文章编号:1007-0249(2001) 03-0051-07

HMM 在说话人识别中的应用*

赵力, 邹采荣, 吴镇扬

(东南大学 无线电工程系,江苏 南京 210096)

摘要: 本文介绍了隐马尔可夫模型在自动说话人识别中的应用,指出了目前说话人识别技术中存在的一些问题和今后需要研究的课题。

关键词: 自动说话人识别;隐马尔可夫模型;语音

中图分类号:TN912.34 文献标识码:A

1 引言

自动说话人识别(Automatic Speaker Recognition :ASR),很久以来就是一个既有巨大吸引力而又有相当困难的课题。说话人识别技术按其最终完成的任务可以分为 2 类:自动说活人确认(Automatic Speaker Verification :ASV)和自动说话人辨认(Automatic Speaker Identification :ASI)。本质上它们都是根据说话人所说的测试语句或关键词,从中提取与说话人本人特征有关的信息,再与存储的参考模型比较,做出正确的判断。不过 ASV 是确认一个人的身份,只涉及一个特定的参考模型和待识别模式之间的比较,系统只做出是或不是的 2 元判决;而对于 ASI 系统则必须辨认出待识别的语音是来自待考察的 N 个人中的哪一个,有时还要对这 N 个人以外的语音做出拒绝的判别。由于需要 N 次比较和判决,所以 ASI 的误识率要大于 ASV,并且随着 N 的增加,其性能将下降。此外,按被输入的测试语音来分,还可将说话人识别分为 3 类,即与文本(Text)无关、与文本有关和文本指定型。前 2 类,一种是不规定说话内容的说话人识别,另一种是规定内容的说话人识别。然而光有这 2 种类型是不完全的,因为如果设法事先用录音装置把说话人本人的讲话内容记录下来,然后用于识别,则往往有被识别装置误接受的危险。而在指定文本型说话人识别中,每一次识别时必须先由识别装置向说话人指定需发音的文本内容,只有在系统确认说话人对指定文本内容正确发音时才可以被接受,这样做可以防止本人的语声被盗用[1]。

说话人的可识别性是基于客观存在的事实。由于各说话人发音器官的生理差异以及后天形成的行为差异,每个人的语声都带有强烈的个人色彩。而且自动说话人识别在相当广泛的领域内可发挥重要作用。如保安领域(如机密场所入门控制)、公安司法领域(如罪犯监听与鉴别)、军事领域(如战场环境监听及指挥员鉴别)、财经领域(如自动转账与出纳等)、信息服务领域(如自动信息检索或电子商务)等等。正因为如此,人们对于该领域的研究始终坚持不懈。虽然面临许多困难,自动说话人识别还是取得了相当的进展,并且在某些领域的实用化取得了一定的成功。近年来,隐马尔可夫模型(HMM)在语音信号处理上得到了广泛应用。由于 HMM 既能用短时模型-状态解决声学特性相对稳定段的描述,又能用状态转移规律刻画平衡段之间的时变过程,所以能统计地吸收发音的声学特性和时间上的变动。因此,在现有多种不同的说话人识别模型中,HMM 已成为目前最佳的说话人识别处理模型。在与文本有关的说话人识别中,最好的结果是用连续 HMM(CHMM)对说话人特征建模而取得的[2]。对于与文本无关任务来说,HMM 模型的瞬态结果是不需要的,所以各态历经 HMM(Ergodic HMM)常被使用。另外,1 状态 CHMM(也称高斯混和模型:GMM)也被广泛用作说话人模型并且取得了比多状态 CHMM 更好的结果[3]。和以语音内容时间序列为识别对象的语音识别不同,说话人识别主要是识别说话人特征,因此怎样的特征参数以及 HMM 结构可以最佳地表现说话人信息是研究的

* 收稿日期:2001-01-08 修订日期:2001-05-16

重点，而且说话人信息特征对时间变化的敏感性以及在说话人识别系统中模型训练数据不可能很多等原因，使得说话人识别 HMM 中的研究课题和语音识别有所不同。本文将介绍一些近年来国内外研究出的利用 HMM 进行说话人识别的新方法，以及对今后的发展作一些讨论。

2 与文本有关、无关和文本指定型说话人识别方法

首先介绍一下利用 HMM 实现的与文本有关、与文本无关和文本指定型说话人识别系统的构成和基本原理。

2.1 与文本有关的说话人识别

基于 HMM 的与文本有关的说话人识别系统的结构如图（1）所示。建立和应用这一系统有两个阶段，即训练（登录）阶段和识别阶段。在训练阶段，针对各使用人对规定语句或关键词的发音进行特征分析，提取说话人语音特征矢量（例如倒谱及Δ倒谱等）的时间序列。然后利用从左到右 HMM(left-to-right HMM)建立这些时间序列的声学模型。因为文本是固定的，所以特征矢量的时间构造是确定的，利用从左到右 HMM 能较好地反应特征矢量时间构造特性。在识别阶段，先和训练阶段一样，从输入语音信号中提取特征矢量的时间序列，然后利用 HMM 计算该输入序列的生成概率，并且根据一定的相似性准则来判定识别结果。对于说话人辨认系统，所得概率值最大的参考模型所对应的使用者被辨认为是发音的说话人。对于说话人确认系统，则把所得概率值与阈值相比较，其值大于（或等于）阈值的，作为本人的声音被接受，小于阈值的作为他人的声音被拒绝。

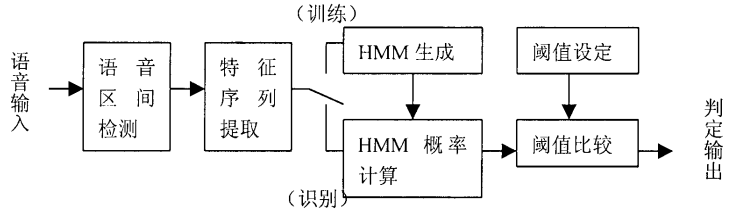

图 1 与文本有关的说话人识别系统构造

在与文本有关的说话人识别当中，由于文本内容是已知的，所以即使利用比较短的语料，也能从中提取出较稳定的说话人特征。而且学习也不需要太多的数据。在实际利用电话语音的说话人识别实验中得到了较高的识别精度[4]。另外，对于不同的说话人，变换文本内容并利用文本内容的差别，也可以进一步提高识别精度。

2.2 与文本无关的说话人识别

对于与文本无关的说话人识别，因为文本内容是不确定的，所以一般需采用各态历经 HMM 建立说话人模型。在学习阶段，对于说话人的各种文本发音提取其特征序列建立模型，HMM 的状态数一般取 5 状态左右。各状态采用混合高斯密度分布，混和数一般取 64 分布左右。识别时和学习阶段一样，例如对于说话人确认识别，先从输入语音中提取特征序列，然后利用本人的 HMM 计算输入特征矢量时间序列的概率值，通过和阈值相比较，判决识别的结果。

除了采用各态历经 HMM 以外，也可以考虑其他结构类型，例如作为两种结构 HMM 的折中，文献[5]利用只有一个状态的混合高斯分布连续 HMM 进行识别，得到了 97.9%的说话人确认率（1 状态 64 混合分布）。而且通过实验表明，在各态历经 HMM 中，状态间的转移对说话人差别的表现贡献不大。所以，增加状态数和增加状态的混合分布数基本上具有同样的效果，识别性能基本上是决定于两者的乘积。另外，利用从左到右 HMM 建立各说话人的基元模型集(音素或音节等基元)，然后在识别搜索中，利用基元模型的自动连接进行识别，也可以得到较好的识别结果。文献[6]报道了利用该方法进行说话人确认识别，得到了 97.1%的说话人确认率。而在同样的实验条件下，利用 1 状态 32 混合分布 HMM 建立说话人模型，仅得到了 91.0%的说话人确认率。

2.3 指定文本型说话人识别

指定文本型说话人识别系统的基本构造如图（2）所示。在该系统中，系统不仅要判别是否是本人的发音，而且还要判定是否是本人所发的指定内容的语音。为了使系统能够随时更换指定文本内容，所以一般系统是以各说话人的基元模型为基本模型，然后由基元模型的连结组成指定文本内容的模型。

在训练基元模型时，为了达到利用有限的说话人发音语料使训练的模型能较好地保持说话人的个人特性，一般是先利用多数说话人的语料训练的非特定说话人基元模型作为初始模型，然后由各说话人的训练语料对初始模型进行自适应训练而得到各说话人的基元模型。并且由于说话人识别系统的自适应训练语料有限，所以在自适应训练时一般仅对混和分布的分歧系数和各高斯函数的均值向量进行重估，协方差矩阵参数则保持不变。在识别阶段，根据系统指定的文本内容，由本人基元模型的连结，组成文本模型，然后利用本人的指定文本模型和输入语音时间序列进行匹配，计算由该模型生成的概率值，并把概率值和阈值比较，进行说话人确认判决。

文献[7]报告了 AT&T 的 Setlur 等人为银行系统研制的 ATM（Automatic Teller Machine）用指定文本型说话人识别系统。指定的文本类型是 4 位数字。基元模型是数字单位的从左到右 CHMM（状态数 10，混和高斯分布数 4）。识别时用户根据 ATM 装置随机指定的 4 位数字发音，识别装置先根据用户申告的话者名，由该说话人的基元模型连结组成文本模型，然后该 4 位数字模型和输入语音进行匹配计算输出概率，并通过和阈值比较进行判决。登录话者 50 名、假冒者 195 名。利用有 6 个月时间差、带宽 0-4KHZ 的语料。采用 12 阶 LPC 倒谱和 Δ 倒谱参数。对于 36 种类的 4 位数字语音的说话人确认率是 98.5%。

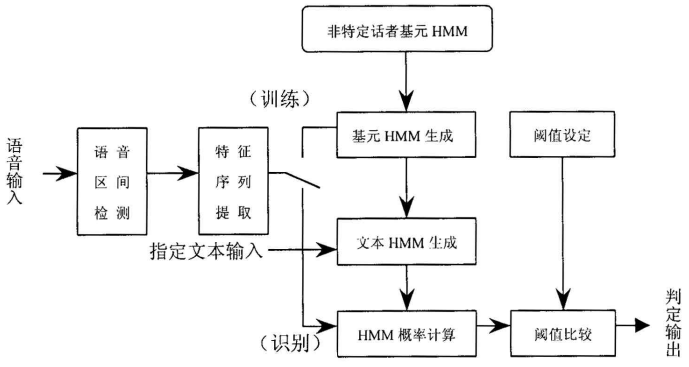

图 2　指定文本型说话人识别系统构造

3　基于 HMM 的说话人识别技术

近年来，在基于 HMM 的说话人识别方面已经取得了许多研究成果。以下结合研究例子，介绍基于 HMM 的说话人识别中的一些新的技术。

3.1　说话人识别所用的特征参数

虽然哪些参数能较好地反映说话人个人特征，现在还没有完全搞清楚，但一般都包含在两个方面，即生成语音的发音器官的差异（先天的）和发音器官发音时动作的差异（后天的）。前者主要表现在语音的频率结构上，主要包含了反映声道共振与反共振特性的频谱包络特征信息和反映声带振动等音源特性的频谱细节构造特征信息。代表性的特征参数有倒谱和基音参数。后者的发音习惯差异主要表现在语音的频率结构的时间变化上，主要包含了特征参数的动态特性，代表性的特征参数是倒谱和基音的线性回归系数 Δ 倒谱和 Δ 基音参数。在说话人识别中，频谱包络特征特别是倒谱特征用得比较多，这是因为一些实验已经证明，用倒谱特征可以得到比较好的识别性能，而且稳定的倒谱系数比较容易提取。和倒谱相比，基音特征只存在于浊音部分，而且准确稳定的基音特征较难提取[4]。

一般来说，人能从声音的音色、频高、能量的大小等各种信息中知觉说话人的个人性。所以可以想象，如果利用复数特征的有效组合，可以得到比较稳定的识别性能。文献[8]利用倒谱特征和可靠性高的区间的基音特征的有效组合进行识别实验，首先对于浊音部、清音部、无音部分别进行编码，在浊音部用倒谱、Δ 倒谱、基音、Δ 基音，在其他区间用倒谱和 Δ 倒谱作为识别特征，然后利用两部分的概率加权值和阈值进行比较。对于 90 名说话人识别系统的说话人确认率达到 98.7%。文献[9]利用提出的动态 HMM 来反映说话人特征的动态特性，对于 10 名说话人识别系统的说话人辨认率达到 96.0%。另外，文献[10]的识别实验证明了，对于与文本有关的说话人识别系统，利用动态特征和静态特征的组合，可以得到比较好的识别结果。而对于与文本无关的说话人识别系统，使用动态特征作为识别特征，并不一定得到好的效果。所以，对于动态特征的有效利用还需要进一步研究探讨。

在说话人识别中，一般电话带宽（0～3KHZ 左右）或者 0～6KHZ 频宽范围内的语声信息用的比较多，比这更高的频带区域内的个人信息利用的研究很少。一般认为，高频区域内的语音频谱能量比

较小，有用的信息比较少。文献[11]通过利用 0～16KHZ 带宽内的特征的说话人识别实验，分析了高频区域对说话人识别的贡献度。结果表明，在至今一直不受重视的语音高频区域的特征，确实存在有用的说话人信息，并且这些信息对于发音时间的变化以及加性噪声都比较稳定。特别地，如果特征中包含了 F 比较大的 5KHZ 附近的频域信息，则对于发音时间差的变动，识别性能更稳定。同时，如果在说话人识别中把高频区域和低频区域的信息有效地进行合理组合，则可以改善识别性能。由于全极点声道模型的最小相位特性，LPC 倒谱系数 C_k 至少以 $\frac{1}{k}$ 的速度衰减，幅度大的倒谱系数集中在 $k=0$ 附近，这样在欧氏距离准则下，低阶的倒谱参数对最终的距离贡献较大，而带有很多说话人信息的高阶的倒谱不能得到很好的体现，所以必须考虑倒谱参数的加权效果，这可以在基于 HMM 的说话人识别中,通过对各 HMM 状态中的输出分布函数的各成分的方差进行加权处理来实现[12]。

说话人识别参数的时间变化对识别率的影响也是一个重要问题。最早提出这一点的是 Luck[13]，日本的古井等人对这一问题做了较系统的研究[14]，他们的结论是，3 周以内基本上没变化；一个月以后开始变化，到 3 个月确认率和辨认率分别下降了 10%和 25%左右；3 个月以后识别率的下降开始变缓，基本上没有太大的劣化。识别参数的时间变化主要是音源特性的变化引起的。可以把音源和声道分离，只用后者组成经得起语音长期变动的说话人识别系统。在基于 HMM 的说话人识别系统中，可以采用类似于非特定人识别的学习方法，采集各个时期的语音数据来训练模型，也可以采用自适应方法，随时更新模型参数。

3.2 说话人识别中判别方法和阈值的选择

对于要求快速处理的说话人确认系统，可以采用多门限判决和预分类技术来达到加快系统响应时间而又不降低确认率的效果。多门限判决相当于一种序贯判决方法，它使用多个门限来作出接受还是拒绝的判决。例如，用两个门限把距离分为三段：如果测试语音与模版的距离低于第一门限，则接受；高于第二门限，则拒绝；若距离处于这两个门限之间，则系统要求补充更多的输入语句再进行更精细的判决。这种方法允许使用短的初始测试文本，系统就能最快的作出响应，而只有当模版匹配出现模糊时才需要较长的测试语音来帮助识别。同样，预分类是从另一个角度来加快系统响应的时间，在说话人辨认时，每个人的模版都要被检查一便，所以系统的响应时间一般随待识别的人数线性增加，但是如果按照某些特征参数预先地将待识别的人聚成几类（如：以平均音调周期的长短来分类等），那么在识别时，根据测试语音的类别，只要用该类的一组候选人的模版参数匹配，就可以大大减少模版匹配所需的次数和时间。在说话人识别实际应用中，有时还要考虑依照方言和某些韵律等超音段特征来预分类。

门限的设定对说话人确认系统来说很重要。比如，门限设得太高了，有可能把真正的说话人拒绝；太低了，又有可能接受假冒者。在说话人确认系统中，确认错误由误拒率（FR）和误受率（FA）来表示。通常由这些错误率决定对门限的估计，这时门限一般由 FR 和 FA 的相等点附近来确定。可以先对正确者和错误者的得分一起排序，然后找到一个点，在这点上，错误者和正确者的得分正好相等。通常，每一说话人的数据都很少，因此，说话人门限确定的统计性不太明显。这就是为何对小数据来说，使用的是全局门限（对每人都一样）的缘故。必须注意，FA 和 FR 都是门限的离散函数，点的个数决定于对真实者的 FR 测试和假冒者的 FR 测试次数。很明显，如果两者的测试点相等，FA 和 FR 会在某一点相交。然而在实际实验中通常假冒者要比真实者多许多，因此用上面的方法，我们会发现 FR 和 FA 不会相等，但会接近。此时，一些实验就将此接近点当作门限。文献[3]对此作了比较详细分析，表明更精确的门限可以由 FR 和 FA 的线性近似函数得到。另外说话人确认是一个 2 值问题,只需判定是否是由申请者所讲即可，在经典的解决方案中,判定是由对申请者模型的语句得分与某一事先确定的门限比较而得到的。这种方案的问题是得分的绝对值并不只是由使用模型决定的，而且还与文本内容以及发音时间的差别有关,所以不能采用静态的门限。文献[15]利用 HMM 输出概率值归一化方法解决这一问题，实验证明可以明显地提高确认率。

3.3　说话人识别 HMM 的学习方法

由于在说话人识别系统中，登录的说话人的发音数据都比较少，所以用于各说话人 HMM 训练的语料就少，给 HMM 的学习带来一定困难。为了建立各说话人的高精度的模型，已有各种说话人 HMM 的学习方法被提出。这里介绍一下文献[16，17]提出的 2 种类型的模型训练方法，一种是仅利用少量的登录说话人学习数据的学习方法；另一种是利用非特定人语音 HMM 和登录说话人学习数据的学习方法。在第一类型学习方法中，首先利用说话人的所有发音数据建立一个和基元类别无关的话者 HMM，然后以此为初始模型，根据各说话人的训练语音文本内容，利用连接学习法，仅仅对各高斯分布的权值进行再推定，而均值和方差不变。因为参加学习的数据少，所以基元模型不能分得太细，例如在汉语中，多元音素不要作为一个基元类别，而只考虑单元音，多元音可由单元音组成。第二类型学习方法是利用非特定人基元 HMM 和各话者 HMM 进行组合的方法。例如设非特定人基元 HMM 是 3 状态 4 混合高斯分布 CHMM，话者 HMM 是 1 状态 64 混合高斯分布 CHMM。先利用非特定人 HMM 收集对应于每一状态的学习数据，利用这些数据对说话人 HMM 的各高斯分布的权值进行再推定，并且把每一状态和相应的推定后的话者 HMM 置换，得到各话者基元 HMM（3 状态 64 混合高斯分布 CHMM），转移概率保持不变。然后以此作为初始模型，根据各说话人的发音文本内容，利用连接学习法，再一次对各话者基元 HMM 的各高斯分布的权值进行再推定。以上仅仅对高斯分布的权值进行再推定是为了保留说话人特性不变。

3.4　鲁棒的说话人识别技术

鲁棒的说话人识别技术一直是一个很重要的研究课题，并且已经有许多研究成果被提出。例如，对于由信号传输信道、滤波器等引起的识别率下降，通过倒谱均值正规化法（Cepstrum Mean Normalization：CMN）可以得到较大的改善[18，19]；由声道特征、发音方式的时间变动等引起的识别率下降，可以通过似然度（或概率）正规化法加以改善[20]。文献[21]比较了对于 LPC 倒谱、MFCC、LPC 美尔倒谱特征参数用上述 CMN 和似然度正规化法的改善程度，结论是 MFCC 的改善最明显；文献[22]利用说话人部分空间影射方法进行语音中语义内容和说话人个人性的分离，提取只含有说话人个人信息的特征进行说话人识别，利用很少的数据得到了较好的识别效果。文献[5]把鲁棒的距离尺度 DIM（Distortion-Intersection Measure）应用于说话人识别 HMM，把 HMM 的各高斯分布的两端用一定值（如 $3s$ ）平滑，结果能较好地吸收特征参数的变动。

HMM 对噪声的鲁棒性较低，所以很多在实验室里具有很好识别性能的基于 HMM 的说话人识别系统，在实环境下识别性能会显著降低。另外在利用电话语音的说话人识别系统中，3KHZ 频带以外的说话人信息的丢失，包括电话机在内的传输线路特性的变化，来自不同干线的话音质量存在差异以及通话环境的噪音等，都严重影响说话人识别系统性能。在语音识别领域里，利用 HMM 合成法在特定的信噪比（SNR）条件下，用语音模型和噪声模型合成出耐噪声性能好的语音模型，取得了很好的效果。文献[23]把这一思想应用到说话人识别系统中，利用 HMM 合成法进行与文本无关的说话人识别实验。在他们的实验中,学习阶段利用无噪声环境下录制的说话人语音数据，建立 1 状态混和高斯分布语音 HMM，同时，利用在实环境下录制的噪声信号数据，生成 1 状态混和高斯分布噪声 HMM。在识别阶段，利用 HMM 合成法，根据输入信号的 SNR，把说话人语音模型和噪声模型的分布参数进行加权组合，建立噪声说话人模型，然后利用该模型进行说话人识别，同样得到了很好的识别效果。在利用 HMM 合成法的实环境下说话人识别中，必须预先知道 SNR，然而对于非平稳噪声的情况，则很难正确地测量 SNR。为此文献[24]提出了对于未知 SNR 情况下的改进识别方法。在他们的方法中，使用复数个 SNR 建立多个合成噪声说话人模型，在识别阶段，输入语音对复数个合成模型计算概率，取其最大值作为说话人模型的生成概率，实验结果表明了这种方法的有效性。

4 尚需进一步探索的研究课题

说话人识别具有广泛的应用前景，在几十年的研究和开发过程中尽管取得了很大的成果，但还面临许多重大问题有待解决。过去人们对说话人识别的研究发展前景曾经一度相当乐观，但现在人们对此有了更加清醒的认识，比如目前对于人是怎样通过语音来识别他人的这一点尚无基本的了解；还不清楚究竟是何种语音特征（或其变换）能够唯一地携带说话人识别所需的特征。目前说话人识别所采用的预处理方法与语音识别一样，要根据所建立的模型来提取相应的语音参数。由于缺少对上述问题的基本了解，因此在这样做的过程中，很可能不自觉地丢失了许多本质的东西。这些基本问题的解决还需借助于认知科学等基础研究领域的突破以及跨学科的协作，但都不是短期内能够实现的。现在，对于利用 HMM 的说话人识别系统，在用高品质的话筒，用从安静环境下的语音信号中提取的倒谱参数，利用事后设定的阈值（一般设为错误拒绝率和错误接受率相等的值）进行判定时，对于几十名话者的识别率可以达到 99%以上。然而，对于通过实际网络（例如市话网）传输的电话语音、存在噪音的实环境下的语音进行判定的说话人识别系统性能还有待研究提高。对于说话人确认系统，虽然理论上识别率和登录的话者数无关，但实际上对于利用 2 值（本人或他人 2 阈值）判定的说话人确认系统，怎样提高有许多登录话者的系统的确认率仍然是一个课题。因此，在说话人识别技术中，有许多尚需进一步探索的研究课题。例如，随着时间的变化，说话人的声音相对于模型来说要发生变化，所以要对各说话人的标准模板或模型等定期进行更新的技术；判定阈值的最佳设定方法等；特别地在存在各种噪声的实环境下，以及电话语音的说话人识别技术，还没有得到充分的研究。以下作为文章的结尾，指出一些尚需进一步探索的研究课题。

4.1 基础性的课题

a）关于语音中语义内容和说话人个人性的分离，系统地全面地进行研究的人还很少。现在语音内容和其声学特性的关系已经较明确，但是有关说话人个人特性和其语音声学特性的关系还没有完全搞清楚。个人特性的详细研究，不仅在说话人识别方面，而且在语音识别方面也是非常重要的。b）说话人特征的变化和样本选择问题。对于由时间、特别是病变引起的说话人特征的变化研究的还很少。感冒引起鼻塞时，各种音尤其是鼻音的频率特性会有很大的变化；喉头有炎症时会发生基音周期的变化。因此，由于感冒而不能进公司大门，这是一个大问题。另外对于样本选择的系统研究还很少。根据听音实验，不同的音素所包含的个人信息是不同的，所以样本的合理选择对识别率也有很大影响。c）利用听觉和视觉的说话人识别研究是用计算机进行说话人识别的基础，例如什么样的特征对说话人识别有效，语音的持续时间和内容与识别率的关系等。利用视觉的说话人识别主要是通过观察声纹（Voice Print）差别来判定说话人，这在公安侦察上很有用[25]。

4.2 实用性的问题

a）说话人识别系统设计的合理化及优化问题。包括在一定的应用场合下对系统的功能、指标合理定义并对使用者实行明智的控制，选择有效而可靠的识别方法，既能正确识别说话人，又能拒绝模仿者。b）说话人识别系统的性能评价问题。需要建立与试听人试验对比的方法和指标；由于目前对于人识别人的性能尚无认识一致的评价方法，所以这一问题的解决还需长期的努力。c）可靠性和经济性。和语音识别系统相比，说话人识别系统的使用者要多几个数量级，例如有信用卡的人可以是几百万或上千万，当然不一定所有的都用一个系统来处理，但是在把说话人识别系统用于社会以前，必须先设想万位以上的说话人进行可靠性实验。在经济性方面，和上面的理由相同，每一说话人的标准模型必须使用尽量少的信息，所以样本和特征量的精选也是急待解决的问题。

参考文献：

[1] 松井知子，古井贞熙. テキスト指定型话者识别のための话者适应化法[A].日本音响学会全国大会论文集[C], 千叶,1994, I (3-7):158-160,

[2] Matsui T and Furui S. Comparison of text-dependent speaker recognition methods using VQ-distortion and discrete/continuous HMMs[A]. IEEE Proc ICASSP' 93[C], 1993, II:391-394，

[3] Markov K and Nakagawa S.Text-independent speaker recognition using non-linesr frame likelihood transformation[J]. *Speech Communication* , 1998, 24:193-209.

[4] 古井贞熙. 音响　音声[M]. 东京：近代科学社，1992

[5] 松井知子，古井贞熙. VQ、离散/连续 HMM によるテキスト独立话者认识法の比较[A]. 电子情报通信学会论文志[C]， 1992, J77-A (4):601-607.

[6] Gauvain J L and Lamel L F. Experiments with speaker verification over the telephone[A]. Proc. Eurospeech[C], 1995, I:651-654

[7] Setlur A and Jacobs T. Results of a speaker verification service trial using HMM models[A]. Proc. Eurospeech[C], 1995, I:53-56

[8] 松井知子，古井贞熙. 音源 声道特征 を用いたテキスト独立话者认识[A]. 电子情报通信学会论文志[C]，1992, J75-A (4):703-709

[9] 林平澜，王仁华. 动态 HMM 在说话人识别中的应用[J]. 信号处理，1993, 9 (4):250-256

[10] Soong F K and Rosenberg A E. On the use of instantaneous and transitional spectral information in speaker recognition[A]. IEEE Proc ICASSP' 86[C], 1986, I:877-880.

[11] 板仓文忠.　音声の高域に含まれる个人性信息を用いた话者认识[A]. 日本音响学会论文志[C]，1995, 51(11):861-868

[12] 赵力，邹采荣，吴镇扬. 汉语连续语音识别中语音处理和语言处理统合方法的研究[J]. 声学学报，2000, 25(6):618-672

[13] Luck J E. Automatic speaker verification using cepstral measurements[J]. *J.Acoust. Soc.Am.*, 1979, 46(4)

[14] 古井贞熙. 音声の个人性パラメ-タの时期的变动と话者认识[A]. 电子情报通信学会论文志[C]，1974, 57-A(12):880-887

[15] Rosenberg A and Lee C. Connected word talker verification using whole word Hidden Markov Models[A]. IEEE Proc ICASSP' 91[C], I:381-384

[16] 松井知子，古井贞熙. テキスト指定型话者认识[A]. 电子情报通信学会论文志[C]，1996, J79-D-II(5):647-656

[17] 松井知子，西谷 隆，古井贞熙. 话者照合におけるモデルとしきい值の更新法[A]. 电子情报通信学会论文志[C]， Vol.J81-D-1998, II(2):268-276

[18] Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification[J]. *J.Acoust.Soc.Am.*,1998, 55(6):485-490

[19] 古井贞. Cepstram Analysis Technique for Automatic Speaker Verification[J]. *IEEE Trans.*.ASSP, 1981, 29(2):1078-1082

[20] Rosenburg. The use of cohort normalized scores for speaker verification[A]. Proc ICSLP[C], 1992, II:821-824

[21] UCHIBE T. A study on the long-term variation in speaker recognition[A]. 日本音响学会全国会议论文集[C],1996, I(3-13):11-112

[22] Tagashira S and Ariki Y Speaker recognition and speaker normalization by projection to speaker subspace[R]. IEICE Technical Report，1995，SP95-28:25-32.

[23].Rose R C, Hofstetter E M andReynolds D A. Integrated models of signal and background with application to speaker identification in noise[J]. *IEEE Transactions on SA*,1994, 2:245-257

[24] Matsui T andFurui S. Speaker recognition using HMM composition in noise environments[A]. Proc.Eurospeech[C], 1995, I:621-624

[25] 中田和男.音声[M]. 东京:コロナ社,1995

作者简介：赵力，1982 **年毕业于南京航空航天大学自动化系，**1988 **年取得东南大学工学硕士学位，**1998 **年获日本京都理工大学（**Kyoto Institute of Technology**）博士学位，**2001 **年从东南大学信息与通信学科博士后站出站。副教授。主要研究汉语语音识别、自然语言处理、情感信息处理等。中国电子学会、日本音响学会、日本通信与电子学会会员；邹采荣，博士生导师，东南大学副校长，主要研究领域为数字信号处理；吴镇扬，博士生导师，主要研究领域为数字信号处理。**

HMM-based speaker recognition

ZHAO Li,　ZOU Cai-rong,　WU Zhen-yang

（Department of radio engineering，Southeast university, Nanjing 210096,China）

Abstract： Advance of automatically speaker recognition based on HMM is reviewed. Some problems in automatically speaker recognition are discussed. Moreover, topics for further research are also given.

Key words: Automatic Speaker Recognition; Hidden Markov Model; Speech