噪声鲁棒语音识别研究综述 *

雷建军[1], 杨 震[2], 刘 刚[3], 郭 军[3]
(1. 天津大学 电子信息工程学院, 天津 300072; 2. 北京工业大学 计算机学院, 北京 100124; 3. 北京邮电大学 信息工程学院, 北京 100876)

摘 要: 针对噪声环境下的语音识别问题,对现有的噪声鲁棒语音识别技术进行讨论,阐述了噪声鲁棒语音识别研究的主要问题,并根据语音识别系统的构成将噪声鲁棒语音识别技术按照信号空间、特征空间和模型空间进行分类总结,分析了各种鲁棒语音识别技术的特点、实现,以及在语音识别中的应用。最后展望了进一步的研究方向。

关键词: 鲁棒语音识别; 语音增强; 特征补偿; 模型补偿
中图分类号: TN912 **文献标志码**: A **文章编号**: 1001-3695(2009)04-1210-07

Review of noise robust speech recognition

LEI Jian-jun[1], YANG Zhen[2], LIU Gang[3], GUO Jun[3]
(1. *School of Electronic Information Engineering, Tianjin University, Tianjin* 300072, *China*; 2. *College of Computer Science, Beijing University of Technology, Beijing* 100124, *China*; 3. *School of Information Engineering, Beijing University of Posts & Telecommunications, Beijing* 100876, *China*)

Abstract: According to the problems of speech recognition in adverse acoustical environments, this paper reviewed the state of the art of robust speech recognition, and expounded the main problems of noise robust speech recognition. Based on the structure of speech recognition system, classified and summarized robust speech recognition technologies into the signal-space, feature-space and model-space technologies, and outlined the main ideas of the approaches. Finally, pointed out the problems to be further studied and the trends of developments in this field.

Key words: robust speech recognition; speech enhancement; feature compensation; model compensation

0 引言

近年来,伴随着语音识别技术的不断发展,语音识别系统的性能不断提高,纯净语音条件下识别系统取得了较高的识别率。然而,大多数语音识别系统应用于实际噪声环境时,系统性能会大大下降。大量实验表明,如果大多数现有的非特定人语音识别系统,使用不同于训练所处的环境或使用不同于训练时使用的麦克风,性能都会严重下降。而对于马路、餐馆、商场、汽车、飞机等环境中的语音信号来说,现有语音识别系统的鲁棒性更差。语音识别的噪声鲁棒性是指在输入语音质量退化,语音的音素特性、分割特性或声学特性在训练和测试环境中不同时,语音识别系统仍然保持较高识别率的性质。

基于统计模型的语音识别系统中,训练的数据必须具有充分的代表性。然而,当识别系统应用于噪声环境时,纯净的训练数据与被噪声污染的测试数据之间存在着不匹配,识别系统在噪声环境下的性能下降主要归因于这种不匹配。噪声鲁棒语音识别的研究目标就是消除或减少这种不匹配的影响,使识别系统的性能尽量接近匹配条件下的性能。由噪声引起的训练和测试的不匹配可以从信号空间、特征空间和模型空间三个层次来分析[1]。图 1 描述了语音识别中训练和测试时信号空间、特征空间和模型空间存在的不匹配。其中,S 表示训练环境下的语音数据;X 表示从训练环境下的语音数据中提取的特征; Λ_X 表示根据训练数据得到的语音模型;T、Y、Λ_Y 分别表示测试环境下的语音、特征和语音模型。当训练与测试环境不匹配时,噪声使 T、Y、Λ_Y 发生失真,从 S、X、Λ_X 到 T、Y、Λ_Y 的失真函数分别用 $D_1(\cdot)$、$D_2(\cdot)$、$D_3(\cdot)$来表示。各种噪声鲁棒语音识别技术正是从信号空间、特征空间和模型空间三个层次来消除由于训练环境和测试环境不同所带来的影响。

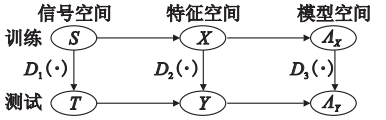

图1 语音识别中训练和测试时的不匹配

1 信号空间鲁棒语音识别技术

信号空间鲁棒语音识别技术关注对原始语音信号的处理,主要包括语音增强和语音激活检测等。

1.1 语音增强

语音增强是信号空间鲁棒语音识别技术中重要的技术之一,多年来一直受到广泛的关注,尤其是在单话筒采集条件下

收稿日期: 2008-07-18; **修回日期**: 2008-09-25 **基金项目**: 国家"863"计划重点资助项目(2006AA010102);国家自然科学基金资助项目(60475007)

作者简介:雷建军(1980-),男,博士,主要研究方向为语音识别(jjlei@tju.edu.cn);杨震(1979-),男,博士,主要研究方向为信息检索、信息安全;刘刚(1973-),男,副教授,主要研究方向为语音识别;郭军(1959-),男,教授,博导,主要研究方向为模式识别、信息检索.

如何消除背景噪声的影响更是许多人研究的课题。语音增强的目的是从含噪语音中提取尽可能纯净的原始语音信号[2]。因为噪声来源很多,特性各不相同,而语音增强处理系统的应用场合又千差万别[3],所以不存在一种可以通用于各种噪声环境的语音增强算法。实际应用时需针对不同的噪声采取特定的语音增强算法,从处理方法上分类,语音增强算法大体上可以分为基于语音周期性的增强算法[4]、基于全极点模型的增强算法[5,6]、基于短时谱估计的增强算法、基于信号子空间的增强算法[7]和基于 HMM 的增强算法[8]等。从目前的发展上看,语音增强最常用的方法是基于短时谱估计的方法,主要包括:

a)谱减法。该方法及其改进算法总体上看运算量较小,易于实时实现,增强效果也较好,是目前常用的一类方法。Boll[9]假设噪声是平稳的或变化缓慢的加性噪声,并在语音信号与噪声信号不相关的情况下,从带噪语音的功率谱中减去噪声功率谱,从而得到较为纯净的语音频谱,建立了谱减法(spectral subtraction,SS)。Berouti 等人[10]在传统谱减法的基础上增加了调节噪声功率大小的系数和增强语音功率谱的最小值限制,提高了谱减法的性能。Lockwood 等人[11]在谱减法的基础上提出了非线性谱减法(nonlinear spectral subtraction,NSS),它根据语音信号的信噪比自适应调整语音增强的增益系数,提高了语音的信噪比。Virag[12]将人耳的掩蔽效应应用到非线性谱减法语音增强算法中,部分解决了谱减法残留音乐噪声大的问题。

b)Wiener 滤波。它是一种比较传统的算法。采用 Wiener 滤波的好处是增强后的残留噪声类似于白噪声,几乎没有音乐噪声的残留[13],可以看做时域波形的最小均方误差估计。欧洲电信标准化协会(ETSI)于 2002 年 10 月发布了分布式语音识别的基于两级维纳滤波算法的噪声鲁棒性算法[14,15]。该算法应用 Mel 域三角滤波器组将维纳滤波系数转换到与语音感知相关的 Mel 域,然后在时域对语音信号进行滤波,并采用两次维纳滤波来实现噪声的消除,使得残余噪声较小,且信号各帧之间有较好的连续性,在噪声鲁棒语音识别应用中取得了较好的性能。

c)最小均方误差估计。Ephraim 等人[16]对最小均方误差(MMSE)估计进行了详尽的描述和改进,并通过实验验证了相应的一些改进算法,如最小均方误差对数谱幅度(MMSE-LSA)估计[17]。目前,对非平稳环境下的语音增强算法研究还较少。Cohen 等人[18]首先估计语音信号概率密度分布函数,然后在此基础上改进了对数谱幅度估计算法,使得改进的算法对非平稳的噪声具有良好的抑制作用。该算法的缺点是语音信号的概率密度函数较难估计。

国内外的许多学者对语音增强算法进行了研究,在平稳的声学环境及信噪比较高的情况下,语音增强得到了较好的效果。但是在低信噪比以及非平稳的噪声环境下,含噪语音信号的增强仍然是一项非常有挑战性的工作。

1.2　语音激活检测

语音激活检测的目的在于从数字语音信号中区分出语音信号和非语音信号。在语音识别时通过语音激活检测准确的区分出语音信号和非语音信号,对于提高语音识别率、节省处理时间是非常重要的。在早期的基于实验室背景的孤立词识别系统中,采用基于能量和过零率的方法可以准确地区分语音信号和噪声。但现实中的语音常常被较大的环境噪声所污染。在这种情况下,上面的方法性能开始恶化,甚至无法区分语音和噪声。在传统的基于短时能量和短时过零率的语音激活检测算法的基础上,针对不同的应用需求,研究者提出了诸多语音激活检测的改进算法,包括基于基频、谱熵、倒谱特征、高阶统计量、似然比测试等方法。另外,文献[19,20]中还研究了如何确定综合规则,综合多种方法的检测结果,以提高系统检测性能。

1)基于基频的方法　基频是一个重要的代表语音生成模型的激励源周期性的参数[21],它表示语音信号的韵律信息。由于浊音有明显的周期性,可以通过检测浊音来检测语音信号的端点。计算基频的方法很多,常用的是短时自相关法和短时平均幅度差函数法。实验结果证明,在安静的背景下,这种方法有较高的准确度;但是随着信噪比的降低,性能下降很大,而且在某些噪声环境下很难准确提取基频参数,因此不能解决这种噪声环境下的检测问题。

2)基于谱熵的方法　广泛应用于编码理论的信息熵代表信源的平均不确定性,语音的熵必定与噪声的熵存在较大差异。基于谱熵的方法[22]首先计算每帧信号的 FFT 系数,然后将每个频率点的频谱能量除以所有频带的能量总和的值作为概率密度函数。通过计算信息熵的公式得到谱熵。谱熵的方法较能量方法在低信噪比和非平稳噪声下,尤其是机器噪声环境下更为有效。但是谱熵不能解决 babble 噪声和音乐噪声背景下的检测,因为 babble 噪声和音乐噪声的谱熵与语音近似。结合能量和谱熵两种特征的方法[23],以能量弥补谱熵在 babble 噪声和音乐噪声背景下的不足,检测准确度较能量方法有显著提高。在基于谱熵的方法中引入正常数 K,改变原有的频谱概率密度函数计算形式[24,25],使得检测门限更加易于优化和确定,算法更加准确实用。

3)基于倒谱特征的方法　由于倒谱特征参数比短时能量等其他参数对语音环境的适应力强,可以利用语音信号的倒谱特征作为判决抽样信号帧是否为语音信号的依据,并使用倒谱距离测量法或循环神经网络法[26]完成对语音信号的检测。

4)基于高阶统计量的方法　由于高阶统计量本身具有的对高斯信号的抑制和相位保持的特性,使得高阶统计量被用于语音信号的处理中[27]。实验证明,基于高阶统计量的方法优于 ITU 的 G.729B[28]的性能,但在周期型噪声环境下性能有所下降,原因是这种噪声有非零的高阶统计量。

5)基于似然比测试的方法　基于似然比测试的语音激活检测算法[29,30]基于假设检验理论,引入对噪声的降噪处理,表现出较好的噪声鲁棒性。基于平滑 LRT 的检测算法[31]引入平滑参数,对基于 LRT 的方法进行改进,得到更加平稳的似然比。基于多观测的 LRT 检测算法[32,33]利用多个观测矢量进行判决,改进了 LRT 算法的性能。基于多统计模型的 LRT 算法[34]采用多个分布对语音进行建模并在线选择模型,提高了 LRT 算法的适用范围,改进了系统性能。

如何在噪声环境下准确地区分出语音信号和噪声至今仍是一个难题,目前已有的算法能够适用于一定的应用环境,但是在强背景噪声下,已有算法仍无法准确地区分出语音信号和噪声。

2　特征空间鲁棒语音识别技术

特征空间鲁棒语音识别技术力求在特征空间减小训练和测试的不匹配所带来的影响，包括鲁棒特征提取、特征补偿和特征规整等。

2.1　鲁棒特征提取

鲁棒特征提取主要研究噪声对语音的影响，试图找出抗噪能力强的特征参数。这类技术的优点是对于噪声的假设很弱，所以适用于大多数噪声环境；缺点是不能充分地利用特定噪声的性质。基于人耳听觉特性的鲁棒特征提取方法，通过对人耳听觉系统的仿真和研究，获得符合人耳听觉特性的语音特征表示，取得了较好的效果。当今，很多基于人耳听觉的特征提取方法，如 MFCC、PLP 已经成为主流的鲁棒性特征提取方法[35]。由于 PLP 特征的提取是基于语音短时谱，易受传输信道的影响。RASTA-PLP 可用来抑制这种线性谱失真。实验表明这种特征能够有效降低错误率[36]。线性鉴别分析（linear discriminant analysis，LDA）也被引入到语音特征提取中[37]。LDA 通过线性变换一方面可以最小化类内差距、最大化类间差距；另一方面可以降低特征的维数，在保证系统识别性能的基础上，提高特征的环境鲁棒性。

2.2　特征补偿

特征补偿通过对训练与测试环境之间差异的研究，在特征空间中修改测试语音的特征，使得修改后的测试语音特征能够更加接近训练语音特征。特征补偿可以分为如下两大类方法[38]：

a）基于数据驱动的特征补偿。该方法事先需要 stereo 数据库，即同时在训练环境和多个具有代表性的测试环境下录制相同内容的多套语音库，并对训练环境与这些测试环境的每一帧语音倒谱特征作比较，将差值存储起来。当系统应用到实际测试环境中，找出差值，对实际测试环境进行补偿。这样的补偿常常只适合于对应的噪声环境，测试环境变化会导致补偿效果不佳，具有较大的局限性。补偿方法主要有 SDCN、FCDCN、PDCN、RATZ 和 SPLICE 等[39,40]。SDCN（SNR-dependent cepstral normalization）事先将测试环境的每一帧语音按照瞬时信噪比的不同分成多个子集，然后在特定信噪比下计算测试环境与训练环境特征参数之间的平均差值。测试环境中，首先估计出瞬时 SNR，然后根据瞬时 SNR 将平均差值加入到含噪语音倒谱特征中，得到纯净语音特征估计值。FCDCN（fixed codeword-dependent cepstral normalization）[41] 对差值作进一步细化，在特定信噪比下，将测试环境与训练环境特征之间的差值用 VQ 聚类量化得到码本，这样不同的 SNR 对应一套码本，因此在实际应用中可调入相应的码本。PDCN（phone-dependent cepstral normalization）[42] 原理上与 SDCN、FCDCN 相似，事先需要确定每个声学单元的补偿矢量。当系统应用于实际环境中，先利用解码器解码获取假定的声学单元序列，并提取给定的补偿矢量补偿实际环境。RATZ 对纯净语音的倒谱矢量分布建立更为精确的高斯混合模型。在补偿前计算出每个混合分量所对应的均值和方差的校正项。补偿时，根据含噪语音得到不同混合分量的后验概率，从而在最小均方误差意义下计算出纯净语音特征的估计值。SPLICE（stereo-based piecewise linear compensation for environments）[43] 是在 FCDCN 基础上发展起来的，不同的是它对含噪语音的倒谱矢量建立高斯混合模型，并利用 stereo数据得到对应的每个混合分量的校正项。识别阶段根据含噪语音选择最优的混合分量，从而由该分量的校正项计算得到纯净语音特征的估计值。

b）基于统计模型的特征补偿。该方法将语音描述为参数化的统计模型，根据环境模型和最优准则估计纯净语音特征值，不需要特定环境下录制的 stereo 数据，因此具有广泛的适用性，成为当前特征补偿研究的主流。补偿方法主要有 VTS、VPS 和 SLA 等[44]。Moreno 等人[45] 采用 VTS（vector Taylor series）方法补偿噪声环境对语音识别系统性能的影响。该方法假设纯净语音和噪声分别服从高斯混合模型（Gaussian mixture model，GMM）和单一高斯分布，利用矢量泰勒级数展开方法对非线性环境模型进行线性化，保证含噪语音也服从 GMM 分布。在给定测试环境下的含噪语音序列和假设环境为平稳的基础上，利用基于最大似然的批处理 EM 算法估计噪声统计量，然后根据 MMSE 准则估计出纯净语音特征。在用 VTS 方法线性化的过程中，高阶项的忽略会带来一定的误差。VPS（vector polynomial series）[46] 采用了更为一般的函数即分段三次函数去逼近非线性函数；SLA（statistical linear approximation）[47] 采用了统计线性近似方法去逼近非线性函数。在一些噪声环境下，噪声明显与语音相关，因此采用简单的环境模型无法刻画复杂的环境。Deng 等人[48] 采用基于相位敏感性的环境模型描述噪声对语音干扰的过程，将噪声和语音信号的相关性进行了细致的分析研究。近年来，基于统计模型的特征补偿方法不断发展，针对非平稳噪声环境下的环境参数估计问题，提出了一些使用序列 EM 算法的补偿方法[49,50]，在非平稳噪声环境下取得了较好的效果。

2.3　特征规整

为了减小训练环境与测试环境之间不匹配的程度，可以对训练或者测试的语音特征进行某种变换，以使得它们的概率分布尽量接近，从而减小训练和测试的不匹配程度。特征规整也称为特征归一化、特征后处理等，是指在提取特征后，通过对特征的归一化等处理，进一步降低训练语音特征与测试语音特征之间的不匹配，提高识别系统的噪声鲁棒性。可以通过使得两者的概率密度函数的积分——累积分布函数匹配[51] 来做到这一点。根据这个原理，变换函数可以由数据的累积分布函数获得。设参数变换函数为 $x = T[y]$。其中：y 是规整前的特征参数；x 是规整后的特征参数。设 x 的累积分布函数为 $C_X(x)$，y 的累积分布函数是 $C_Y(y)$，则参数变换函数应该使得

$$C_Y(y) = C_X(x)$$

由此可以得到

$$x = T[y] = C_X^{-1}(C_Y(y))$$

实际应用中，为了算法实现的方便，经常把训练和测试的数据概率分布都变换到同一个事先给定的标准分布。这一过程即实现了对特征参数的规整。

特征规整算法主要包括倒谱均值归一化（cepstrum mean normalization，CMN）、倒谱方差归一化（cepstrum variance normalization，CVN）、倒谱均值、方差归一化（mean-variance normalization，MVN）、倒谱直方图均衡（cepstral histogram equalization，HEQ）、MVA（mean-variance normalization，ARMA filter）特征规

整等。CMN 方法[52]是特征规整算法的一个典型代表,它通过归一化处理,使得处理后倒谱特征的均值为 0,一般只能用来补偿信道畸变的影响,这是它的局限。CVN 通过归一化处理,使得倒谱特征的方差为 1,它通常与 CMN 同时使用,构成了 MVN 方法[53]。MVN 方法同时归一化特征矢量的均值和方差,因而对加性噪声也有一定的效果。HEQ[54]是一种利用特征参数的累积直方图的规整算法,它提供一个变换将含噪语音概率密度分布转换为纯净语音的标准参考概率密度分布(一般均值为 0,方差为 1),取得了比 MVN 更好的结果。此外也有人将直方图均衡方法进一步发展,提出了基于分位数的直方图均衡方法[55]。这种方法只用少量的数据便可获得数据分布的累积直方图,或者将直方图均衡与其他方法(如谱减法[56]、VTS[57]等)结合起来,综合提高系统性能。MVA[58,59]在归一化特征矢量的均值和方差之后,采用 ARMA 滤波对特征进一步进行平滑处理,提高了特征的噪声鲁棒性。将 MVA 用于不同语音特征的规整实验[60]表明,MVA 算法在多种特征后端都取得了较好的效果。

3　模型空间鲁棒语音识别技术

模型空间鲁棒语音识别技术改变训练模型的参数以适应测试语音,包括模型补偿和自适应技术等。

3.1　模型补偿

模型补偿通过对训练与测试环境之间差异的研究,在模型空间通过调整纯净语音模型参数来适应含噪的测试语音。常用的模型补偿方法有 PMC(parallel model combination)、Jacobian自适应和 VTS 方法等。PMC[61,62]将纯净语音模型和噪声模型组合,产生与噪声环境匹配的含噪语音模型。常规的 PMC 中,对纯净语音和噪声分别建立各自的 HMM 模型,然后将它们的参数转换到对数频谱域和线性频谱域中。倒谱域中高斯分布的矢量在线性谱域中为 Log-Normal 分布。对于加性噪声,可以假设两个 Log-Normal 分布的变量之和也是 Log-Normal 分布。根据这个假设,只需估计含噪语音数据在对数频谱域的均值和方差,然后经过适当的逆变换即可得到含噪语音在倒谱域的分布。PMC 的优点在于纯净语音模型和噪声模型是独立并行的,单独的噪声模型可以处理很多非稳态噪声情形,同时当背景噪声发生变化时,无须获得含噪语音数据,仅仅对背景噪声进行重估即可;缺点是当噪声很复杂时,噪声模型的状态会变多,由此带来的运算量会非常大,并且这种方法很难直接用于动态倒谱参数的补偿。文献[63]讨论了把动态倒谱参数引入到 PMC 的情况,将静态参数的连续时间导数作为动态参数以推导补偿的形式。VTS[64,65]在对数频谱域或倒谱域中采用有限长泰勒级数展开来近似计算含噪语音模型的参数。VTS 的计算量取决于泰勒级数的长度和模型参数的维数,增加泰勒级数的长度可以取得更精确的结果,但计算量也会相应增加。实验表明,VTS 要比 PMC 方法中的 Log-Normal 分布近似精确,大多情况下 VTS 方法的性能优于 PMC 方法。Jacobian 自适应[66]假设纯净语音受加性噪声的干扰,含噪语音的特征可以看成纯净语音特征和噪声特征的二元函数,后者的变化可以通过 Jacobian 行列式以反映含噪语音特征的变化。因此对于模型参数来说,含噪语音对应的模型参数就可以用噪声模型的均值和方差通过 Jacobian 行列式转换得到。Jacobian 自适应可以看做一个简化的 VTS 算法,适合模型参数的快速调整,有着与 PMC 接近的性能。

3.2　自适应技术

传统的说话人自适应技术同样可以用于噪声环境下的模型自适应。自适应技术可以利用针对使用环境的一些自适应数据对纯净语音模型参数进行更新,使得系统在该使用环境中的识别性能显著提高。目前自适应技术主要分成两大类[67],即基于变换的方法和基于最大后验概率(maximum a posteriori, MAP)的方法。前者估计非特定模型与被适应模型之间的变换关系,对非特定模型作变换,减少非特定模型与被适应环境之间的差异;后者是基于后验概率的最大化,利用贝叶斯学习理论,将非特定模型的先验信息与被适应环境的信息相结合实现自适应。还可以将两类方法结合起来,充分发挥各自的优点。

1)基于变换的方法　目前常用的基于变换的方法主要是 MLLR(maximum likelihood linear regression)[68,69]。HMM 模型中最重要的参数是混合高斯的均值和方差,MLLR 的思想就是通过一组线性回归变换函数对均值和方差进行变换,使得自适应数据的似然值能最大化。由于变换函数的参数只需较少的数据就可以估计出来,能有效地实现快速自适应。MLLR 应用最广泛的场合是将一个新的说话人或者新的环境加入到现有的模型中。一般来说,MLLR 自适应的速度要比 MAP 快,而且在数据量较少时,MLLR 要好于 MAP,但随着数据增多,MAP 会表现出一定的优势。

2)基于 MAP 的方法　基于 MAP 的自适应算法[70,71]采用基于最大后验概率准则,具有理论上的最优性,它仅对自适应语音数据出现过的语音模型进行更新,而对未出现过的语音模型不能作自适应调整。MAP 的一个明显优点是能够解决数据稀少的问题,因为它能够很好地利用模型的先验信息。对于有限的训练数据,MAP 在模型先验概率的辅助下调整模型参数。一般来说,在这种情况下,模型参数不会发生大的变化,除非这些训练数据提供了强有力的证据。MAP 其实可以看做最大似然的结果和先验知识的一个加权平均,反映了先验知识与训练数据之间的相互平衡。MAP 的缺点在于实际中一般难以得到精确的先验知识,而且只有在自适应数据中能观测到的模型参数才会被调整。当自适应数据非常多时,MAP 估计会非常接近最大似然估计,因为此时先验知识的影响已经很小了。

4　其他技术

4.1　区分性训练技术

传统声学模型训练采用基于最大似然准则(maximum likelihood estimate, MLE)的训练方法[72],算法比较成熟,语音训练时有快速算法;但 MLE 只使用与被训练模型相关的数据,忽略了模型之间的相互区分性,因此这种方法并不一定能够获得最佳的分类性能,而且对于噪声环境中的语音信号来说,其分布有可能与高斯分布的假设相差较远。为了提高声学模型在噪声环境的鲁棒性,可采用区分性训练方法,如基于最大互信息(maximum mutual information estimation, MMIE)[73]、基于最小分类误差准则(minimum classification error, MCE)[74]、基于最小音素错误率(minimum phone error, MPE)[75]等。其中,MMIE

通过最大化所有句子的期望辨识率来优化模型参数;MCE 通过直接最小化损失函数来达到最小化分类错误的目标;MPE 最大化所有句子的期望辨识率,强调音素层次的正确率,借着最大化所有可能语句的音素正确率,达到最大化所有句子辨识率的效果。

4.2 采用含噪语音进行模型训练

造成语音识别系统在噪声环境中性能下降的根本原因是在纯净环境中训练的语音模型与噪声环境中语音的统计特性不匹配。为了减少这种不匹配,一种解决方法是将实际环境的噪声叠加到训练语音数据中,用含噪的语音数据来训练语音模型。如果已知测试噪声环境,采用测试环境下的含噪语音数据进行训练可以取得较好的效果。文献[76]中采用了多种噪声数据训练方法,实验表明,语音识别系统的性能得到明显的改善。采用含噪语音直接进行训练,在小词表的情况下效果比较理想,但对于大词汇量连续语音识别效果有限。因为在大词汇的情况下,很多语音单元本身比较接近,被噪声污染后,这些语音单元的特征会发生变化,导致不同语音单元之间的区分度下降,影响系统的识别性能;而且训练和测试噪声类型、噪声水平的匹配情况将直接影响识别系统的性能,在无法预知实际应用环境的情况下,为了构造包容不同噪声类型、噪声水平的声学模型,训练数据就需要包含不同类型、不同信噪比的噪声数据。

5 结束语

本文对多年来噪声鲁棒语音识别技术进行了综合阐述,并根据语音识别系统的基本框架及训练和测试的不匹配层次,将噪声鲁棒语音识别技术按照信号空间、特征空间和模型空间的鲁棒语音识别技术进行了分类总结,详细讨论了各种鲁棒语音识别技术的特点、实现以及在语音识别中的应用。可以看到,尽管人们已经提出了多种噪声鲁棒语音识别技术,但噪声环境下的语音识别性能还远远没有达到实用的要求,特别是在低信噪比、非平稳噪声环境下,如何提高系统的识别率以及如何针对不同环境利用不同的鲁棒性方法仍需要进一步研究。近几年噪声鲁棒语音识别技术发展迅速,根据目前发展的现状,需要进一步研究的工作主要包括以下几个方面:

a)现有方法主要针对加性噪声进行研究,利用加性噪声模型实现语音模型和特征参数的建模。实际环境往往是非常复杂的,语音识别系统除了要考虑加性噪声的影响外,还需考虑卷积噪声的影响。

b)噪声鲁棒语音识别研究中,对噪声的性能研究是至关重要的,现有的研究工作主要针对平稳噪声,而对非平稳噪声考虑不多。应针对非平稳噪声环境,研究相应的噪声估计算法及鲁棒语音识别技术,以提高语音识别系统的实用性。

c)现有方法主要研究语音与噪声不相关的情况,而有些噪声与语音信号是相关的,例如在一些会议场所,语音信号会沿着墙壁的不同路径反射,产生很多与语音信号相关的干扰噪声,因此有必要考虑信号之间的相关信息。

d)信号空间和特征空间的鲁棒语音识别技术与识别系统的词汇量无关,无须对识别软件进行自适应,具有广泛的适用性。模型补偿更接近识别核,能够取得较好的效果,因此应考虑对语音增强、特征补偿、模型补偿结合算法的研究,通过对多空间算法的有效结合以综合提高识别系统的噪声鲁棒性,特别是低信噪比情况下的识别性能。

e)语音识别面临的一个重要挑战是对真实口语语音的识别,这一任务有一些区别于朗读式连续语音识别任务的问题。因为在真实的口语环境下,词汇不受约束、语音是自然的、有重叠、使用的是不明显的麦克风设备,这都对语音识别的鲁棒性产生了更高的要求,需要研究更具鲁棒性的语音识别技术。

参考文献:

[1] SANKAR A, LEE C H. A maximum-likelihood approach to stochastic matching for robust speech recognition[J]. IEEE Trans on Speech and Audio Processing, 1996, 4(3):190-202.

[2] EPHRAIM Y, LEV-ARI H, ROBERTS W J J. A brief survey of speech enhancement[K]//The electronic handbook. [S. l.]: CRC Press, 2005.

[3] EPHRAIM Y, COHEN I. Recent advancements in speech enhancement[K]//The electrical engineering handbook. [S. l.]: CRC Press, 2006.

[4] MALAH D, COX R. A generalized comb filtering technique for speech enhancement[C]//Proc of ICASSP. 1982:160-163.

[5] LIM J S, OPPENHEIM A V. All-pole modeling of degraded speech [J]. IEEE Trans on Acoustics, Speech and Signal Processing, 1978, 26(3):179-210.

[6] PELLOM B L, HANSEM J H L. An improved (Auto:I, LSP:T) constrained iterative speech enhancement for colored noise environments[J]. IEEE Trans on Speech and Audio Processing, 1998, 6(6): 573-579.

[7] EPHRAIM Y, TREES H L van. A signal subspace approach for speech enhancement[J]. IEEE Trans on Speech and Audio Processing, 1995, 3(7): 251-266.

[8] EPHRAIM Y. A Bayesian estimation approach for speech enhancement using hidden Markov models[J]. IEEE Trans on Signal Processing, 1992, 40(4): 725-735.

[9] BOLL S F. Suppression of acoustic noise in speech using spectral subtraction[J]. IEEE Trans on Acoustics, Speech, and Signal Processing, 1979, 27(2): 113-120.

[10] BEROUTI M, SCHWARTZ R, MAKHOUL J. Enhancement of speech corrupted by acoustic noise[C]// Proc of ICASSP. Washington DC: [s. n.], 1979:208-211.

[11] LOCKWOOD P, BOUDY J. Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust recognition in cars[J]. Speech Communication, 1992, 11(2-3): 215-228.

[12] VIRAG N. Single channel speech enhancement based on masking properties of human auditory system[J]. IEEE Trans on Speech and Audio Processing, 1999, 7(2): 126-137.

[13] LIM J S, OPPENHEIM A V. Enhancement and bandwidth compression of noisy speech[J]. Proceedings of the IEEE, 1979, 67(12): 1586-1604.

[14] AGARWAL A, CHENG Yan-ming. Two-stage Mel-warped wiener filter for robust speech recognition[C]//Proc of International Workshop on Automatic Speech Recognition and Understanding. 1999:67-70.