - 216.

[57] Bour-Ghazale S, Hansen J H L. Synthesis of stressed speech from isolated netural speech using HMM-based models [A]. ICSLP '96 [C]. USA: IEEE Press, 1996. 1860 - 1863.

[58] Bour-Ghazale S, Hansen J H L. A source generator based modeling framework for synthesis of speech under stress [A]. ICASSP '95 [C]. USA: IEEE Press, 1995. 664 - 667.

[59] Womack B D, Hansen J H L. N-channel hidden markove models for combined stressed speech classification and recognition [J]. IEEE Trans on Speech and Audio Processing, 1999, 7(6): 668 - 677.

[60] Arslan L M, Hansen J H L. Selective training for hidden markov models with application to speech classification [J]. IEEE Trans on Speech and Audio Processing, 1999, 7(1): 46 - 54.

作者简介:

张 磊 女, 1973 年出生于七台河市, 博士研究生.

韩纪庆 男, 1964 年出生于哈尔滨, 博士, 教授, 博士生导师.

王承发 男, 1940 年出生于山西汾阳市, 教授, 博士生导师.