第十七卷 第六期
2001 年 12 月

信 号 处 理
SIGNAL PROCESSING

Vol.17. No.6
Dec. 2001

稳健语音识别技术发展现状及展望

姚文冰 姚天任 韩涛

（华中科技大学电子与信息工程系 武汉 430074）

【摘要】本文在简单叙述稳健语音识别技术产生的背景后，着重介绍了现阶段国内外有关稳健语音识别的主要技术、研究现状及未来发展方向。首先简述引起语音质量恶化、影响语音识别系统稳健性的干扰源及其影响。然后分别介绍语音增强、稳健语音特征的提取、基于特征和模型的补偿技术、麦克风阵列、基于人耳的听觉处理及听觉视觉双模态语音识别等技术路线及发展现状。最后讨论稳健语音识别技术未来的发展方向。

关键词：稳健语音识别 抗噪性

一、稳健语音识别的提出

随着计算机技术日新月异地发展，最近十五年中语音识别技术的研究有了实质性的突破，许多成功的语音识别系统相继问世。例如，Cardin 等研制的基于 TIDIGIT 数据库的非特定人连接数字语音识别系统，误识率仅为 0.5%；而 Das 等研制的 20,000 单词的特定人孤立词语音识别系统，误识率仅为 1%。据统计，现有的语音识别系统以每年 2" 的速度降低误识率。目前，这些系统部分或全部地克服了特定说话人、孤立词、小词汇量、有限语法这四个约束，达到了很高的识别率。更重要的是，这些系统中的绝大部分已经走出实验室成为商品。其中，最具代表性的当属 IBM 公司研制的 ViaVoice 大词汇量连续语音识别系统。它的误识率，在一定的环境下，可以低于 5%。

然而，大多数类似的系统都只适合于识别“干净”的语音，当它们应用于噪声环境中时，性能大大下降。Lockwood 等人发现传统的语音识别系统用“干净”语音训练，可达到 100%的识别率，而用时速 90 公里的汽车中的语音信号训练后，只能达到 70%的识别率。大量实验表明，大多数现有的非特定人语音识别系统，如果使用不同于训练时所用的麦克风或处于不同于训练时所处的外部环境时，即便是在安静的办公室内测试，性能都会严重下降。而对电话信号，汽车、工厂内或室外环境中的语音信号来说，现有识别系统的稳健性更差。

产生上述现象的主要原因在于语音信号在受到各种实际影响后而表现出的多变性，这些多变性包括：

1. 音素可变性：最小语音单位--音素的确定严重地依赖于上下文。比如，英语单词 *two, true* 中/t/需分别划分到音素/t/和/tr/中。单词边界的发音受上下文影响发生了变异。比如，*gas shortage* 往往发成 *gash shortage*。

2. 声学可变性：环境、声音传感器(麦克风或电话)的位置及传输特性的变化将导致语音的不同。

3. 说话人本身的可变性：说话人自己情绪、身体状况、语速、音质的变化导致语音的变化。

4. 说话人之间的可变性：说话人不同的社会背景、方言、声道形状、长短也会影响识别结果。

针对以上问题，1992 年美国国家自然科学基金的一次会议提出了有关语音识别技术的十大关键课题。其中语音识别的稳健性、识别系统的可移植性、识别系统对环境，说话人、麦克风的自适应及语言模型的建立等五个问题被放在了前五位。

语音识别的稳健性是指在输入语音质量退化，或语音的音素特性、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。其中声学特性(如声道、麦克风、电话特性)的差别及环境的差别是研究的重点，也有人将这一部分研究称为“抗噪声语音识别”。随着语音识别技术进入实际应用，稳健语音识别系统，即能在复杂且动态时变的环境中保持较好识别率的语音识别系统的开发变得越来越重要。

八十年代以来，基于隐马尔可夫模型(HMM)的统计模型匹配技术和动态搜索算法的应用使语音识别系统的研究迈上了一个新的台阶。然而，基于统计的声学模型和语言模型需要训练数据具有充分的代表性。当训练环境与测试环境失配时，由训练数据所得模板的代表性降低，识别系统的性能因而大幅度下降。虽然增大训练数据量，尽量覆盖所有失配的情形，可部分解决问题，但不是最终的解决方案。因此，稳健语音识别系统除稳健性之外的另一个重要目标就是降低对大量训练语音数据的依赖性，更有效地利用有限的数据，提取准确的统计模型以适应不同声学环境的变化。

二、主要技术路线及研究现状

稳健语音识别的研究早期曾一度受到语音增强技术的影响。然而语音增强虽然能够提高信噪比，但却不一定能够提高语音识别的识别率。稳健语音特征的研究伴随着语音识别的发展从未停止过。最近有很多稳健语音特征，例如 Mel 频段倒谱系数(MFCC)、感知线性预测(PLP)等都被证明具有很好的抗噪特性。但是，没有哪一种抗噪特征可以消除所有噪声的干扰，为了进一步提高识别系统的稳健性，人们又对训练和测试环境之间差异的补偿进行了深入地研究，并提出了各种特征及模型补偿技术。近年来，其它许多新的方法对上述技术构成了补充。例如，多麦克风阵列利用信号源和噪声源空间位置的不同来改善语音信噪比，以提高识别的稳健性。基于人耳听觉的信号处理方法通过对人耳听觉系统的仿真和观察，获得符合人耳听觉特性的语音特征表示。实验表明，基于人耳听觉的信号处理算法可以显著的改善语音识别系统的稳健性。除此之外，听觉视觉双模态语音识别(AVSR)的方法独辟蹊径，在听觉单模态语音识别系统的基础上，加入视觉子系统，摄取说话者的面部图像，从脸部主要特征中提取与发音有关的视觉特征，与声学特征一起作为识别器的输入。在噪声环境下，视觉信息的补偿对语音的感知性能有较大的改善。

下面在简单描述干扰源并分析其引起的训练与测试环境失配对识别系统造成的影响的基础上，对当前最流行的五类处理技术分别加以介绍，其中稳健语音特征的提取和特征及模型补偿技术是综述的重点。

1. 干扰源对识别系统的影响

影响语音准确识别的干扰源很多，最重要的两类是：未知加性噪声（例如各种机器、气流等引起的背景噪声、背景环境中其他说话人的干扰语音）和未知线性滤波效应（亦称为未知卷积噪声，例如房间内表面反射引起的回声、麦克风和说话人声道不同引起的语音谱形状的改变等）。其他干扰源包括语音信号受到的瞬时干扰（例如关闭房门或电话铃声产生的噪音），（由碳阻麦克风或电话系统中随机相位跳动引起的）语音非线性畸变，以及几个人同时说话造成的“串话”干扰等。到现在为止，稳健语音识别中大部分的研究工作针对的还是加性噪声和未知线性滤波效应干扰源。

由加性噪声和未知线性滤波效应引起的训练与测试环境的失配对识别系统的影响可从信号空间、特征空间和模型空间三个层次来分析，如图 1 所示。其中 **S** 是原始的训练语音，**X** 是从训练数据中提取出的语音特征，Λ_X 是根据训练数据得到的统计模型参数。类似的 **T**、**Y**、Λ_Y 分别是测试语音、测试语音特征和测试语音模型。当训练环境与测试环境失配时，干扰使 **T**、**Y**、Λ_Y 发生畸变，畸变影响用 **S**、**X**、Λ_X 到 **T**、**Y**、Λ_Y 的畸变函数 $D_1(\cdot)$、$D_2(\cdot)$、$D_3(\cdot)$来模拟。可以发现，前述的语音增强等除 AVSR 之外的稳健处理技术力图从信号空间、特征空间、模型空间三个层次消除畸变的影响。由于人耳的许多听觉特性被分别应用于稳健语音特征提取技术中，因此本文将

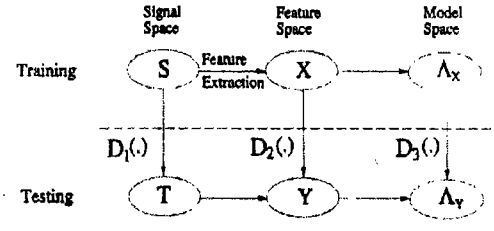

图 1　训练环境与测试环境失配时的影响

这一部分内容放在“稳健语音特征”一节中介绍，而在“基于人耳的听觉特性”一节中着重介绍有关人耳听觉模型的研究近况。

2. 语音增强

语音增强技术早期曾一度影响稳健语音识别研究。在语音识别系统中，它一般都作为预处理或前端处理(front-end processing)模块存在。语音增强以抑制背景噪声或保护和提高感知语音质量为目的，现有的语音增强技术大多在信号空间和特征空间两个层面进行处理。

传统的语音增强算法包括改进的 SNR 特征化(improved SNR characterization)，线性或非线性减谱算法，维纳滤波(Wiener filtering)等。其中，减谱算法和谱归一化技术在处理未知噪声和线性滤波干扰的道路上具有非常重要的影响。

Boll 为补偿加性噪声而提出的减谱（spectral subtraction）算法试图在没有语音的信号中估计加性噪声的功率谱，然后从实际语音信号的功率谱估计中将其减去，以增强识别的稳健性。Berouti 等进一步扩展了这个算法，主要通过“过度减去”噪声功率谱以达到避免“乐声”噪声的目的。最近， Lockwood 将非线性谱减（nonlinear spectral substrac- tion）引入汽车语音的稳健识别系统，Yoma 用一个 IIR 滤波器为加性噪声建模，减谱增强后，又将一个跟噪声功率的有关的阈值作为加权系数引入模型匹配过程。

Stockham 等提出的谱归一化（spectral norma- lization）技术，通常首先估计语音在训练和测试环境中的平均功率谱，然后用线性滤波器将测试语音功率谱“最优”地转换为训练语音功率谱。现在广为应用的未知线性滤波效应的补偿算法倒谱均值归一化(CMN)，虽然不直接变换测试语音的功率谱，但其直接强迫训练域和测试域中的倒谱系数的均值为零的做法显然也受到了谱归一化思想的影响。至今为止，减谱和谱归一化技术仍然受到广泛的关注。

传统的语音增强技术主要采用数学的方法，与语音的感知特性没有关系，近来许多语音增强算法利用了听觉特性（auditory information），如模仿噪声掩蔽（Noise masking）效应，当信号能量低于噪声能量时，令所有滤波器的输出等于噪声电平；将语音谱分成一些符合人耳听觉特性的子带，在每个子带中分别估计噪声特性和滤波，以增强汽车语音。

其他各种利用语音分类知识的有限迭带算法（constrained iterative methods）、变换域去噪算法，如小波去噪（在小波域内区分污染语音的清、浊音，分别用门限进行不同的处理，然后将去噪后的小波系数反变换，进行语音识别）等越来越多的语音增强新技术正在逐步应用到稳健语音识别中来。

语音增强技术最大的问题在于它们往往可以大幅度的提高信号的信噪比，但增强后的信号由于频率特性改变了，相对于识别器来说的最佳语音特征也往往被破坏了，因此采用语音增强技术并不一定能带来更高的识别率。

3. 稳健语音特征的研究

长期以来，稳健语音特征的研究伴随着语音识别的发展从未停止过。语音识别系统希望提取出稳健的语音特征，在干扰环境中尽可能地保留那些对识别有重要意义的信息，同时最大限度地摒弃那些无用的、冗余的信息，以便集中区别不同类的语音信息。利用语音特征的稳健性来抵抗噪声干扰的方法通常对噪声没有（或只有很弱）的假设条件，一般情况下不需要噪声统计特性的具体描述。稳健语音特征的研究包括以下两个方面：

- 稳健语音特征的提取
- 倒谱高通滤波

3.1 稳健语音特征的提取

相对于计算机而言，人耳的识别能力非常强，在噪声环境下，甚至是许多人说话的环境中都能够一下子

听到自己所关心的人或关心的话题的内容，因此人们期望利用人耳的听觉特性找到一些对噪声不敏感的语音特征。最近有很多稳健语音特征，例如 Mel 频段倒谱系数（MFCC：Mel-Scaled Frequency Cepstral Coefficients）、感知线性预测（PLP: Perceptual Linear Prediction）系数、符合时频掩蔽效应的掩蔽谱、调制谱（Modulation Spectrum）、听觉谱（Perceptual Spectrum）等等都被证明具有很好的抗噪特性。

MFCC 用 Mel 系数划分中频段，并采用了反映谱动态信息的一阶及二阶倒谱。PLP 模拟了人耳听觉感知 3 个方面的特点，即人耳听觉感知和频率的非线性关系、响度和频率的非线性关系、响度和声强的关系。

掩蔽谱利用了人耳的掩蔽效应，即一个很窄频带的刺激，消失后，还会对后续时间及周围频带的信号产生影响。掩蔽效应消失时间越长，掩蔽阈幅度越小，但掩蔽频带越宽。掩蔽阈可看作是一个相对时间、频率的可变函数。利用时间-频率掩蔽的掩蔽谱因此而得名。

目前的语音识别特征利用的大多为语音的短时谱信息。短时谱容易受说话人、环境噪声等的改变而改变。Kanedera 提出的调制谱反映了谱的长时相关特征。它是语音频谱的时间序列的频谱。在调制谱中 0-8Hz 的成分反映了一些重要的语义信息。实验表明，4Hz 的调制谱相对其他频率的调制谱对语音识别来说更为重要。

最近台湾科学家提出了一种听觉谱，模仿了人耳掩蔽效应、最小可听域（MAF：minimum audible field）及 Mel 系数的特性。在汉语元音识别实验中，听觉谱表现出比 MFCC、倒谱系数、反射系数更适应说话人的变化，对白噪声更不敏感，但识别准确率却不如后者。

3.2 倒谱高通滤波

倒谱高通滤波技术通过对语音特征进行高通滤波来改进稳健性，其计算消耗几乎为零。Hermansky在他著名的RASTA(RelActive SpecTrAl)处理中就采用了一个高通(或带通)滤波器对语音倒谱系数进行滤波。RASTA与PLP的结合，RASTA -PLP通过对压缩后的谱作高通或者带通滤波，然后将滤波后的谱采用指数函数变换回线性谱域。通常的压缩方法或者为对数压缩，或者为立方根压缩。由于在压缩后的频率域作了滤波，故而恢复到线性谱域的谱反映了原来谱中的变化信息。根据此线性谱求得的线性预测系数即为RASTA-PLP。实验表明RASTA-PLP比标准的PLP，线性预测系数LPC，或MPFCC的稳健性好。

在倒谱均值归一化(CMN)的方法中，高通滤波是以从输入倒谱系数中减去其短时均值的形式来完成的。RASTA 和 CMN 算法直接强迫训练域和测试域中的倒谱系数的均值为零，从而实现对未知线性滤波效应的补偿。倒谱高通滤波的方法经济有效，当前几乎所有的稳健语音识别系统中都有它的某种形式的存在.

4. 特征及模型补偿技术

许多情况下语音增强和稳健的语音特征并不能完全消除训练和测试环境失配的影响，补偿技术由此产生。它通常是在特征空间中修改测试语音的特征 $\mathbf{Y}$，使得测试语音的模型能够更加接近训练模型 $\Lambda_{\mathbf{X}}$；或反过来，动态修改训练模型的参数、结构，令所得到的补偿训练模型更加接近测试语音，如图 1 所示。这一类技术统称为特征及模型的补偿技术，可粗略地分为以下三类：

- 经验补偿技术；
- 盲补偿；
- 基于模型的补偿；

4.1 经验补偿技术

基于经验的补偿(亦称为基于训练的补偿)方案一般需要一组双声道立体声训练数据集。两个数据集是同时采集的，其中一个声道记录清晰无畸变干扰的高质量语音，另一个声道记录有畸变干扰的退化语音信号。通过比较两个训练数据集中语音的特征或模型的差异，利用经验设计补偿方法在识别过程中补偿训练和测试环境的差异，如图 3 所示。

Stern 等提出的经验倒谱补偿(empirical cepstral compensation)是最典型的例子。环境的变化和说话人的变化被看作对语音特征的加性扰动，为了消除这种加性扰动，一般采用一个经验校正矢量修正输入语音的语音特征或识别系统内语音模板的统计参数。若校正矢量能随着信噪比或特定信噪比下语音特征在特征空间内位置的变化而变化，则可进一步提高系统的稳健性。当事先无法确定测试环境时，可预先训练一个对应于不同测试条件的校正矢量集合，测试时通过比较找到最可能的那个校正矢量；若无法找到，则可通过对该集合内相邻元素的内插来构造一个新的校正矢量，从而完成补偿。

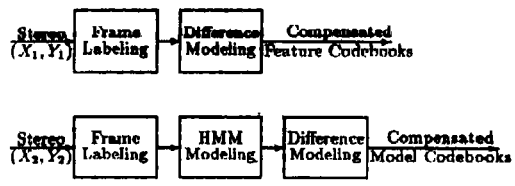

图 3 经验补偿算法原理示意

以Chien的自适应方差似然测量算法(var- iance adapted likelihood measure)为代表的仿真模型补偿(simulated model compensation) 对经验倒谱补偿构成了补充。它在无法获得立体声训练数据库的情况下，利用Monte Carlo算法根据清晰的高质量语音生成畸变的退化语音。

基于经验比较的补偿过程非常简单，并且在测试条件与校正矢量的某个训练条件非常相似时特别有效。在使用多个未知麦克风的 5000 单词的连续听写测试中，采用经验比较补偿方案的系统比采用倒谱均值归一化(CMN)的系统的误识率低大约 40%。然而，这种方法最大的问题在于，它需要在训练和测试环境下同时录得的立体声语音数据库。实际应用中，干扰源的数学仿真模型一般无法确定，使得采用 Monte Carlo 仿真数据的方法也无法实用。

4.2 盲补偿

盲补偿的方法与经验补偿法不同，它对于训练与测试环境的差异不需要准确的描述，并且摆脱了训练模型必须准确表示训练样本的前提。例如 1993 年 Merhav 和 Lee 提出的最小最大分类方法(minimax classification)假设测试是在与训练环境最不匹配的环境下完成的，并且认为对应于测试样本的最优模型参数 Λ'_X 和(或)最优测试特征 Y' 处于参数对(Λ_X,Y)的邻域内。通过同时修改最优判据和判决参数(Λ_X,Y)，该算法在低信噪比的环境下具有较好的补偿能力。1995 年 Moon 进一步发展了最小最大分类法，提出了盲特征补偿和递推特征补偿。

最小最大分类的原理如图 4 所示。假设测试语音模型 x 由概率密度函数 $p_\lambda(x)$，$\lambda \in \Lambda_i$，$i=1, \cdots, M$ 产生。λ 的邻域被分为 M 个区间，Λ_i 是其中第 i 个。同样的 x 的邻域也被分为 M 个区间，Ω_i 是其中第 i 个。假设最优模型参数随机分布在 x 的邻域中，故选取最不匹配的条件下最可能的参数是合理的，也就是说要找出最大错误概率最小时的参数值。最大错误概率可以表示如下：

$$p_\Omega(e) = \sum_{i=1}^{M} p_i \max_{\lambda \in \Omega_i} \int_{\Omega_i^c} p_\lambda(x)dx$$

其中 p_i 第 i 个 source 的先验概率。Ω_i^c 是 Ω_i 的补集。最小化最大错误概率可以通过最小化其最大值

$$\tilde{p}_\Omega(e) = \sum_{i=1}^{M} p_i \int_{\Omega_i^c} \max_{\lambda \in \Omega_i} p_\lambda(x)dx$$

来实现，等价于最大化下式：

$$\sum_{i=1}^{M} p_i \int_{\Omega_i} \max_{\lambda \in \Omega_i} p_\lambda(x)dx$$

故具体算法分为两步，首先用最大似然 (ML) 估计求出每个 Λ_i 内的最优参数，如 $\hat{\lambda}_i = \max_{\lambda \in \Lambda_i} p_\lambda(x)$，

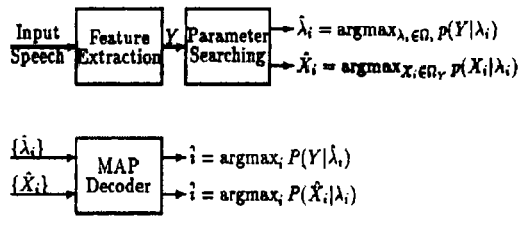

图 4 最小最大补偿原理

然后根据其结果，用最大后验概率(plug-in MAP)估计求出结果：

$$\hat{\Omega}_i = \left\{ x : p_i \max_{\lambda \in \Lambda_i} p_\lambda(x) = \max_{1 \le j \le M} \left[p_j \max_{\lambda \in \Lambda_j} p_\lambda(x) \right] \right\}$$

最小最大分类方法中采用的邻域搜索意味着它在分类过程中不仅采用了点估计 λ_i 还采用了区间估计 Λ_i，这使得它的判决函数更为稳健。同时，由于它仅仅对失配的状况作了一个最保守的假设，分类的性能只与所选邻域的形状大小有关，所以它适用的范围也非常广泛。然而，同样是这一点使得最小最大分类法的补偿性能无法与那些利用了部分失配先验知识的技术相比。而且，对于连续语音识别来说，模型参数所对应的邻域范围太大导致最优判决时的计算量将会非常大。另外，重叠的邻域还常常会引起最小最大分类的失败。

4.3 基于模型的补偿

目前，基于模型的补偿算法是补偿技术中最丰富的一类。与盲补偿对环境失配的保守假设不同，它利用先验知识为失配误差建立模型，然后通过该模型来进行补偿。这类方法一般同时调整失配误差模型的所有参数，由于参数很少，所以可以直接从给定的训练模型和测试数据中通过最优估计得到，而不需要进行额外的训练。下面分三类进行介绍。

· 特征补偿

Acero 提出的依赖于码字的倒谱归一化算法(codeword-dependent cepstral normalization)是一种早期的特征补偿技术。它选择混合高斯概率密度函数表示语音，并将环境的影响表示为

$$z = x + q + r(x,n,q)$$

其中 z 表示样本语音的倒谱矢量，x 表示“干净”语音的倒谱矢量，q 表示未知线性滤波效应，$r(x,n,q)$ 表示加性噪声。加性噪声和线性滤波效应对“污染”语音的倒谱来说是加性扰动，如何从倒谱矢量 z 中估计出 x 的问题，转化成了如何估计 q 和 $r(x,n,q)$的问题。具体算法中 q 和 $r(x,n,q)$是通过 ML 估计同时完成，x 的估计则通过最小均方误差(MMSE)估计完成。实验表明，依赖于码字的倒谱归一化算法在 SNR 比较高时效果较好。

Rahim 和 Huang 简化了上述模型，令 n=0，并引入矢量量化，在电话语音识别中获得了较好的效果。由于仅仅根据语音数据一般无法建立语音模型，现在广为使用的方法是采用离散、连续、半连续隐马尔科夫模型作为语音模型。Sankar 和 Lee 的统计匹配利用了 HMM 语音模型来补偿特征。另外许多仿射变换(affine transform)的方法也被用来进行非线性特征补偿。

· 基于预测的模型补偿

基于预测的模型补偿算法，假设训练模型一般都是由安静环境下高质量语音数据训练得到的，其主要的目的是将训练语音模型 Λ_X 转换成污染的测试语音的模型 Λ_Y。

早期曾通过往纯净训练数据上叠加测试环境中的噪声，然后重新训练模型来实现这个目标，这种方法与减谱法相比可以避免负功率的问题，但是却无法解决 Lombard 效应，而且不适用于大词汇量语音识别。为此人们提出用描述训练/测试环境差异的失配函数(mismatch function)将语音的训练模型和噪声模型合并，直接利用噪声信息补偿训练模型。事实上，这个失配函数就是图 1 中的 $D_2(\cdot)$或 $D_3(\cdot)$。Gales 和 Young 提出的并行模型联合(PMC: parallel model combination)算法，Minami 和 Furui 的模型组合(model composition)算法、Varga 和 Moore 的模型分解(model decomposition)算法都属于这个类型。

Gales 提出的并行模型合并算法在假设噪声不会影响路径的转移概率的前提下，对整个状态输出概率密度进行补偿。状态输出概率密度采用混合高斯概率密度函数来逼近。补偿则反映在均值与方差的调整上。在噪声合并过程中，纯净语音模型的均值与方差通过逆离散余弦变换从倒谱域变换到对数谱域，然后在线性谱域合并。合并后的均值与方差反映了污染语音的均值与方差。同样的过程反过来，再变回倒谱域，就得到了

污染的训练模型。并行模型合并算法对加性和卷积性噪声的影响都可进行补偿。

通常情况下，测试过程中能用的语音数据不多，所以失配函数的形式一般都很简单，多半是线性变换。Moreno 在他提出的矢量泰勒序列扩展(vector Taylor series expansion)中采用了更复杂的函数形式。许多失配函数还采用了非线性变换的形式，如多层感知机神经网络。

· 自适应模型补偿

预测模型补偿与自适应模型补偿的区别在于，前者是只使用测试语音的自适应过程，后者则是在真正测试前，用大量标注好的增量数据来自动提高训练模型性能的过程。如 Gauvain 和 Lee 提出的最大后验自适应(MAP adaptation)，Leggetter 和 Woodland 提出的最大似然线性回归(MLLR: maximum likelihood linear regression)等 HMM 参数自适应算法，Digalakis 和 Neumeyer 为说话人自适应提出的均值矢量的自适应仿射变换等。

还有用贝叶斯自适应(Bayesian adaptation)做参考模型和测试模型之间均值矢量域映射的矢量域平滑(vector field smoothing)算法，在说话人和环境的适应上有较好的效果。与贝叶斯自适应批量处理算法相对应的是，在线(online)或增量(incremental)自适应算法。Digalakis 用增量估计(incremental estimation)进行 HMM 参数的自适应估计，Huo 和 Lee 用增量自适应算法同时更新先验/后验分布的参数(hyperparameter)和 HMM 参数。这类算法最近引起了很多研究小组的关注。

传统的 HMM 参数估计假设不同 HMM 的参数之间相互独立。这样，若自适应数据的数量相对于不同模型不均等，就会造成有些模型训练得多，相对比较准确，有些模型训练得少，相对不那么准确的现象。为此人们试图在不同模型的参数之间建立某种联系，这样即便某个模型的训练数据不存在也能够同时一致地调整所有 HMM 模型的参数。有三类方法解决这个问题：1) 将每个自适应数据划入多个类，而不单是一类；2) 建立具备相关关系的模型，并估计模型的相关参数，例如 Hung 的 HMM 的平滑(Smoothing HMM)，随即轨迹模型(stochastic trajectory model)等；3) 在 HMM 的参数矢量之间引入相关(correlation)。

· 统计匹配

统计匹配(stochastic matching)的方法首先由 Sankar 和 Lee 提出。其主要思想是通过以下两种途径减少声学失配的影响：

1) 将畸变的测试特征 **Y** 变换成训练空间的特征 **X**，即，$\boldsymbol{X}=\boldsymbol{F}_{\nu}(\boldsymbol{Y})$，然后利用训练得到的模板 Λ_X 进行识别。或，

2) 将训练空间的模板 Λ_X 变换成由测试语音得到的模板 Λ_Y，即，$\boldsymbol{\Lambda}_{\boldsymbol{Y}}=\boldsymbol{G}_{\eta}(\boldsymbol{\Lambda}_{\boldsymbol{X}})$，以便于更好地与测试语音特征 **Y** 匹配。

统计匹配的算法是用具体的函数来逼近失配的影响，因此是比并行模型合并更为严格的补偿算法。补偿之前需要利用部分声学失配的先验知识确定映射函数 $\boldsymbol{F}_{\nu}()$、$\boldsymbol{G}_{\eta}()$的形式。给定 Λ_X，采用 EM(expectation maximum)算法即可递推地实现对 ν 或 η 的最大似然估计 ν' 或 η'，

$$\nu'=\max_{\nu} p(Y|\nu,\Lambda_X),\quad \eta'=\max_{\eta} p(Y|\eta,\Lambda_X)\cdot$$

通常为了计算简便起见，失配函数设为线性的。Surendran 和 Lee 提出采用非线性失配函数，用神经网络来逼近，并用扩展 EM 算法来估计神经网络的参数。

5. 麦克风阵列

近年来，针对汽车或房间内语音的识别系统广泛地采用了基于麦风阵列的语音增强处理。相对于单一麦克风来说，麦克风阵列由于在声场中不同位置放置多个对声音方向敏感的麦克风，有更多的信息可以利用。信号来自空间中不同的方位，即便频带相互覆盖交叠，也可以通过对阵列信号的时空滤波(spatiotemporal filtering)来增强或分离。与单麦克风稳健处理相比，基于麦克风阵列的稳健语音识别技术的研究的侧重于彩色

加性噪声抑制、回声抑制以及空间中不同位置说话人的语音竞争问题，在识别系统中，它主要起语音增强的作用。

传统的语音增强技术，如减谱算法，空间滤波(temporal filtering)，噪声消除(noise can- celing)在麦克风阵列处理中都有应用。在事先已知语音与噪声来向的前提下，通过调整各个麦克风信号的增益可以提高信噪比。基于最小均方误差（MMSE）的传统自适应滤波技术也被用来提高被加性独立噪声和回声干扰污染的语音的SNR。另外,通过基于互相关的算法还可以增强来自声场中特定角度的声音，但迄今为止，这种方法获得的性能的改善只比延迟求和(Delay and Sum)的方法稍好一点。阵列技术、麦克风的设计与位置摆放技术对回波干扰的补偿性能很好。比较麦克风阵列技术与单麦克风补偿技术对混响干扰的补偿性能，可以发现在麦克风阵列在低信噪比时对混响干扰的抑制明显优于单麦克风补偿技术，信噪比在0dB以下时，前者的系统识别率甚至比后者高一倍，达到80%-90%。若将两者结合起来，性能更优。

国内有关麦克风阵列(或称多话筒)语音处理的文献不多，98年以来只看到哈工大赵以宝等采用多话筒分别识别一个语音,然后用数据融合技术对识别结果进行处理的报道。

6. 基于人耳听觉的信号处理

随着听觉心理学和生理学研究的发展，基于人耳听觉的语音识别系统逐渐成为近年来研究的热点。许多学者对人耳听觉特性进行了深入的研究，并分别将其应用于语音处理，尤其是在提取符合人耳听觉特性语音特征和利用听觉特性的稳健前端处理中。表1归纳了常用的听觉生理、心理现象及对应的工程仿真算法。

表1　常用的听觉生理、心理现象及其仿真

名称	听觉生理心理现象	声音信号处理
外耳、中耳	具有外耳的音响系，中耳的机械系频率传输特性	在2kHz-3kHz频段上作频率预加重
基底膜	掩蔽	噪声掩蔽
	临界频带	Bark(或Mel)频率刻度上等间隔并行临界带宽滤波器
	声音响度	符合Phon近似对数变换关系的响度曲线
内耳毛细胞	仅具有基底膜的振动的半周期上的兴奋极性	半波整流 内耳毛短时自适应
听觉神经	频率侧抑制特性	声音频谱的强调音调

表2　主要的人耳听觉模型

表2提出者	特　征	实　验
Deng & Geistler (1987) Wisconsin Univ.	模仿中耳、基底膜运动、内耳毛细胞突触特性的建模，依照神经网络的通道间的相互关系而表现类似频谱 频宽: 0-800Hz 通道数: 1400	在CV音节上神经脉冲的时间模式的分布其生理数据一致
Ghitza (1988) AT&T Bell Lab	带通滤波器输出和设定阈值的交错作用而模拟神经放电，求取交错的时间间隔的直方图(EIH). 频宽: 200-3200Hz 通道数: 85	识别英语39个单词，作为识别系统的前端处理模块。输入语音SNR较低时，识别效果较好。
Seneff (1988) MIT	听觉神经的同期性和平均放电频率相结合的模型，由同步检测器(GSD: Generalized Synchronize Detector)和振幅包络检测器构成 频宽: 140-6400Hz 通道数: 40	各种生理数据一致 在频谱图所表示的话音素频率的分解度高
Shamma (1988) Maryland Univ.	依HPF及Sigmond而实现其饱和性，向各个通道进行2种的侧抑制，以侧抑制网路而进行特征频率锐化 128通道的基底膜模型	母音及破裂音的识别特征的表现

与此同时，完整的人耳听觉模型的建立和完善推动着基于人耳听觉的语音识别系统的研究逐渐成为一个相对独立的研究领域。它模仿人类听觉生理和心理机制建立听觉模型,对语音进行预处理，一般作为自动语音识别系统的前端处理模块。表 2 列出了最近十几年以来影响最广的四类听觉神经模型。这些模型都以带通滤波器的输出信号所具有的周期性或同步性为前提，包括一组模仿人耳耳蜗的临界频率带通滤波器，和紧接其后的模仿内耳毛传导、侧抑制等作用的通道/相邻通道的非线性处理器。

近年来，这四种模型得到了进一步的研究和发展。在 Ghitza EIH 模型的基础上，引入了峰值检测和非线性处理(compressive nonlinearity)。利用小波表示听觉模型的方法在中也有深入的介绍。我国科学家在人耳听觉模型的研究方面也做了大量的工作，但尚无突破性的成果出现。

最近的报道表明，当输入语音质量下降或训练测试条件失配时，基于听觉模型的识别系统的确比采用倒谱特征的识别系统的识别率高。但比起现在效果最好的动态自适应或模型补偿技术来说，则不行。而且它的计算量也比后者要大。当然，这也有可能是因为隐马尔柯夫模型（HMM）分类器不适合听觉模型所提取出来的语音特征的原因造成的。总之，虽然大多数研究人员都认为基于人耳听觉的语音处理是个非常有前途的研究方向，但迄今为止，尚无法证明它的识别效果最优。

7. 听觉视觉双模态语音识别

自从 1984 年 Petajan 开拓性地将视觉信息引入语音识别研究后，听觉视觉双模态语音识别(AVSR)技术逐渐吸引了众多研究者的注意，并逐步形成了语音识别技术的一个分支。

它的生物物理学基础在于：视觉信息在声学环境恶劣时，可以作为声学信息的补充；在良好的声学环境下，也有助于对言语的识别。视觉信息对人的言语感知的贡献为如下三类：引起注意、冗余和补充。在平常的生活中，大多数人都有类似的体验，在对语音的听觉感知存在障碍，如听力受损、环境噪声太大时，常常自觉不自觉地以视觉感知(如说话人的唇形)作为补充，这样对语音的识别能力就会增强。听觉有障碍的人士，如耳聋患者，常常需要通过唇读来帮助自己对语言的理解。有些人甚至只通过唇读就可以完美地理解句子的意思。可见，在语音识别系统中正确地引入视觉信息，将提高识别系统的稳健性。

一般来说 AVSR 系统包括视觉子系统和听觉子系统。在视觉处理子系统中，摄像机获取说话者的图像，然后进行数字化和图像处理，以得到语音识别用的视觉特征。同时语音经麦克风录入，经数字采样后得到听觉特征。最后，识别系统综合听觉和视觉两个子系统的数据进行分类识别。虽然双模态识别系统与传统的听觉单模态识别系统有相似之处，但前者的研究重点放在视觉特征的提取、融合策略及识别算法的研究上。

视觉特征的提取主要分为：基于像素的方法和基于模型的方法。

基于像素的方法，是将整幅原始图像，或者经过某种图像变换的变化域图像，作为语音的视觉特征。基于像素的方法具有较好的识别率和较高的稳健性。其缺点是，图像数据量太大，所以许多基于像素的方法采用了降低维数的方法，但尽管如此，得到的特征矢量维数仍然很高。另外，基于像素方法对于光照变化的稳健性很差。基于模型的方法，用少量的参数表示提取出来的主要发音器官如唇、下颌的轮廓，并将其作为特征矢量送入识别器。其优点是，特征矢量维数低，并且对平移、旋转、光照等变化具有移不变性，因而识别的速率快、稳健性好。但究竟哪些参数与语音的区别密切相关，目前还不是特别清楚，现有系统中采用的参数也不全相同。而且，轮廓的提取和跟踪算法实现复杂，其稳定性也易于受到图像质量的影响。一旦轮廓的定位跟踪错位，识别时将产生不可恢复的错误。

判决融合策略是近来 AVSR 研究中的另一热点。所谓判决融合，就是将来自声学和视觉两个通道的信息结合到一起，对音子进行分类判决。由于声学信息和视觉信息来自不同的通道，他们反映的问题本质不尽相同，时间先后上不完全同步(内在的)，所受的噪声干扰也不相同，因此需要一个判决融合系统来进行分类。判决融合系统的结构可分为数据到数据(早期融合)、判决到判决(晚期融合)、数据到判决三种。

国外文献报道，在高斯白噪声污染的孤立元音识别的实验中，AVSR 系统的抗噪能力比单纯的语音识别系统好 6dB-12dB。汉语元音音素的口形识别的正确率可达 80%以上。

国外关于听觉视觉双模态语音识别技术的研究已经开展了多年，但由于其涉及到图像处理、理解技术及不同信息源的融合等难点，导致目前的研究尚处于初级阶段。我国中科院声学所语音交互技术中心、浙江大学、哈工大都已开始从事这方面的研究，并取得了一定的进展。

三、未来发展方向

尽管稳健语音识别技术极其重要，但它也只是在近几年才发展起来的重要研究领域。目前，环境自适应技术还只在一个相对较窄的应用领域（比如，在半稳态加性噪声和/或线性滤波环境中，或可获得大量“特定环境”中训练数据的情况下）获得了成功。语音识别系统在识别非母语说话人时，效果还是非常差，即便使用说话人自适应技术也没有多大改观。前面提到的语音识别所面临的几大关键课题迄今为止都没有得到满意的解决。只有成功的解决这些关键问题，语音识别技术才能真正进入实际应用。具体说，以下几个方面的问题是今后研究的重点。

电话语音识别

电话语音识别比较困难的主要原因在于电话线路有其独特的信噪比和频率响应。除此之外，电话语音还受到瞬时干扰和非线性畸变的影响。因此，电话语音识别系统必须能够自适应新的信道和训练样本很少这样的特点。

低信噪比环境语音识别

根据现在的补偿技术，语音识别系统在 SNR 低于 15dB 的环境中，识别率会大幅下降。有趣的是，人类本身在更低的 SNR 环境下，仍然能够准确的识别语音。

语音串话干扰

相对于宽带噪声干扰来说，串话干扰是对稳健语音识别的一个更大的挑战。然而，到现在为止通过提取说话人特定语音特征来降低串话干扰的效果却非常不成功。

非母语说话人的快速自适应

在现在这个地域限制越来越弱、流动性越来越强的社会，成功的语音识别系统必须能够同时处理母语和非母语说话人的语音。并且随着语音识别系统的一步步商品化，对非母语说话人口音的快速自适应也必须得到进一步的发展。

真实环境语料库的建立

稳健语音识别技术的持续快速发展必须依靠包含真实环境中各种语音数据的语料库的收集、整理、发布。只有通过理论研究者、系统开发人员和最终用户的通力协作才能选择出各种具有代表性的环境、各个领域的具有代表性的语句从而得到一个可以广为应用的实用的语料库。

Development and Prospect of Robust Speech Recognition

Yao Wenbing　Yao Tianren　Han Tao

(Dept. of Elec. & Info. Eng., Huazhong University of Sci. & Tech., Wuhan 430074)

Abstract: This paper introduces the backgroud of the proposal of robust speech recognition, compares several differenct approaches to robust automatic speech recognition and the prospect future direction of robust speech

（下转第 497 页）

结果来看，利用镜像频谱来解调是切实可行的。另外系统实验结果也从实践上证明了直接中频采样、数字 I/Q 正交解调较之模拟正交解调具有很大的优越性。本系统的研制为中频数字接收技术应用于实际的系统作出了有益的尝试，本系统可以应用于移动通信基站和窄带雷达系统中，以取代此类系统中的模拟零中频处理部分，从而提高现有系统的性能。

参考文献

[1] Sophocles J.Orfanidis, Introduction to Signal Processing, Prentice-Hall, 1996

[2] 曾志刚、钱亚生，现代通信原理，清华大学出版社，1992 年 8 月

[3] 杨小牛、楼才义、徐建良，软件无线电原理与应用，电子工业出版社，2001 年 1 月

[4] 张健，软件无线电的基本理论构架，电子科技大学博士后研究报告，2000 年 10 月

[5] W.M.Waters, B.R.Jarret, Bandpass Signal Sampling and Coherent Detection, IEEE Trans. On AES, Vol.AES-18, NO.4, Nov 1982: 731-736

[6] HSP50214B Data Sheet, Harris Semiconductor, 1999

A Kind of IF Digitized Receiver Based on Software Radio Technology

Lu youxin　Lei Ting　Zheng Ligang　Xiang Jingcheng

(College of Electronic Engineering,UESTC,Chengdu　610054)

Abstract: This paper discussed a kind of IF（Intermediate Frequency） digitized receiver based on software radio technology. A new method of using image spectrum to demodulate signal is applied in the system realization, thus we can demodulate the signal which is modulated at high frequency in lower frequency. The experiment' sresults show that the performance of the system is excellent.

Key words: Software Radio　Image spectrum　IF digitized receiver

（上接第 493 页）

recognitioin. The ongoing research in the use of acoustical pre-processing to achieve robust speech recognition is firstly reviewed. Approaches based on direct cepstral comparisons, on parametric models of environmental degradation, and on cepstral high-pass filtering are emphatically discussed and compared. The effectiveness of two complementary methods of signal processing for robust speech recognition: microphone array processing and the use of physiologically- motivated models of peripheral auditory processing are also described.

Key words: Robust speech recognition　Noise robust

通　知

本刊从 2002 年开始已由邮局发行，邮发代号为：18－143，读者可直接到全国各地邮局办理订阅手续，也可到本编辑部订阅。