语音识别前端鲁棒性问题综述*)

刘放军　王仁华

(中国科学技术大学电子工程与信息科学系　合肥 230027)

摘　要　随着手持设备的日益小型化以及一些特殊场合的限制,使用语音识别这种自然的人机接口技术愈发显得迫切。基于HMM架构的语音识别技术经过几十年的发展,在实验室环境下已经取得了很高的识别率。当前已经取得的技术要想走向实用化,所面临的最大障碍来自于语音识别前端的鲁棒性问题。本文对语音识别的前端鲁棒性问题做了比较深入细致的分析,并在此基础上比较全面地介绍了解决这些棘手问题所采取的一些措施。文章最后对语音识别前端鲁棒性问题给出了一定的讨论和展望。

关键词　语音识别,鲁棒性,人机界面,语音识别前端,隐马尔科夫模型

The Speech Recognition Front-End Robustness: Review

LIU Fang-Jun　WANG Ren-Hua

(Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027)

Abstract　Along with the miniaturization of the handy devices and some uses limitations in some especial places, it is very exigent to use speech recognition technology as man-machine interface—the most natural communication style. The speech recognition system based on HMM has got rapid progress in lab circumstance after few decades development. The most severe obstacle of the technology way to application lies in the front-end robustness problems. In this paper, detailed analysis has been done about the speech recognition front-end robustness problems and the main means used in resolving such problems. Some discuss and expectations of the speech recognition front-end robustness technologies have been made in the end.

Keywords　Speech recognition, Robustness, Man-machine interface, Speech recognition front-end, Hidden markov model

1　引言

随着无线手持设备和无线网络的迅速普及,人们将可以实现在任何时候、任何地方、跟任何人、以任何方式传递任何信息。而手持设备的日益小型化却带来了输入困难的尴尬局面;同时,在一些特殊场合,比如驾车过程中的打手机问题,很多国家法律明令禁止。更有很多信息服务领域,急切需要实现信息咨询的自动化。凡此种种,都对语音识别技术产生了巨大的需求。经过几十年的努力,语音识别技术已经取得了巨大的进步。

然而,一旦这些技术使用在实际环境中,因为环境噪声、信道和说话人等方面的影响而使识别率大幅度下降。语音识别前端鲁棒性技术就是在系统的前端解决这种环境影响的技术。

本文首先较深入地分析了鲁棒性问题的起因,接着对语音识别前端鲁棒性方面的现有各种主流技术进行了比较全面的分析和比较,力图清晰展示这方面研究的现状。文章最后对语音识别前端鲁棒性的现有各种技术进行讨论,并对它的进一步发展进行了展望。

2　语音识别的研究现状

2.1　语音识别的基础理论

自动语音识别(Automatic Speech Recognition, ASR)是指让计算机听懂人的语音的技术。对语音识别的研究可以追溯到大约50年前。最早的语音识别系统多基于声学语音学理论,且通常是特定说话人的简单孤立词识别系统。上世纪60年代,动态规划被引入到语音识别的模版匹配方法之中,导致了DTW算法的提出,并成为六、七十年代语音识别的主流基础算法之一。80年代,隐马尔科夫模型(Hidden Markov Model, HMM)被引入到语音识别当中,是语音识别发展过程中的一个里程碑。目前大部分实用系统都基于这个统计模型。

一个典型的语音识别系统包括图1所示的几个部分:语音特征提取,基于HMM结构的声学模型训练,模式匹配,语言模型对语音单元相关性的表达和利用。

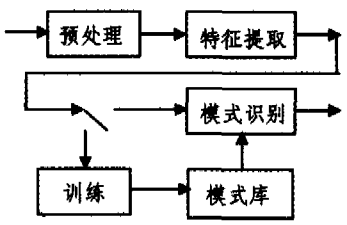

图1　语音识别原理框图

*)本课题得到了自然科学基金(编号:60275038)的资助。刘放军　硕士研究生,研究方向为语音识别的前端鲁棒性;王仁华　教授,博导,研究方向为人机语音通信、数字信号处理、多媒体通信等。

1)预处理:包括语音信号采样、反混叠带通滤波、去除个体发音差异和因设备、环境而引起的噪声影响等,并涉及到语音识别基元的选取和端点检测问题。

2)特征提取:用于提取语音中反映本质特征的声学参数,如平均能量、平均过零率、共振峰等。

3)训练:在识别之前通过让话者多次重复语音,从原始语音样本中去除冗余信息,保留关键数据,再按照一定规则对数据加以聚类,形成模式库。

4)模式匹配:它是整个语音识别系统的核心,根据一定规则(如某种距离测度)以及专家知识(如构词规则、语法规则、语义规则等),计算输入特征与库存模式之间的相似度(如匹配距离、似然概率),判断出输入语音的语义信息。

2.2 语音识别前端鲁棒性问题

目前,许多在实验室环境下性能很好的系统一旦放到实际环境中,系统性能往往就急剧下降。产生这种现象的原因是实际环境极为复杂多变,在训练过程中通过训练数据获知的语音信息无法反映实际环境中的语音信息,从而给语音识别系统的各个部分都带来巨大挑战。归纳起来,实际环境中,影响语音识别系统性能的因素主要有以下一些:

1)信道影响(线性滤波噪声)

这主要由不同麦克风之间的频率响应的差异引起。通常可以把麦克风的频响对语音的影响等效于一个 LTI 滤波器(当然这并非完美,许多麦克风还存在相当程度的非线性),不同的麦克风相当于与不同的线性滤波器进行卷积。这个卷积过程所带来的影响与语音本身内容完全无关,因而是需要用一些方法尽量设法去除的。

2)加性噪声

这是最常遇到的对语音的干扰。实际环境之中,总是存在各种各样的环境噪声。这些环境噪声对语音的影响通常可以用一个叠加模型来进行较好的描述。当环境噪声级别较高时,就很有可能对语音质量或者语音识别器带来较大的影响。环境噪声的分类非常多样。可以将它们分为窄带(带限)噪声和宽带噪声。也可以分为平稳噪声和非平稳噪声,通常平稳噪声更容易处理,而非平稳噪声的影响大多很难去除。

3)口音

不同的人说话所带的口音将会导致他们的语音的特征参数处于不同的参数空间中,使不同的语音单元的分布出现更大的交叠,并给识别器带来更大的困难。但是,对一个实用的可能需要面对成千上万潜在用户的非特定人语音识别系统而言,可能的使用者常常是带有一定口音的。因而,对口音的处理也就成为这类系统的设计与实现中一个很重要的方面。

4)Lombard 效应

在较强噪声环境中,为了使自己的语音更为清晰可懂,人说话时通常会发生一些变化,比如音量普遍增高、语音持续时间更长等等,在语音频谱上通常也会有程度不同的变化。通常把这类现象称为 Lombard 效应。这种现象也会使得噪声环境下的语音与安静环境下的语音之间出现差异,从而给识别器带来麻烦。

5)语音的复杂多样

语音是非常复杂多样的,尤其是在自然语音(Spontaneous speech)中,往往充斥了连读、轻声、省略、插入语以及各种各样的语气词等。要很好地处理这些多变的语音现象是非常困难的,这又成为当前语音识别的又一个难点。

6)说话人自身声音的变化

人的成长以及衰老、生病等生理变化,都可能会引起语音的变化。这种不稳定性会给语音识别系统,特别是一些特定人识别系统带来较大的影响。

上述因素中,信道影响与加性噪声是两种最为常见的因素。各种因素也经常在同一环境中一起出现。但在不同的具体环境之中,各种因素对识别系统影响程度可能有不同:如在电话语音识别环境下,信道影响与口音问题往往更加突出;而在手机等移动设备上,加性噪声往往会起到主导作用;如果噪声很强,Lombard 效应的影响也不能忽略。

2.3 环境模型和环境因素对特征参数的统计特性的影响

根据上面所述的 6 类引起鲁棒性问题的原因,经过一定程度的简化,着重考虑最常见的加性噪声和卷积噪声,用下面的模型来模拟环境因素对干净语音信号的影响。

其中加性噪声一般像:风扇声,机器声,砰门声或者其他说话人的声音等;

卷积噪声一般像:由于墙壁等物反射的回音,麦克风的频率响应,A/D 转换器的滤波特性,电话线的回波等。

一种噪声环境的环境模型如图 2 所示。

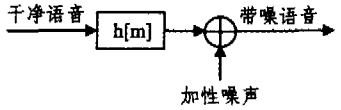

图 2 噪声环境模型

$x[m]$,$h[m]$,$n[m]$,$y[m]$分别表示干净语音、卷积噪声、加性噪声和带噪语音。这些信号在时间域的关系为:

$$y[m]=x[m]*h[m]+n[m] \tag{1}$$

这些信号在频率域的关系为:

$$Y(f_k)=X(f_k)H(f_k)+N(f_k) \tag{2}$$

其中 $0\leqslant k\leqslant K$

在倒谱向量之间的关系为:

$$y=x+h+C\ln(1+e^{C^{-1}(n-x-h)}) \tag{3}$$

$$g(z)=C\ln(1+e^{C^{-1}z}) \tag{4}$$

其中 C 为 DCT 变换矩阵。

可以很明显看出来,如果知道干净语音、加性噪声和卷积噪声的倒谱,可以计算出带噪语音的倒谱。

一般情况下,如果干净语音信号服从高斯分布,则带噪语音信号不一定呈高斯分布。带噪语音信号的均值相对于原来的干净语音的均值有一定的偏移,而方差相对于原来的干净语音信号的方差有压缩或者扩张。

3 解决这些鲁棒性问题的主流方法

噪声鲁棒性问题的根源可以归结为语音识别训练和测试环境之间的不匹配,大多数噪声鲁棒性方法都可以从设法减小两者的不匹配这个角度来解释。现有的进行语音识别前端鲁棒性的方法很多,但是最基本的还是最经典的那些,别的方法都可以看作是在这些方法的基础上的改进或者某些方法的结合。对于这些经典的前端鲁棒性方法,可以分为 4 个大类。

鲁棒性特征提取:使得特征参数只与语音信号有关,或者说是抗噪声的,特征参数具有相对的稳定性,在面对环境变化时尽可能不变化或者变化很小,可以对噪声情况和纯净情况使用相同的系统配置。

语音增强:消除或者减小语音数据中的噪声的影响,以使得到的特征参数尽可能接近无噪声环境下的情况,然后使用

纯净的语音模型进行识别。

模型补偿:将已经训练好的声学模型从训练环境变换到测试环境,使模型尽量与实际环境下的特征参数匹配。

麦克风阵列:它近年来鲁棒性语音识别方面特别是车载语音识别方面的一个研究热点,它试图利用麦克风阵列的空间分辨能力来获得进一步的语音和噪声信号。

3.1 鲁棒性特征提取

常用的特征有倒谱特征(Cepstrum)、PLP(Perceptual Linear Prediction)特征、MFCC(Mel Frequency Cepstrum Coefficients)特征。实验表明,倒谱系数用在语音识别中比传统的线性谱系数要好得多,而模拟耳朵某些听觉机理的 PLP 特征和 MFCC 特征都取得了比 Cepstrum 较好的效果。

CMN(Cepstral Mean Normalization)。在不存在加性噪声和信道影响稳定的情况下,这种方法能非常有效地消除卷积噪声,性能相当好。做法是在训练数据和测试数据中将每句话的倒谱均值减去,就可以实现消除的效果。这种方法的不足之处在于,当语句比较短的时候,倒谱均值可能估计不准,从而导致错误率增加。当非平稳加性噪声出现时,性能就会出现较大的下降,并且它无法区分静音段(包括噪声段)和语音段。为了弥补这个缺陷,先用 VAD 检测出静音段和语音段,再分开做 CMN。

MVN(Mean and Variance Normalization)。由 CMN 方法进一步扩展得到,也就是在均值和方差两个方面同时对语音特征进行规整,可以近似消除加性噪声的影响。但由于噪声的多变性,因此适用的范围不大,对识别率的提高相对有限,一般比 CMN 好一点。实际中噪声的影响可能不仅仅改变分布参数,甚至连分布都会改变。

MVA(MVN+ARMA)。它是 MVN 方法加上 ARMA 滤波器的结果,ARMA 滤波器实质上是一个低通滤波器,起平滑作用。事实上,干净语音中的一些突变的峰值往往代表着很重要的信息,而带噪语音中的一些毛刺现象则常是由噪声引起的,因此平滑时应该兼顾两方面。这样一来,滤波器阶数 M 就应该有个最优值:如果太小,则会保留一些短时倒谱信息,但对噪声难以容忍;如果太大,则会有较好的抗噪声性,但短时倒谱信息可能也会随之丢失。往往采取实验的方式来寻找最佳的 M 值。

双高斯方法(Double Gaussians)。实验表明,噪声影响后的语音分布经常出现双峰的结果[33],原因在于语音段和噪声段的特征参数的统计分布特征差别较大,故不同段的特征参数集中于不同区域。基于这种现实,使用 2 个高斯分布的线性组合来描述带噪语音的分布,并使用 EM 算法估计双高斯分布的参数,最后使用归一化方法对训练数据和测试数据进行规整,以降低它们间的不匹配。

直方图均衡方法(Histogram Equalization)。通过统计特征参数各维的分布情况来得到语音特征的加性密度分布函数,利用这个分布函数对训练数据和测试数据进行归一化操作,以减小它们之间的不匹配。这种方法和谱相减方法的结合,效果会更好。

动态倒谱规整方法(Dynamic Cepstrum Normalization)。它可以消除缓变的噪声成分。因为动态特征反映的是连续帧之间的相关性,而这种相关性能在一定程度上消除缓变成分。

LDA(Linear Discriminant Analysis)。它是一种参数空间变换方法,考虑到一般的特征中都多少存在冗余信息,LDA 变换就是要丢掉容易引起混淆的分量,只留下对分类有用的分量,通过变换不但能降低维数以降低运算复杂度(特征维数和运算量呈一种指数关系),而且能提高特征参数的噪声鲁棒性。这种方法的缺点是对训练环境和测试环境的不匹配敏感,尤其是对训练数据和测试数据的信噪比不匹配很敏感,训练 LDA 矩阵的数据需要得比较多,不然就不能得到鲁棒性很好的 LDA 矩阵。这种方法对新的环境需要重新估计变化。

3.2 语音增强

3.2.1 消除加性噪声

维纳滤波方法(Wiener Filtering)。它实质上就是最小均方误差准则在信号滤波中(时间域)的应用,按照这个准则来自适应调整滤波器的系数。这个方法降噪比较有效,不过它和谱相减一样,只对加性噪声很有效,所以大多数时候要和其它处理方法结合起来使用,并且它们有可能会损害语音信号。

谱相减方法(Spectral Subtraction)。假设噪声信号可以获得或者可以估计到,并且语音和噪声统计独立,那么干净语音的功率谱可以通过带噪语音的功率谱减去噪声的功率谱。谱相减方法可以看作是维纳滤波方法的一个特例。需要注意的是,计算中的 SNR 的估计要进行时间域和频率域上的平滑。这种方法对静止或慢变的加性噪声很有效,但也存在一些问题[30]。首先是噪声估计的问题,如果静音段检测不准确,就会导致噪声估计产生偏差;其次,带噪语音谱和噪声谱的差可能出现负值,此时通过门限的设置来解决,但同时这种操作会带来所谓的 Musical Noise,这种方法不能有效地处理信道影响。

3.2.2 倒谱域特征参数补偿

根据实验结果和理论分析发现,模型域方法对识别性能的提高有个极限问题。SPLICE(Stereo-based Piecewise Linear Compensation for Environments)的出现打破了这个极限,而且性能相当好。它属于倒谱域特征参数补偿方法,从带噪倒谱中得到干净倒谱的一个估计,因而也是一种语音增强方法。CMU 的 Acero 在 1990 年的博士论文中[52]提出 SDCN(SNR-Dependent Cepstral Normalization)、CDCN(Codebook-Dependent Cepstral Normalization)方法以及它们的改进方法 ISDCN(Interpolated SDCN)和 FCDCN(Fixed CDCN)之后,CMU 的 Moreno 在 1996 年的博士论文中[53]对这些方法做了进一步的改进。此后 Acero 等人又提出了 MFCDCN(Multiple FCDCN)[54],最终产生 SPLICE 方法[56,57,59~62]。FCDCN 结合了 SDCN 和 CDCN 两种方法的优点,即同时获得 CDCN 的精度和 SDCN 的低运算量。

SPLICE 方法仍然是基于式(3)、(4),主要是从带噪倒谱中估计出干净倒谱,是对信道影响和加性噪声的联合补偿,所以一般性能都比较好。它的优点是不需要对噪声进行建模,这是因为噪声特性已经隐含在双通道数据之间的映射中。它能处理很多噪声,包括加性噪声、卷积噪声、非稳态噪声,甚至在时域上引起的非线性畸变。SPLICE 能成功使用的关键是测试环境的畸变要尽可能和从双通道数据中估计出来的偏移矢量所反映的畸变相似,但是这个方法对双通道数据的依赖仍然是个问题。

3.2.3 非完整特征法

这类方法将受噪声干扰的特征参数看作非完整的特征,从而设计针对该非完整特征的识别方法。

频谱子带法。这个方法基于这样的事实:人的听觉系统独立处理语音信号的各个频率子带,信息可能只在一些频带被污染,子带的特征参数可能比全频带的更有优势[18],语音

信号各频带的短时平稳性的转移可能异步[15]，各个子带可以构造适合该子带的特征。它的做法分成串行和并行两种方式。对于并行模式，从语音信号得到语音频谱，然后分成若干个子带，分别通过不同的识别器进行识别，最后把各个识别器识别的结果合并起来得到最终的结果。对于级联模式，与并行模式不同之处在于不同子带的谱先经过级联以后才送入识别器得到最终的识别结果。

3.2.4 Missing Feature Approach

这个方法基于如下认识：人的听觉系统能处理不完整语音并且它具有屏蔽效应以及对听觉场的分析。实现的方法是将特征参数分为非可靠部分与可靠部分，然后可使用2种方法解码：一种称作数据内插(data imputation)方法，内插非可靠特征，再使用全部特征解码；另一种所谓的边缘法(marginalization)，只使用可靠特征解码。

该法存在的问题是，如何准确划分特征参数为非可靠部分和可靠部分。

3.3 模型补偿

3.3.1 带噪语音训练模型

使用带噪数据训练模型，训练分为用与测试环境匹配的带噪数据训练模型和用多条件下的带噪数据训练模型两种方法，后者比前者应用灵活，但是前者的性能一般来说要好于后者。其实现方法可分为直接录制带噪数据和先录制噪声数据、后叠加到纯净训练数据上两种方式。

3.3.2 HMM分解

PMC方法(Parallel Model Combination)。假设干净语音和噪声都是混合高斯分布，从而通过变换组合得到带噪语音的分布[30]。在这个方法里面，干净语音数据和噪声各使用一套模型，如果已知干净语音和噪声在倒谱域的模型参数(均值向量和协方差矩阵)，这里只考虑加性噪声，在线性频率域通过这两个模型参数的直接相加得到带噪语音的模型参数，从而得到合成的带噪语音模型。PMC的优点在于干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，我们不需要获得带噪语音数据，仅仅对背景噪声进行重估就可以了。PMC的缺点是当噪声很复杂时，此时噪声模型的状态会变多，由此带来的运算量会非常大，并且这种方法不能直接用于1阶、2阶倒谱特征。

Vector Taylor Series。由(3)和(4)式根据一阶Taylor级数展开可得到近似的结果。实验表明，这个近似要比PMC方法中的对数正态分布近似要精确，很多情况下一阶VTS方法性能都比PMC方法好。

3.3.3 HMM自适应

3.3.3.1 模型参数变换

基本思想：调整模型参数，使得训练环境和测试环境之间的不匹配最小。

MLLR(Maximum Likelihood Linear Regression)[12]。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要较少量的数据就可以估计出来，因此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。

最大后验估计MAP(Maximum A Posteriori)[8]。它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[30]。

MAP其实可以看作ML的结果和先验知识的一个加权平均，反映了先验知识和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ML估计，因为此时先验知识的影响已经很小了。

3.3.4 HMM扩展

3.3.4.1 动态贝叶斯网络(DBN)

动态贝叶斯网络(Dynamic Bayesian Network)是将BN扩展为能描述时间过程的动态形式。HMM仅仅是DBN的一个特例。DBN的优点有：

非线性。各随机变量的关系通过条件概率表(CPT)反映，定义灵活。

解释性。每个随机变量对应具体物理量。

因子化。联合概率分布易于因子化，具有统计有效性和计算有效性。

扩展性。图状结构易于扩展。

在基于HMM的识别系统中，一般使用上下文相关模型单元对上下文相关性建模。在基于DBN的系统中，可以同时使用上下文相关模型单元和上下文相关DBN。

这种方法有复杂而庞大的结构和参数，它在语音识别中的应用现在还不成熟，推导过程中做了大量的近似，目前尚没有非常明显地提高性能。

3.3.4.2 因子化HMM

因子化HMM(Factorial HMM，FHMM)是DBN的一个特例，也是标准HMM的扩展形式。HMM将时间域的信息编码在一个隐含状态串中，每个状态可以对应所有的观测值。这种结构限制了HMM的表达能力。比如为了表达30比特的时间历史信息，一个HMM需要个2^{30}状态。如果在HMM中使用分布式的状态形式，则使用30个状态即可，大大减少了状态的数量。这种含有分布式状态结构的HMM即为FHMM。FHMM的动机有两方面：一是模型能自动分解状态空间以对应观测数据中的多种因素；二是如果对观测数据中的隐含因素具有先验知识，分布状态形式可以简化对多因素数据的表达。

3.3.4.3 HMM误差模型

HEM(HMM Error Model)有2个流：一个流仍然为隐含状态序列；另一个流为混合高斯模型(Gaussian mixturemodel，GMM)。第一个流将观测数据从特征空间转换到一个归一化的空间，其中的数据是独立同分布(independent and identical distribution，iid)的。这个流将观测数据从一个空间"滤波"到另一个空间，所以被称为"滤波模型"(Filter model)。第二个流模拟归一化空间。因为它模拟的是"残差"数据，所以被称为"残差模型"(Residual model)。标准HMM是HEM的一个特例。当HEM的残差模型为一个高斯分布时，HEM等同于一个标准HMM。为方便起见，通常设该高斯分布是均值为0、方差为1的分布N(0,1)。所以HEM是比

HMM更灵活的模型在于它用GMM模拟了归一化空间的残差数据。

3.4 麦克风阵列(Microphone Array)

对车载和Hands-Free环境下的语音增强和语音识别等任务,麦克风阵列方法是一个非常常用而且比较有效的方法。特别是对于非稳态噪声,麦克风阵列的空间选择性提供了另一条解决途径。

对于空间非稳态信号(声源位置发生变化),可通过声源定位和阵列的steering得到一定的解决。

麦克风阵列由于利用了空间选择性,在降噪和去混响等方面拥有一些单麦克风系统所不具有的优势。大量实验也证明了麦克风阵列的有效性,以及与一般鲁棒性方法组合运用的可行性。但是麦克风阵列在成本、设备尺寸、所需运算资源等方面要求较高,这影响了它的实际应用。

方法的总结以及对语音识别鲁棒性前端技术的展望 目前处理噪声鲁棒性的方法主要可分为4类:鲁棒性的特征提取、语音增强、模型匹配和麦克风阵列。有些方法已经被证明是有效的且具有较好的鲁棒性。比如特征提取中的CMN方法。CMN除去倒谱参数的均值,能有效地消除缓变信道噪声的影响。CMN对加性噪声也有一定的效果,这是因为在倒谱域上加性噪声经过一定的假设可以近似为卷积噪声。而且对纯净测试数据,CMN一般也不会导致识别率的下降。又如模型自适应方法MLLR,在自适应数据量充足的条件下,能有效地提高对带噪数据的识别率。

大部分噪声鲁棒性语音识别方法均存在着各自的假设条件,因而一般只适用于相应的特定环境。比如语音增强中的谱减法、子带法和忽略特征法等都假设噪声频谱是已知的或可估计的,所以它们的性能受到噪声频谱估计精确度的影响。又如模型补偿中的PMC方法,假设噪声模型是可以训练得到的,这也要求噪声信号是事先可以获得的。用带噪数据训练声学模型的方法则假设带噪数据是足够多的。即使上面提到的两种鲁棒性较高的方法CMN和MLLR,也存在一定的条件限制。在CMN中,如果我们对每个测试句估计倒谱均值,然而测试句不是足够长,可能会导致估计均值出现偏差,进而导致识别性能的恶化。MLLR则在自适应数据量不充足的情况下会导致识别率的下降。

目前的大部分噪声鲁棒性语音识别方法尚只适用于平稳噪声或缓变噪声的情况,对于非平稳的噪声尚不适用。非平稳的噪声如其他说话人的语音和背景音乐声等。对于需要已知噪声的方法,比如谱减法、子带法、忽略特征法和PMC方法等,由于测试句中的非平稳噪声不可能事先获取,因此无法适用于对非平稳噪声的消除。对于HMM自适应方法,虽然它们可以用非监督学习的方式进行在线的自适应,但由于每次的自适应过程实际上仍使用前一次的识别结果作为自适应脚本,因此在线自适应实际上只能有效地处理缓变噪声,而对变化较快的噪声达不到较好的效果。目前有一些研究者已开始着手处理非平稳噪声的影响,比如渐进噪声估计。它是一种跟踪噪声变化的模型自适应算法,但当噪声变化较快时,会由于跟踪难度的增大而影响性能的改善。又如使用动态贝叶斯网络做声学模型的方法,可以将噪声的非平稳性在模型中用状态转移来表达。但动态贝叶斯网络理论上尚有值得探讨的地方,需要更深入的研究工作。麦克风阵列虽然利用空间选择性,能比较有效地处理非稳态噪声,但是成本、运算资源方面的因素限制了它的广泛应用。

综上,前端鲁棒性问题是一个十分复杂的问题,不是一个或者两个方法所能单独有效解决的。要设计出高效的语音识别系统,现有前端方法的有机结合以及语音识别相关信息的有效利用是关键,这也是后续工作的重点。

鸣谢 本课题得到了国家863高科技发展计划(编号:2003AA252031)的资助,在此表示感谢。

参考文献

1 Acero A. Acoustic and Environmental Robustness in Acoustic Speech Recognition:[PhD thesis]. CMU, 1990
2 Allen B J. How Do Humans Process and Recognize Speech? IEEE Trans on Speech and Audio Processing,1994,2(4):567～577
3 Barker J, Cooke M, Green P. Robust ASR on Clean Speech Models: An Evaluation of Missing Data Techniques for Connected Digit Recognition in Noise. Eurospeech,2001
4 Bregman A S. Auditory Scene Analysis. Cambridge, MA: MIT Press,1990
5 Droppo J,Deng L,Acero A. Evaluation of the SPLICE Algorithm on the Aurora2 Database. In:Eurospeech,2001. 217～220
6 Gales M J. Model Based Techniques for Noise Robust Speech Recognition:[PhD thesis]. Cambridge University,1995
7 Gales M J F. Transformation Streams and the HMM Error Model. Computer Speech and Language,2002,16(2):225～243
8 Gauvain J L, Lee C H. Bayesian Learning of Gaussian Mixture Densities for Hidden Markov Models. In: Proc. of the DARPA Speech and Natural Language Workshop, Palo Alto, CA, 1991. 272～277
9 Gong Y. Speech Recognition in Noisy Environments: a Survey. Speech Communication,1995,16:261～291
10 Hermansky H, Morgan N. RASTA Processing of Speech. IEEE Trans of SAP, 1994,2(4):578～589
11 Hunt M J,Lefebvre C. A Comparison of Several Acoustic Representations for Speech Recognition with Degraded and Undegraded Speech. In:ICASSP,1989. 262～265
12 Legetter C J,Woodland P C. Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. Computer Speech and Language, 1995,9:171～185
13 Lippmann P R. Accurate Consonant Perception Without Mid-Frequency Speech Energy. IEEE Trans on Speech and Audio Processing,1996,4(1):66～69
14 Logan B,Moreno P J. Factorial Hidden Markov Models for Speech Recognition: Preliminary Experiments:[Technical Report]. Cambridge Research Lab,1997
15 Mirghafori N,Morgan N. Transmissions and Transitions: A Study of Two Common Assumptions in Multi-Band ASR. In:ICASSP, 1998. 713～716
16 Moore B C J. An Introduction to the Psychology of Hearing. 4th ed. New York:Academic Press,1997
17 Moreno P. Speech Recognition in Noisy Environments:[PhD thesis]. CMU, 1996
18 Raj R. Reconstruction of Incomplete Spectrograms for Robust Speech Recognition:[Ph D Thesis]. CMU,2000
19 Rao S,Pearlman W A. Analysis of linear prediction, coding, and spectral estimation from subbands. IEEE Trans on Information Theory,1996,42:1160～1178
20 Riener K, Warren R,J B Jr. Novel findings concerning intelligibility of bandpass speech. Journal of the Acoustical Society of America, 91(4):S2339
21 Siohan O. On the Robustness of Linear Discriminant Analysis as a Preprocessing Step for Noisy Speech Recognition. In: ICASSP, 1995. 125～128
22 Okawa S,Bocchieri E,Potamianos A. Multi-band speech recognition in noise environments. In:Proc. ICASSP,1998. 641～644
23 Rahim M G,Juang B H. Signal bias removal by maximum likelihood estimation for robust telephone speech recognition. IEEE Trans on Speech and Audio Processing,1996,4(1)
24 Tamura S,Waibel A. Noise Reduction using Connectionist Models. In:ICASSP,1988. 553～556
25 Tibrewala S,Hermansky H. Sub-band based recognition of noisy speech. In:Proc. ICASSP,1997. 1255～1258
26 Varga A P,Moore R K. Hidden Markov model decomposition of

speech and noise. In: ICASSP, 1990, 2. 845～848
27 Warren R, Riener K, J B Jr, Brubaker B. Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits. Perception and Psychophysics, 1995, 57(2): 175～182
28 Zweig G. Speech Recognition with Dynamic Bayesian Networks: [Ph D Thesis]. Berkeley: University of California, 1998
29 Young S, et al. The HTK Book (for HTK Version 3.2), December 2002
30 Young S. Large Vocabulary Continuous Speech Recognition: a Review. In: Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding, Utah: Snowbird, December 1995. 3～28
31 Huang X, Acero A, Hon H. Spoken Language Processing. Prentice Hall, 2001. 375～538
32 Gong Y F. Speech Recognition in Noisy Environments: a Survey. Speech Communication, 1995, 16: 261～291
33 Juang B H. Speech Recognition in Adverse Environments. Computer Speech And Language, 1991, 5: 275～294
34 Openshaw J P, Mason J S. On the Limitations of Cepstral Features in Noise. In: Proc of ICASSP 1994, Adelaide, Australia, April 1994. 49～52
35 Hermansky H. Should Recognizers Have Ears? Speech Communication, 1998. 25: 3～27
36 Chen C P, Filali K, Bilmes J A. FrontEnd Post-Processing and BackEnd Model Enhancement on the Aurora 2.0/3.0 Databases. In: Proc. of ICSLP 2002. Denver, Colorado, September 2002. 241～244
37 Haeb-Umbach R, Ney H. Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition. Proc of IEEE on Acoustics, Speech and Signal Processing, 1992, 1: 13～16
38 Ho D K C. Speech Enhancement: Concept and MethodologyUniversity of Missouri-Columbia
39 Juang B H, Soong F K. Hands-free Telecommunication. In: HSC 2001(International Workshop on Hands-Free Speech Communication), 2001. 5～10
40 Omologo M, Svaizer P, Matassoni M. Environmental Conditions and Acoustic Transduction in Hands-Free Speech Recognition. Speech Communication, 1998, 25: 75～95
41 Elko G W. Microphone Arrays. In: HSC 2001, 2001. 11～14
42 Omologo M. Hands-Free Speech Recognition: Current Activities and Future Trends. In: HSC 2001, 2001. 23～26
43 Mangold H A. Realistic Hands-Free Applications - the Key for Really User-Friendly Applications. In: HSC 2001, 2001. 35～38
44 Gong Y. A Robust Continuous Speech Recognition System for Mobile Information Devices. In: HSC 2001, 2001. 39～42
45 Buchner H, Herbordt W, Kellermann W. An Efficient Combination of Multi-Channel Acoustic Echo Cancellation with A Beamforming Microphone Array. In: HSC 2001, 2001. 55～58
46 McCowan I A, Bourlard H. Microphone Array Post-Filter for Diffuse Noise Field. In: ICASSP 2002, 2002. 905～908
47 Fujimoto M, Ariki Y. Noise Robust Hands-Free Speech Recognition Using Microphone Array and Kalman Filter as Front-End System of Conversational TV. MMSP2002, 2002
48 Saruwatai H, Kajita S, Takeda K, et al. Speech Enhancement Using Nonlinear Microphone Array with Noise Adaptive Complementary Beamforming. In: ICASSP'00, 2000. 1049～1052
49 Acero A. Acoustical and Environmental Robustness in Automatic Speech Recognition: [Ph D thesis]. Carnegie Mellon University, 1990
50 Moreno P. Speech Recognition in Noisy Environments: [Ph D thesis]. Carnegie: Mellon University, 1996
51 Liu F H, Stern R M, Acero A, et al. Environment Normalization for Robust Speech using Direct Cepstral Comparison. In: Proc. of ICASSP 1994, Adelaide, Australia, April 1994. 61～64
52 Deng L, Acero A, Plumpe M, et al. Large-Vocabulary Speech Recognition under Adverse Acoustic Environments. In: Proc. of ICSLP 2000, Beijing, China, October 2000. 806～809
53 Deng L, Droppo J, Acero A. Recursive Estimation of Nonstationary Noise Using Iterative Stochastic Approximation for Robust Speech Recognition. IEEE Transactions on Speech and Audio Processing, 2003, 11(6)
54 Droppo J, Acero A, Deng L. Efficient On-Line Acoustic Environment Estimation for FCDCN in a Continuous Speech Recognition System. In: Proc. of ICASSP 2001, Salt Lake City, Utah, May 2001. 209～212
55 Droppo J, Acero A, Deng L. Uncertainty Decoding with SPLICE for Noise Robust Speech Recognition. In: Proc of ICASSP 2002, Orlando, Florida, May 2002. 57～60
56 Droppo J, Deng L, Acero A. Evaluation of the SPLICE Algorithm on the Aurora2 Database. In: Proc. of Eurospeech 2001, Aalborg, Denmark, September 2001. 217～220
57 Droppo J, Deng L, Acero A. Evaluation of SPLICE on the Aurora 2 and 3Tasks. In: Proc of ICSLP 2002, Denver, Colorado, September 2002. 29～32, 977～980

(上接第 163 页)

上　联：	两只黄鹂鸣翠柳；
HMM 模型生成的下联	九万紫燕舞红梅。
Trigram 模型生成的下联	三正紫燕舞晴空。

上　联：	爆竹一声除千户，
HMM 模型生成的下联	梅花千树新万里。
Trigram 模型生成的下联	梅花千色安万家。

上　联：	大江东去，浪淘尽千古英雄；
HMM 模型生成的下联：	江山春来，山河增百年盛世。

上　联：	九天揽月，华夏英豪驰宇宙；
HMM 模型生成的下联：	四海迎春，神州崛起舞天下。

上　联：	天王盖地虎；
HMM 模型生成的下联：	地焗成天牛。

从实验结果来看，用机器来应对联语的下联取得了比较满意的结果，生成的下联比较符合语义、语法，意境也大致切合。从两种模型的实验对比来看，基于 HMM 模型的对联语料机器学习和下联生成的时间都较长，Tri-gram 语言模型的机器学习时间和下联生成的时间都较短，但是 HMM 模型生成的下联明显比 Trigram 语言模型的质量好。

结论和展望　文学语言的处理与生成，是计算机艺术的重要内容，本文在对联应对生成方面所作的探索，再一次印证了机器学习在模拟人类智能方面的能力。下一阶段的研究，我们还将继续改进联语应对生成的方法，融合传统的对联创作方法，综合语义关系、语法关系、意境、主题等语言和文学的信息，并尝试用其他的机器学习方法来研究对联的应对生成问题。

参 考 文 献

1 Yi Yong, He Zhongshi, Li Liangyan. Studies on Traditional Chinese Poetry Style Identification. In: the Proc. of ICMLC04. Shanghai, 2004. 2936～2939
2 Christopher D. Manning, Foundation of Statistical Natural Language Processing. Pearson Education, 1998
3 Allen J. Natural Language Understanding, Second Esition. Pearson Education, Inc., 1995
4 Mitchell T M. Machine Learning, McGraw-Hill Companies, 1997
5 朱承平. 诗词格律教程. 暨南大学出版社, 1999