摘　　要

基于生物特征的身份识别技术是当前国际上的重点研究内容，自动说话人识别通过语音识别说话人的身份，在系统安全认证、司法鉴定、金融服务以及电子侦听等领域有着广泛的应用价值。本文在对现有说话人识别技术分析的基础上，运用互信息理论进行说话人识别的研究，提出了可实际应用的语音信号互信息计算方法，并针对基于文本和文本无关的说话人识别分别提出了相应的说话人语音模型和互信息匹配算法，实验证明了本文提出的语音信号互信息计算方法的有效性。

本文的主要研究内容如下：

一、对自动说话人识别原理以及相关的语音产生机理和语音信号处理方法作了全面的描述与分析。特别在特征参数选择与提取、说话人语音模型建立、模式匹配以及语音的声学特性方面进行了详细的分析。

二、从信息量的角度考察分析语音信号之间的特征相关性，提出随机干扰信号的概念来解释和描述语音信号之间的失真，并从随机信号的特征以及随机信号分析理论推导出这一信号的统计分布特性，最终，语音信号之间互信息的计算归结到随机干扰信号的熵并得到解决。

三、研究了语音信号互信息计算的具体算法，提出了基于模式的线性映射匹配算法LPM和非线性搜索匹配算法NLM。

四、对互信息测度的聚类特性进行分析，通过类内凝聚度、类间耦合度和类间重叠三大指标对互信息测度的分类特性进行了详细分析，并与其它常用测度Euclidean、Itakura-Saito和Mahalanobis进行比较，结果显示出互信息测度的模式分类有效性和优越性。

五、针对不同识别要求研究适合互信息测度应用的说话人模型，提出应用于基于文本的说话人识别的多模板模型MTM和应用于文本无关说话人识别的全特征矢量集模型CFC，实验证明这些模型能够充分表达说话人的语音特征。

六、对于文本无关的说话人识别，综合考虑距离空间和信息空间的特性，提出多级最小最大

搜索匹配算法MMS计算全特征矢量集模型CFC和语音信号的互信息，实验证明该算法有效。

本文提出的基于互信息理论的说话人识别方法综合运用了语音信号的时变分布与统计分布特征，在基于文本和文本无关的说话人识别实验中显示出比基于 GMM 模型的识别方法优越的识别性能。本文的研究工作有助于自动说话人识别技术的完善、发展和提高，有利于基于生物特征的身份识别技术的实际应用。

关键词：说话人识别，互信息，匹配，语义特征，个性特征

Abstract

Speaker recognition as one of biometric identification research aims to identify living persons from their voice. It is useful in person authentication, forensics and speaker tracking, etc. Many scientists and engineers have contributed their wisdom and enthusiasm in this challenge research, but still there are many problems such as speaker model optimization and adaptation, feature selection and detection, pattern measure and matching left for further study. This thesis proposes a new approach based on mutual information theory to investigate the speaker recognition problem. The most attention focus on mutual information estimation of speech signals, speaker model and pattern matching scheme, performance evaluation and analysis with comparison to Gaussian based method. The main research work and achievements are as following.

The previous work and results in speaker recognition research and its fundamental principle are introduced with discussion and analysis. Based on mutual information theory and analysis of statistical distribution and stochastic property of speech signal, the mutual estimation method was derived by defining a random interference signal to describe the distortion between speech signals. Two practical calculation algorithms were proposed as Linear Projection Matching (PLM) algorithm and Non-Linear search Matching (NLM) algorithm. Both time-varying and statistical distribution features can be well processed by these algorithms, and it make proposed method more meticulous and robust than traditional VQ and GMM methods which did not take process of neither one of the two features.

Speaker models named as multi-template model (MTM) and complete feature corpus model (CFC) were proposed respectively for text-dependent speaker recognition and text-independent speaker recognition. MTM represents central templates of a speaker's text-dependent voice in the pattern space, CFC is designed as an adequate description of speaker's phonetic and pronunciation properties and practically trained by a clustering algorithm in feature vector space with sufficient samples.

Text-independent speaker recognition scheme is an integration of CFC and a matching algorithm as Multi-step Mini-max Search algorithm (MMS). MMS algorithm makes the input speech and CFC speaker model sequentially match in distance space and information space with minimum distance and maximum mutual information

criteria respectively.

Experiments on clustering and classification property analysis show that the proposed mutual information measure has larger intra-class compactness and smaller inter-class intersection than traditional Euclidean, Mahalanobis and Itakura-Saito measures. This result is also demonstrated by the speech digits recognition experiment.

Speaker identification experiments based on proposed mutual information method are examined and analyzed. The results both of text-dependent and text-independent speaker identification experiments were compared with the method based on Gaussian Mixture Model. As can see from Chapter 6 and 7, the proposed mutual information method is effective and has better performance than GMM. From our experiments, mel-frequency cepstrum coefficients are more effective than linear prediction cepstrum coefficients.

In summary, investigating speaker recognition from viewpoint of mutual information theory is successful. The proposed speaker models with corresponding matching algorithms provide a new way to make the speaker recognition system more consummate.

Keywords: Speaker recognition, Mutual information, Matching, Linguistic property, Individual property

上 海 大 学

本论文经答辩委员会全体委员审查，确认符合上海大学博士学位论文质量要求

答辩委员会签名：

主任：

委员：

导师：

答辩日期：2004年9月13日

原 创 性 声 明

本人声明：所呈交的论文是本人在导师指导下进行的研究工作。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已发表或撰写过的研究成果。参与同一工作的其他同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签名：[signature] 日期：2004.9.13

本论文使用授权说明

本人完全了解上海大学有关保留、使用学位论文的规定，即：学校有权保留论文及送交论文复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容。

（保密的论文在解密后应遵守此规定）

签名：[signature] 导师签名：[signature] 日期：2004.9.13

第一章 绪 论

本章提要：

- 说话人识别的概念
- 说话人识别的应用
- 说话人识别的特点与难点
- 本文研究工作的意义、主要内容与指导思想

语音是人类最自然的通信方式，说话人识别研究的目的是使机器能够通过语音来判断说话人的身份。在我们的日常生活中，人们经常通过电话等各种方式交流信息，当一方在线路的一端对着话筒说话时，另一方能够很快判断出对方是否是熟悉的人，如果熟悉的话还能够很快知道是哪一位。这是一个日常生活中典型的说话人识别事件，通过话筒传来的语音进行说话人身份的识别。

在当今世界进入信息化时代的过程中，关于身份鉴定与识别的需求越来越多，一般可以通过以下三种方式进行：（1）钥匙或信用卡；（2）PIN码或密码；（3）签字、指纹、声音或人脸。其中，前两种方法是已经使用了几个世纪的传统方法，这些方法的缺点是容易丢失和遗忘，甚至被错误使用。第（3）种方法是一种基于生物特征的身份鉴定识别方法[1,2,3]，签字、指纹、声音或人脸这些生物特征都反映了个体的生理、心理特性以及长期的文化与生活习性，是自然唯一的、具有随身携带和不会丢失遗忘的特点。

在过去的10年里，随着计算机运算速度的提高以及超大规模集成电路体积越来越小，研究开发基于生物特征的身份识别系统越来越受到重视。本文探讨通过语音信号特征分析进行说话人识别的方法，研究如何运用互信息理论分析语音特征，建立说话人语音模型以及匹配识别的具体途径。

1.1 说话人识别基本概念

说话人识别根据具体的任务可以分为说话人辨认和说话人确认两大类[4,5]。在说话人辨认中，一个未知说话人的语音特征与N个已知说话人的语音特征进行比较，进行1—N匹配，获得

最佳匹配的说话人作为识别结果。在说话人确认中，需要将未知说话人的语音特征与其所声称的说话人的语音特征进行比较，实行1：1匹配，判断两者是否为同一个人，如果语音特征之间的距离小于预设阈值或似然度大于预设阈值，则接受，反之则拒绝。

一般认为说话人辨认是一个比说话人确认更困难的任务。这一推论的直观性在于，随着登记的说话人人数增加，错误判决的概率会上升[1,6,7]。而对于说话人确认来说，理论上并不会因为人数的增加导致性能下降，因为比较匹配的只是两个人。

1.1.1 面向闭集和开集的说话人辨认

说话人辨认可以进一步分为面向开集（open-set）的说话人辨认和面向闭集（closed-set）的说话人辨认两种情况。如果所需识别的说话人都在预先登记的说话人集合中，则称为面向闭集的说话人辨认，但如果所需辨认的说话人有可能不属于预先登记的说话人集合，则称为面向开集的说话人辨认。一般来说，面向开集的说话人辨认问题难度更大些。对于面向闭集的说话人辨认而言，通过输入语音与各说话人语音模型之间的一一匹配，依据最佳匹配准则来决策，辨认结果是具有最佳匹配值的语音模型所对应的说话人，而不管这个所谓的最佳匹配值具体多少。然而，在面向开集的说话人辨认中，必须预先设置一个阈值，如果最佳匹配值超过这一阈值，则进行决策辨认，反之，则认为说话人为未登记的未知说话人而加以拒绝。因此，说话人确认实际上是面向开集的说话人辨认的一个特例，只是预先登记的说话人集合中只有一个说话人。

1.1.2 基于文本和文本无关的说话人识别

说话人识别根据对输入语音的要求可以分为基于文本（text-dependent）的说话人识别和文本无关（text-independent）的说话人识别两大类。对于基于文本的说话人识别来说，识别时输入语音所对应的文本预先是知道的。而对于文本无关的说话人识别而言，输入语音文本可以是任意的。显然，后一种情况的难度要大些，说话人模型必须能够反映说话人的声道和发音特征，而不仅仅是发某个特定语音的特征。

一般，基于文本的说话人识别性能较高，因为在语音匹配时不仅可以利用语音特征，还可以利用语义特征。因此，语音识别机制可以被用来判别说话人所说的语音与所提示的是否一致，实现语音确认，并可以与说话人确认综合运用[8]。对于说话人确认系统来说，输入语音可以是固定的，也可以是变化的，系统可以在不同的时候采用不同的文本，并提示用户按

新的文本输入语音。例如，系统可以随机地从一个设计好的文本数据库中选择一个文本作提示。文本数据库可以选择由一些单词或语句段构成，也可以采用更灵活的方式，即在识别时根据一些基本单元（如单字）随机组合一个单词或语句段。这样的方式称作文本提示（text-prompted）说话人确认，其好处是任何人无法在事先知道系统所提示的文本，也无法通过回放事先录音的方式来仿冒真正的说话人，并且，由于系统要求用户在提示后很短的时间内输入语音，仿冒者也无法通过软件合成语音等手段进行诈骗。

1.1.3 说话人识别的其它方式

上面的内容是从自动说话人识别的观点而言的。从一个更广义的角度来看，说话人识别的方式还有基于听觉的说话人识别（auditory）和介于听觉与自动识别之间的半自动说话人识别（semi-automatic ）。

（1） 基于听觉的说话人识别

在我们的日常生活中每天都在运用听觉进行说话人识别。当我们听到熟悉的人的声音时，能够很容易地识别他们的声音。另外，即便没有事先足够的“语音训练”，我们仍然能够从声音上大体估计出说话人的一些特征，如年龄、性别等。

在司法鉴定中，如果有人在犯罪嫌疑人作案期间听到过他（她）的声音，那么，说话人识别将非常有用。但是，由于各人的听觉特性有差别，因此，不同的人进行说话人识别的能力是很不一样的[4,9]。另外，随着先后两次听音时间间隔的增加，人类的识别能力将下降[10]。

有若干种方法可以对人类和机器的说话人识别性能进行比较分析[9,11,12]。Crystal 和 Schmidt-Nielsen曾经做过一个大数据量的比较，对65个听众组进行了共50000次听觉测试，每一个听众组包含8个人，实验的结果与一个计算机识别系统的识别结果进行了比较。实验发现，不同人的说话人识别能力是有很大区别的，并且，不同的人所使用的判决阈值是不一样的，也就是说错误接受率（FA: False Acceptance）和错误拒绝率（FR: False Rejection）完全是因人而异的。关于识别能力的比较方面，Schmidt-Nielsen 和 Crystal 发现在纯净环境下，人类和机器的说话人识别能力是相差无几的，但在有背景噪声、线路扰动和多人说话等噪声环境下，人类的识别能力要好些。

Schmidt-Nielsen 和 Crystal 两人的研究结果发表在 NIST 的1998年说话人识别系统评估报告中[13]。但是，由于近年来机器说话人识别技术的迅速发展，他们的结论也许已经过时了，一些新的、更加有效的技术的出现使得很多研究人员对自动说话人识别投入了大量的热情

[14,15,16,17]。最新的说话人识别系统已经同时使用低级和高级的说话人特征信息[16,17]，运用了韵律统计特性、N-grams、发音模型、多分辨率分析等进行识别处理，因此，与以前的说话人识别系统相比，识别性能有了很大的提高。

（2） 半自动说话人识别

利用信号处理分析技术和人类的感知特性进行说话人的识别可以增加识别结果的可靠性和认同度。Rose 的研究指出 [18]，在司法鉴定中，声音应该采用多种不同的手段进行分析，进行比较的两个语音样本必须在语言学角度看是有可比性的，必须仔细挑选用于比较的音素、单词等单元。这需要具有语音学专业知识的研究人员对语音样本进行手工分割，包括使用听和看的方法，如观察信号波形和语谱图等。

图1.1显示的是几张语谱图，一般的人很难凭肉眼判别其中哪两个图属于同一个说话人，这说明了实际应用中声音比较的复杂性。在语言专家进行语音分割和语音单元选择时听觉比较是很有帮助的。很明显，在司法鉴定中，语音比较必须采用半自动方式处理，因为任何被告不愿让计算机来作判决，最终的分析必须通过细致的主观评判进行，并结合合理的统计分析。

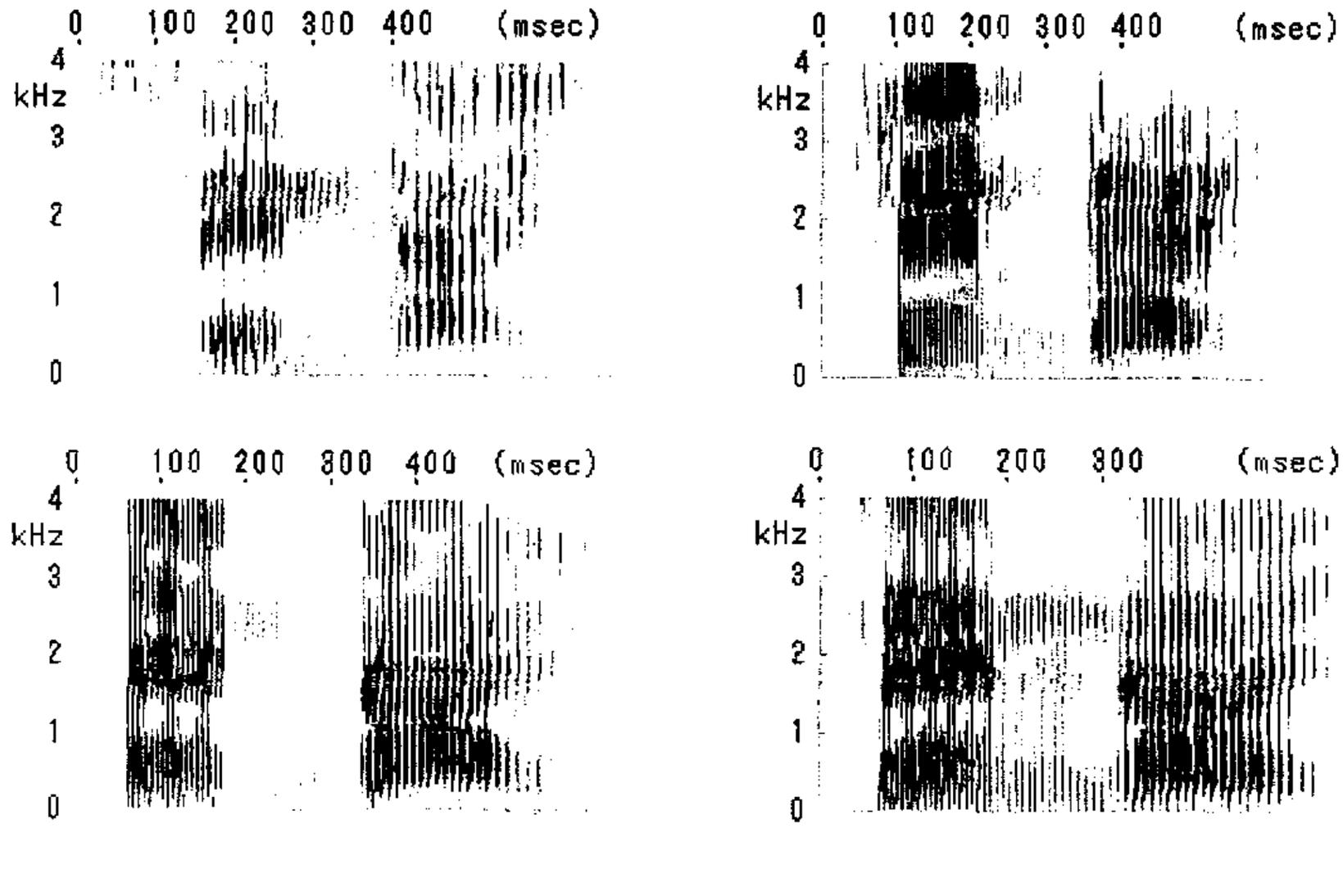


图1.1　三名男性说话人的“HIROSHIMA”发音，其中两个由同一说话人发音

1.2 说话人识别技术的应用

说话人识别技术的应用主要包括以下几个方面：（1）个人身份认证鉴定；（2）电子侦

听与司法鉴定；（3）多说话人环境下的话者检测；（4）语音识别系统的话者自适应；（5）个性化的人机界面。个人身份认证鉴定是所有生物特征识别技术的最主要应用。另外，说话人识别技术可以与人脸识别技术等结合应用于银行信用交易，计算机和数据库系统登录，安全部门的身份确认检查等。司法鉴定也是说话人识别的一个重要应用。如果犯人的声音在作案期间被录音，那么，采用说话人识别技术就可以通过犯罪嫌疑人的声音与所录声音的对比分析来判断两者是否一致[18]。

语音识别的研究目的是将声音转换成文字，自50 年代开始以来，世界各国的研究人员都对此进行了大量的研究[19,20,21,22,23]，但到目前为止，还没有一个语音识别系统能够实现无限词汇的非特定人识别。语音识别和说话人识别是两个既有联系又有区别的问题，两者通过语音信号分析分别对语义和说话人身份进行识别。在语音识别中，说话人的变化是影响识别性能的一个主要问题，而对说话人识别来说却主要是语义的影响。基于说话人识别技术的说话人自适应可以在语音识别系统中起作用，减少说话人变化波动的影响[24]。例如，语音识别系统可以包含一个“说话人选通”模块，其识别当前说话人（图1.2），这样，语音识别系统可以调整系统参数适应当前说话人或选择说话人相应的特定语音模型进行识别。

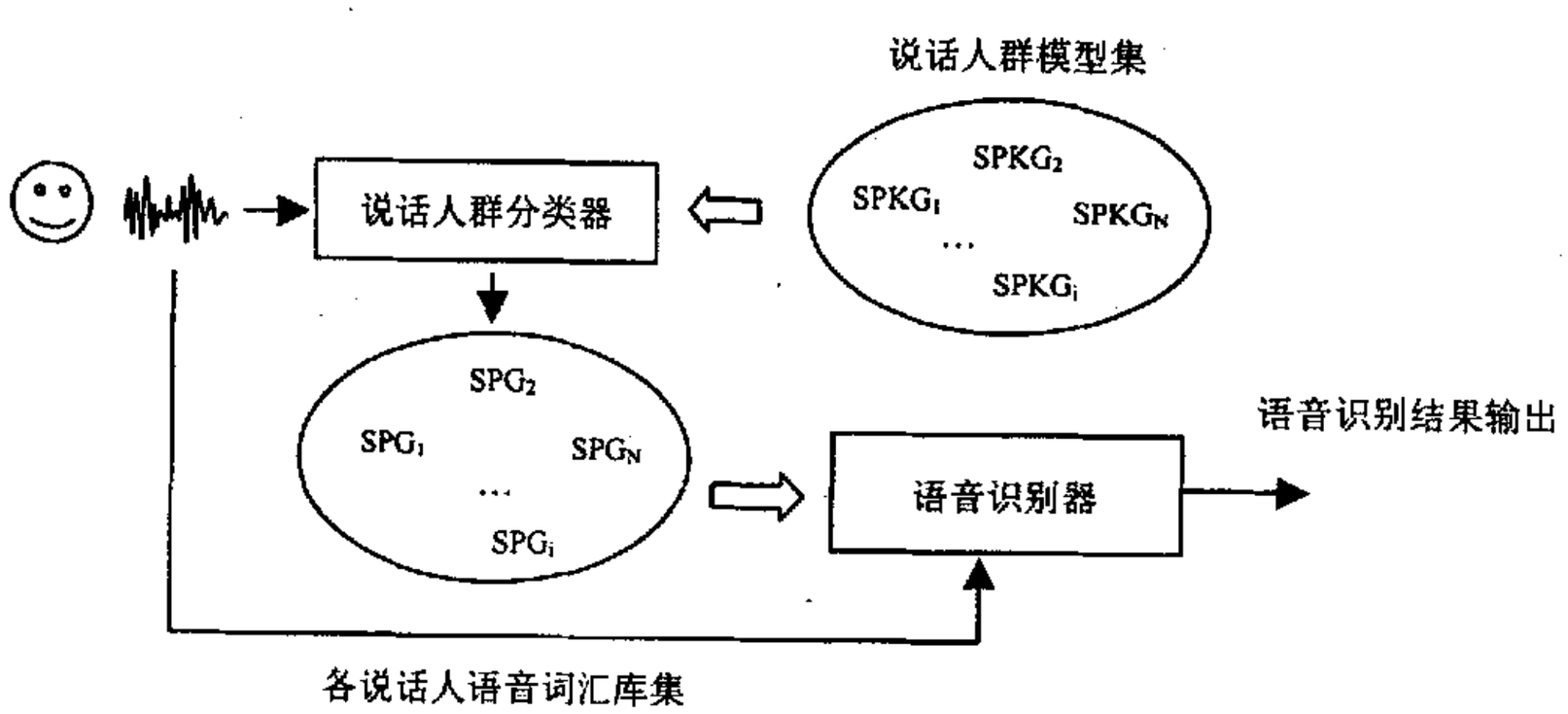


图1.2 说话人识别应用在话者自适应处理

在一个多说话人语音环境中，系统语音的输入涉及到多个说话人，例如，讨论会、法庭辩论、电话会议以及电视广播等场合，在这些情况下，说话人识别技术将十分有用。根据具体的特征，可以将多说话人环境下的说话人识别分成三种情况[25]：（1）说话人检测；（2）说话人跟踪；（3）说话人分割。说话人检测的目的是判断某个说话人是否正在说话；说话人跟踪是指判断说话人的语音持续轨迹，包括停顿的位置；说话人分割则是检测每一个说话人

的语音片段位置。在大部分情况下，都无法预先知道说话人的先验知识和数量[26]。有关说话人交替分割的技术已经被应用于新闻播音的语音分割中[27,28]。

另外，随着语音信号处理技术的发展，语音人机界面正变得越来越普遍，例如语音电子邮件。通过说话人识别技术，系统可以自动地进行调节以适应用户的需要和喜好。

1.3 说话人识别技术的特点

说话人识别的一个最大优点是自然。语音是人类最自然的通信方式，因此，说话人识别技术在具体应用中是最容易被接受的。另一个主要优点是其低成本特性，说话人识别技术的应用不需要特殊的设备。指纹识别和虹膜识别等其它生物特征识别方法都需要特殊的扫描输入设备，而说话人识别仅仅需要一个话筒。另外，说话人识别中，信号处理和模式匹配算法只需要较低的存储量，计算也并不十分复杂，因此适合移动通信终端设备的应用。

当然，说话人识别相对虹膜识别等其它生物特征识别技术而言有一定的非稳定性，因为人类的“声纹”并不象虹膜那样具有完全的唯一性。语音和虹膜虽然都是属于生物特征，但它们之间存在一定的区别，虹膜是一种生理性生物特征，它直接通过人体特征的测量而获取，而语音是一种行为性生物特征，它是通过人体发音器官的运动而生成，是发音器官生理特性的一种变换形式，即便是同一个人连续地发同一个音，前后两次发音过程中发音器官的运动或生成的语音信号不可能绝对一致。

已经证明，语音和其它生物特征相结合的多模态身份识别技术将会大大提高识别性能[29]。在最近一次AVBPA会议上，语音和人脸、指纹一起成为生物特征识别的三大热点研究内容[30]。在会议上还介绍了九种不同的多模态生物特征识别系统，说话人识别系统也是其中之一。

有关说话人识别议论最多的一个问题是：语音很容易被假冒。但实际上这个说法是很主观的。一般，假冒者经常会夸大特定的生物特征，并经常使用可视信息来强化仿冒的真实度。当一个假冒的说话人说话时，人们一般会感觉到异样，并会更加注意辨别声调等发音细节[31]。说话人的个性特征通常包含多种参数以便于相互区分，人类也许仅仅只利用了其中一部分。因此，所谓的不可靠问题也只是一种基于人们自身主观印象的论断而已。

1.4 说话人识别技术的难点

影响说话人识别性能的因素很多，有些直接与说话人本身有关，而有些则是由技术条件

的限制引起。

1.4.1 说话人本身的变化因素

众所周知，一个人生理上和心理上的变化会引起声音的改变[32]，情绪、刺激和药物同样影响人的发音（见图1.3）。当说话人感到紧张的时候，与放松状态相比，其声音也会发生很大的不同，特别是声调和语调会发生较大的改变。同样，声音会随着年龄的增长，体重的增加以及其它生理改变而变化，实际上，这些变化是导致说话人声音发生变化的最大因素。根据作者的经验，即便是在相同的技术条件下，同一天不同时刻录制的语音也会发生不匹配的情况，一些说话人的训练效果比其他人要难些[33,34]。一般，训练数据应该考虑语音特性的平衡性，以使得其包含所有语言单元所对应的声音以及不同上下文时声音的多个样本，这样，当任意一个语音输入时都能够由系统来识别。

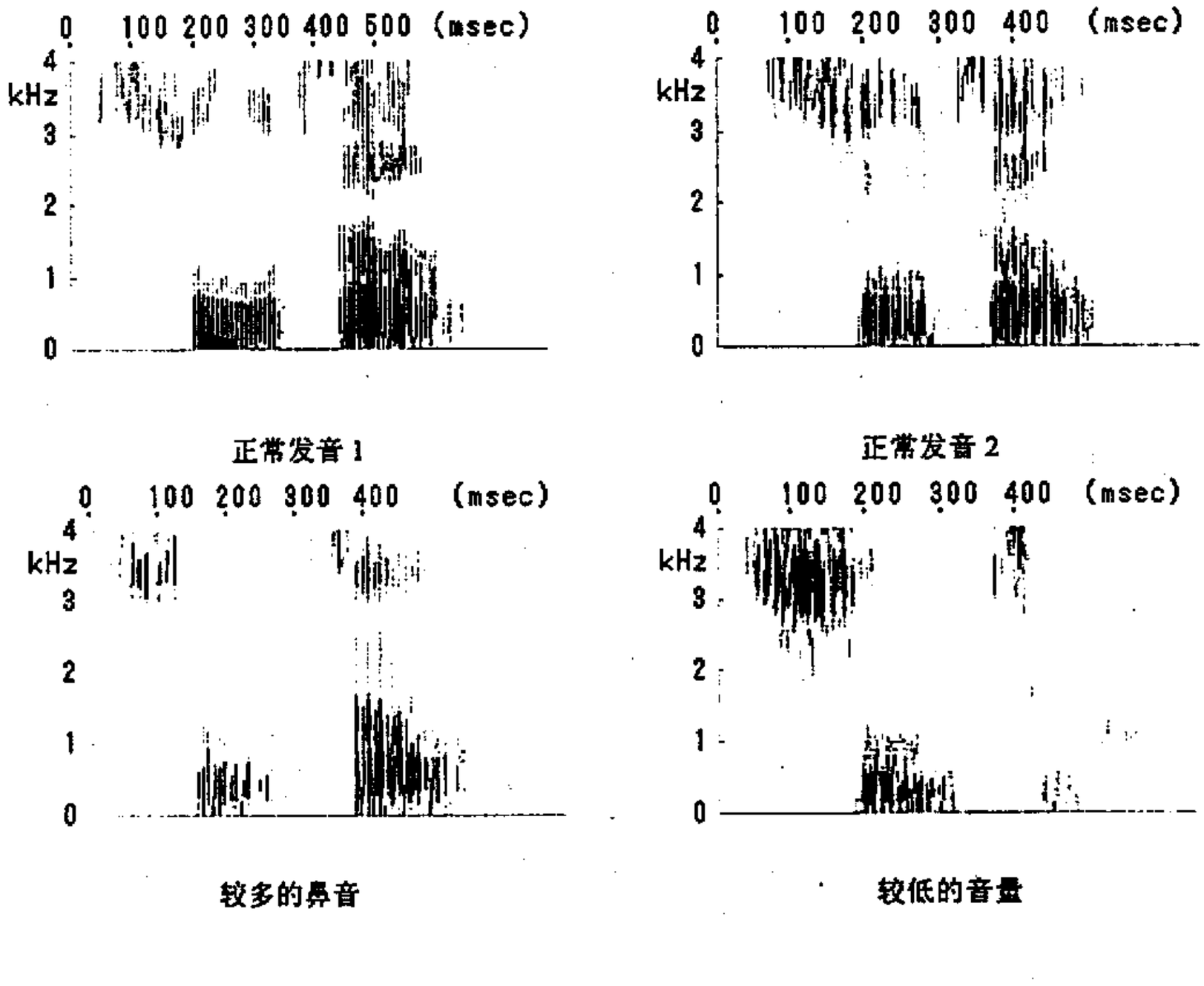


图1.3 同一说话人以四种不同方式发“苏州”的语谱图

1.4.2 声音掩饰与仿冒

声音掩饰指故意改变声音，以便其不能与同一个说话人的正常语音很好匹配。声音掩饰在司法鉴定中比较普遍。例如，当犯罪嫌疑人打匿名电话时会刻意地通过改变发音器官的自

然形态来掩饰真实的声音，或在警方调查时改变声音。有关这方面的研究已经有了一些成果，例如，Majewski 等曾经对三种特征参数的抗掩饰能力进行了分析比较[35]，得出了哪一种参数最具有鲁棒性得结论。实验分析中所采用的声音掩饰方式包括耳语和口腔含物等形式。但是，该研究仅仅对同一说话人的变化进行了分析，因此并不能将其结论应用于不同说话人之间的变化。即使一个特征参数对于声音掩饰的鲁棒性很强，对同一个说话人无论怎样掩饰变化都很小，也可能并不适合其它的说话人。

仿冒说话人是声音掩饰的一个特殊形式，仿冒者企图象另一个说话人一样发音，达到以假乱真的目的。

显然，无论是刻意的声音掩饰，还是仿冒，都会影响说话人识别系统的性能。研究表明，通过声音映射将声音变换为另一个人的声音会降低说话人识别的性能[36,37]。

1.4.3 技术因素

一些技术因素同样会引起说话人识别性能的下降，包括基于听觉特性的说话人识别和自动说话人识别，主要因素如图1.4 所示。

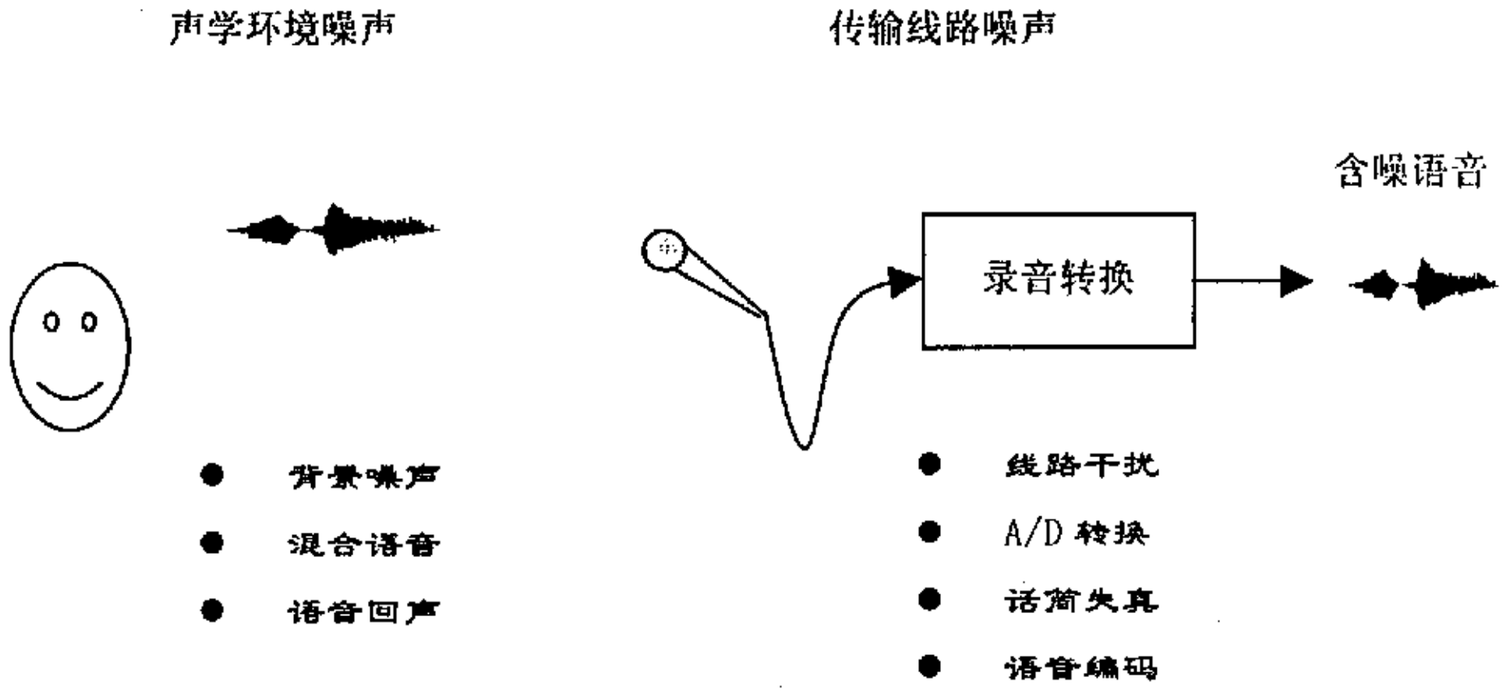


图1.4 引起说话人识别错误的技术因素

首先，语音通过话筒或电话信道录制，环境噪声（计算机、汽车、键盘、开关门、音乐、背景声音等）将被叠加到语音信号中。反射将引起一个延迟语音信号与所录制的语音信号的叠加[38]；低质话筒将在语音频谱上引起非线性失真；Quatieri等通过对高质量话筒和低质话筒下所录制的一组相同文本语音进行比较发现，低质话筒会引起一些假共振峰等谱畸变，这些假共振峰往往出现在真共振峰的和及乘积处，并且，共振峰带宽增大，频谱形状变得扁平[39]。

A/D转换器本身存在失真，录音设备容易受到移动电话电波的干扰。如果语音通过电话线传输，有损数据压缩技术（数字传输时）同样会带来噪声。语音信号编码也会引起说话人识别性能的大幅下降[40,41]。综上所述，人耳听到的语音与识别系统所得到的语音信号是不一样的，在各种传输变换过程中，信号质量不断降低。不同的环境是影响识别性能的另一个重要的因素[42,43,44]，由于系统训练和识别处于不同时期，不仅说话人本身特征发生了变化（如基音、声调），技术条件上也发生了变化[45,46,47]。这些变化的因素如下：（1）环境声学特性不一致；（2）背景噪声能量和形式不一致；（3）话筒不一致；（4）录音质量不一致。环境不一致的现象很容易发生，例如，系统本身是在安静环境下训练的，但使用却在有噪环境中的情况。

关于噪声环境下的识别算法研究和鲁棒性特征参数提取的算法研究一般在以下假设下进行[38,48]：（1）噪声是短时平稳的；（2）噪声的均值为零；（3）噪声与语音信号是独立的。

1.5 本文研究工作的意义、基本思路与主要内容

1.5.1 研究意义

作为一种基于生物特征的身份识别方法，说话人识别通过语音来识别说话人的身份，在电子银行、远程网络系统和数据库系统的用户身份确认、电子侦听中说话人身份的自动检测与判别及其它各类安全系统的身份认证中有很大的应用价值，并具有其它生物特征身份识别方法所不具备的特点，即数据采集设备的非接触性和简易性，是当前语音信号处理研究领域的热点之一。本文的研究工作有助于自动说话人识别技术的完善、发展和提高，有利于基于生物特征的身份识别技术的实际应用。

1.5.2 国内外研究现状

说话人识别的系统性研究开始于 20 世纪 80 年代，但对于语音信号特征参数和说话人模型的分析并不充分。进入 90 年代后，随着基于数据挖掘技术的研究方法不断发展，美国的ＡＴ＆Ｔ、ＭＩＴ－ＬＬ以及日本的ＮＴＴ、ＡＴＲ、名古屋大学等研究机构和大学都相继开展了研究，并取得了一定的成绩[4,46,49,50,51]。我国清华大学、中国科技大学、解放军信息工程大学等高校和研究机构也对说话人识别进行了一定的研究[52,53,54]。但由于语音信号的复杂性，说话人识别技术仍未成熟，主要的原因有：（1）说话人特征难以分离和提取；（2）

缺乏可靠的、鲁棒性好的说话人模型；（3）缺乏精确高效的似然度或失真度计算方法。

语音是一种复杂信号，包含了语义、语言、声学和韵率等多种信息，而说话人的特征主要反映在声学和语音韵率这两个方面，由说话人的声道结构、声源激励特性和发音习惯决定。语音识别需要提取信号中所包含的语义和语言特征，而说话人识别需要提取有关说话人的特征信息。目前所采用的特征参数主要有 LPC、LPCC 以及 MFCC，基音等，这些参数同时表达了语义和说话人的特征信息，无法实现两者的分离描述[55,56]。作为说话人模型，目前采用的几种类型可归纳为：（1）模板结构模型（template model）；（2）码本模型（codebook model）；（3）统计模型（statistical model）；（4）人工神经网络模型（artificial neural network model）。前两类模型训练较简单，对数据量要求不高，但没有考虑语音的统计特征，而后两类模型对训练数据量有较高要求，而且训练时间较长，对新的应用环境适应能力较差，但考虑了语音的统计特征。与各类模型相对应的模式匹配算法有(1)动态规划方法 DTW（Dynamic Time Warping），适用于模板结构模型[57]；（2）矢量量化匹配方法 VQ（Vector Quantization），适用于码本模型[58]；（3）隐马尔可夫模型 Viterbi 算法（Hidden Markov Model），适用于统计模型[59]；（4）径向基函数网络 RBFN 方法（Radial Basis Function Network），适用于神经网络模型[60]。ＤＴＷ没有考虑语音信号统计特征的利用，仅适合基于文本的说话人识别；VQ 同样没有考虑语音信号统计特征的利用，但可以实现语义的归一化，因此，可适用于基于文本和文本无关两种情况；ＨＭＭ（GMM）通过 Viterbi 算法计算似然度得分，完全依赖统计特征，说话人发音的时变特征没有考虑，计算时间较长，适合文本无关的说话人识别；RBFN 通过对两层前向径向基函数网络的输出计算似然度得分，利用了语音的统计特征，匹配时间较长。

1.5.3 研究思路

互信息可以揭示两个随机信号之间相互携带对方的信息量，也可以衡量一个随机信号经过变换或传输之后信息的损失程度。互信息具有非负性、对称性和有界性。语音信号是一种短时平稳随机信号，其中包含了语义和说话人特征信息，在传输过程中这些信息会因为生理、心理和环境等因素的干扰而产生一定的损失，引起信号的失真，而这些信息的损失和失真程度完全可以运用互信息理论来进行分析与测量。在国际上，信息熵和互信息等信息理论在语音识别中的应用进行了一定的研究，主要用于系统、参数性能的评价和模型训练。例如，Lee 说明了几种常用距离测度与信息熵表示的关系[61]，Okawa 等提出了利用互信息准则提高音素模式训练精度以及多频带分配与组合的方法[62,63]，Bahl、Nakagawa 等说明了互信息测度在

HMM 模型参数训练中的应用[64,65,66]，但如何计算语音信号之间的互信息，如何建立可实际应用于语音识别或说话人识别的互信息匹配算法，如何建立相应的说话人模型以及基于互信息理论的说话人识别系统还未有研究结果。

本文基于互信息理论，从信息量这一更广泛的角度考虑说话人识别问题，通过综合分析语音的时变分布特征和统计分布特征，研究直接计算语音信号之间互信息的方法，提出能够全面反映发音特征的说话人模型以及相适应的互信息匹配算法。

1.5.4 主要研究工作与创新

（1） 对自动说话人识别原理以及相关的语音产生机理和语音信号处理方法作全面的描述与分析。特别在特征参数提取、说话人模型建立、模式匹配以及语音的声学特性方面进行详细的分析。

（2） 从信息量的角度考察分析语音信号之间的特征相关性，提出随机干扰信号的概念来解释和描述语音信号之间的失真，并从随机信号的特征以及随机信号分析理论推导出这一信号的统计分布特性，最终，语音信号之间互信息的计算归结到随机干扰信号的熵的计算并得到解决。

（3） 对互信息测度的聚类特性进行分析，通过类内凝聚度、类间耦合度和类间重叠三大指标对互信息测度的分类特性进行详细分析，并与其它常用测度Euclidean、Itakura-Saito和Mahalanobis进行比较，观察其模式分类的有效性和优越性。

（4） 研究语音信号之间互信息计算的具体算法，提出基于模式的线性映射匹配算法LPM和非线性搜索匹配算法NLM。LPM算法通过线性映射将语音信号特征矢量序列规整到相同的时域进行互信息的计算，而非线性搜索匹配算法则通过动态规划方法将两个语音信号特征序列以非线性的方式进行匹配，计算互信息。

（5） 针对不同识别要求研究适合互信息测度应用的说话人模型，提出应用于基于文本的说话人识别的多模板模型MTM和应用于文本无关说话人识别的全特征矢量集模型CFC，并通过实验对模型的特性进行分析。

（6） 对于文本无关的说话人识别，综合考虑距离空间和互信息空间的特性，提出多级最小最大搜索匹配算法MMS计算全特征矢量集模型CFC和语音信号的互信息。

本文研究工作得到江苏省教育厅自然科学基金资助（01KJD510001 基于互信息理论的说话人识别研究），也得到了国家自然科学基金（60172016）的部分资助。

第二章　自动说话人识别原理与分析

本章提要：

- 特征参数选择与提取：LPC，LPCC，MFCC
- 说话人模型与训练：码本模型 CBM，高斯混合模型 GMM
- 模式匹配算法：矢量量化 VQ，最大似然算法
- 判决准则：说话人辨认与确认的不同判决方法，归一化的重要性

典型的自动说话人识别系统结构如图2.1所示。不管是辨认系统还是确认系统，其工作状态都是两种，即训练状态和识别状态。在训练状态下，已知说话人的语音信号经过预处理后提取特征，并登记到说话人数据库。而在识别状态下，一个未知身份的说话人输入语音到系统，并有系统判别说话人的身份。

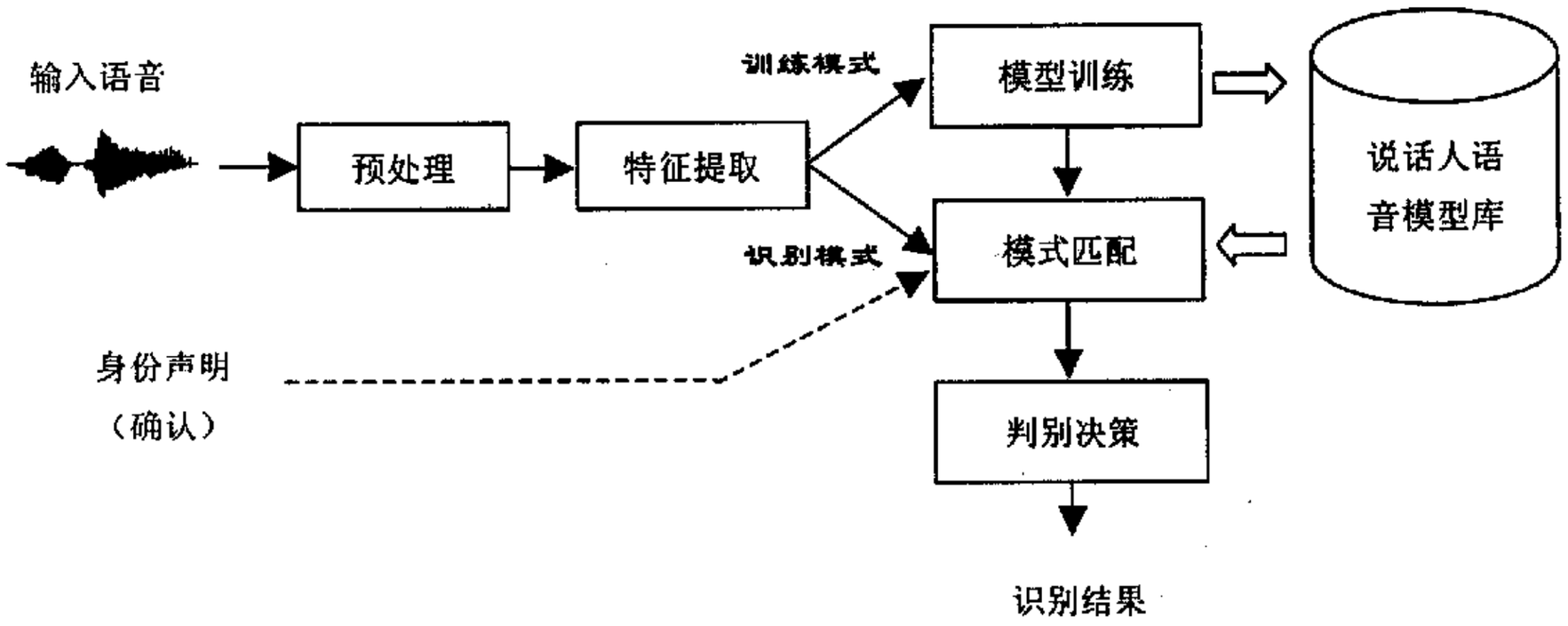


图2.1 说话人识别系统结构框图

系统的训练和识别状态都包括语音信号特征参数提取等处理，通常叫做系统前端预处理。特征提取将原始语音信号变换为特征矢量序列。特征矢量与原始语音信号相比，在表达语音特征上更加稳定、鲁棒性强、描述更紧凑。在说话人识别中，特征提取可以被看作是一种数据压缩处理过程，原始数据被压缩成一些反映信号基本特征的数据。

训练的目的是根据特征提取部分得到的特征矢量建立说话人模型[67]。说话人模型在总体

上描述了说话人的语音特征，虽然这样的描述仅仅是通过训练数据得到的，但一般认为一个未知的说话人语音能够通过与这个模型的匹配被正确分类。在识别状态下，未知说话人的输入语音信号经过特征提取被送到模式匹配模块，该模块采用一个或多个算法计算输入特征矢量序列与每个说话人模型之间的匹配值。识别模式的最后一个模块是决策模块，该模块的输入是模式匹配值，输出则是说话人的身份，甚至还可能有一个置信度值[68,69]。决策的方式取决于具体的识别任务。说话人确认的决策是一个二值决策，或者接受，或者拒绝。另一方面，说话人辨认则有两种可能性。对于面向闭集的辨认，决策的结果是具有最佳匹配值的说话人模型所对应的说话人。对于面向开集的辨认，还需要一个附加的决策，即判断说话人是否为数据库中已登记的说话人。

2.1 特征提取

特征提取是一种将高阶矢量变换为低阶矢量的处理，即是一种映射 $f: R^N -> R^D$，这里，D<<N 。特征提取的意义主要有两个方面。第一，能够根据较少的训练样本数据量建立可靠并具有鲁棒性的说话人模型。模型训练数据量一般以指数级正比于特征矢量的维数，如果数据量太大会引起所谓的维数灾难问题[70,71]。第二个原因就是减少了运算复杂度。

对说话人识别来说，好的特征参数必须具有以下特点：（1）类间耦合度小；（2）类内凝聚度大；（3）容易计算；（4）对刻意掩饰和假冒具有鲁棒性；（5）对失真与噪声具有鲁棒性；（6）与其它特征非相关。前面两个特点要求特征具有最大的可分性，图2.2是一个二维特征集的例子。显然，第二组特征集比第一组的可分性要强很多。即便在第一组特征中，依据特征2也可以很好地将说话人3和其他说话人加以区分。

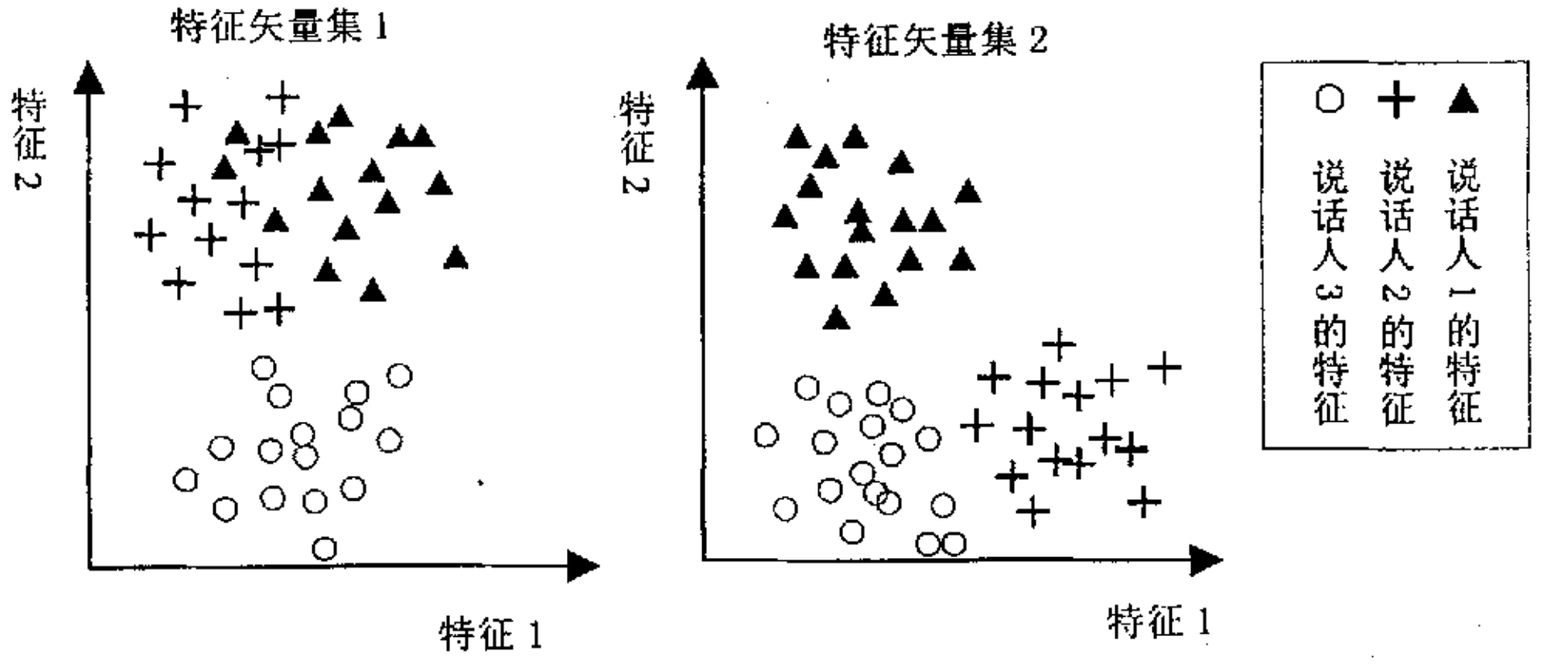


图2.2 两个特征矢量集

特征参数必须是便于提取和测量的，这包括两层含义：（1）特征应该是普遍和自然地存在于语音信号中，这样就可以利用较短的信号样本提取特征参数；（2）特征提取的方法本身应该简单可行。

一个好的特征应该对刻意的发音掩饰、失真和噪声有抵御能力。另外，从语音信号提取的不同特征参数之间应该具有很强的独立性。如果将两个相关的特征组合起来构成新的特征矢量，那么，不仅没有好处，反而会降低识别性能。

但是，迄今为止还没有一种参数能够满足以上所有要求，并且也缺乏一种很好的客观评价方法。相对而言，由于MFCC考虑了人类的听觉感知特性，实验结果显示具有较好的分类性能和鲁棒性[33,72]。

在自动说话人识别中或语音处理中经常被采用的特征参数，如线性预测系数LPC（Linear Prediction Coefficients）、线性预测倒谱系数LPCC（Linear Prediction Cepstrum Coefficients）、Mel频率尺度倒谱系数MFCC（Mel-Frequency Cepstrum Coefficient）、线谱对参数LSP（Linear Spectrum Pair）各分量之间具有很强的非相关性。有一些特征变换方法可以将原始特征变换到一个新的特征空间，在这个新的子空间中，特征参数之间的相关性变小，可分性更强。例如：线性判别分析LDA（Linear Discriminant Ananlysis）[73,74]，KLT（Karhunen-Loeve）变换 [73]，独立成分分析ICA（Independent Component Analysis）[75]。

特征提取和特征选择是两个不同的概念。在特征提取中，新的特征是所有原有特征的函数。相反，对于特征选择来说，是从既有特征集合中选择一个子集，这个子集具有较好的模式分类特性[76]。

2.1.1 线性预测系数 LPC

语音信号线性预测编码LPC（Linear Prediction Coding）分析方法运用全极点自回归AR（Auto-Regressive）模型对语音短时帧进行拟合，并在最小频谱均方误差LSE（Least Square Error）准则下进行最优化。全极点自回归AR模型如下：

$$H(Z)=\frac{G}{1-\sum_{i=1}^{p}a_i Z^{-i}} \tag{2.1}$$

模型系数 a_i　$i=1\sim p$ 称为线性预测系数。如果语音信号的频谱为 $S(e^{j\omega})$，声源信号的频谱为 $E(e^{j\omega})$，则有关系如下：

$$S(e^{j\omega}) = H(e^{j\omega})E(e^{j\omega}) \tag{2.2}$$

$H(e^{j\omega})$ 反映语音频谱的包络，也是语音发音时的声道频谱特性，而 $E(e^{j\omega})$ 反映频谱的细节。

最优化预测系数可以通过 Durbin 叠代算法计算，具体步骤如下：

$$\begin{aligned} &E^0 = r(0) \\ &k_i = \{r(i) - \sum_{j=1}^{i} \alpha_j^{i-1} r(|i-j|) / E^{i-1}, \quad 1 \le i \le p \\ &\alpha_i^i = k_i \\ &\alpha_j^i = \alpha_j^{i-1} - k_i \alpha_{i-j}^{i-1}, \quad j = 1 \sim i-1 \\ &E^i = (1-k_i^2)E^{i-1} \end{aligned} \tag{2.3}$$

以上各步计算进行 $i = 1 \sim p$ 叠代处理，并在叠代结束后得到如下最优预测系数以及其它相关参数：

$$\begin{aligned} &\text{线性预测系数：} \quad a_m = \alpha_m^p, \quad 1 \le m \le p \\ &\text{PARCOR 系数：} \quad k_m, \quad 1 \le m \le p \\ &\text{对数面积比系数：} \quad g_m = \log(\frac{1-k_m}{1+k_m}), \quad 1 \le m \le p \end{aligned} \tag{2.4}$$

2.1.2 线性预测倒谱系数 LPCC

所谓倒谱是语音信号幅度谱对数的Fourier变换，其特点是可以从语音信号频谱中较好地提取谱包络，即可以更好地提取语音信号发音时的声道特性。其定义如下：

$$\log|S(e^{j\omega})| = \log|H(e^{j\omega})| + \log|E(e^{j\omega})| \tag{2.5}$$

$$c(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log|S(e^{j\omega})| e^{j\omega} d\omega \tag{2.6}$$

由于 $H(e^{j\omega})$ 反映谱包络，而 $E(e^{j\omega})$ 反映谱的细节，因此，上式的低次系数对应谱包络。可以利用LPC系数直接推算LPCC系数，具体公式如下：

$$\begin{aligned} &c(0) = \log G \\ &c(n) = \begin{cases} a_n + \sum_{k=1}^{n-1} (\frac{k}{n}) c(k) a_{n-k} & 1 \le n \le p \\ \sum_{k=1}^{n-1} (\frac{k}{n}) c(k) a_{n-k} & n > p \end{cases} \end{aligned} \tag{2.7}$$

许多语音识别和说话人识别实验表明，采用LPCC作为特征参数比LPC系数更加有效，识别性能有很大的提高。

2.1.3 Mel 频率尺度倒谱系数 MFCC

从人的听觉感知特性来看，线性频率尺度与实际的听觉感知灵敏特性并不一致，比较合理的是Mel频率尺度，其与线性频率的关系式如下：

$$Mel(f) = 2595\log_{10}(1+\frac{f}{700}) \tag{2.8}$$

显然，人类的听觉具有很强的语音分辨能力，并且对噪声的鲁棒性极强。因此，计算基于听觉特性的Mel倒谱系数MFCC具有很重要的意义。

通过Mel频率与线性频率的比较，可以得出Mel倒谱系数MFCC，其计算步骤如下：

（1） 在Mel频率分布范围$[0, Mel(f_s/2)]$等间隔地选取L个中心频率 $f_c(i)$　$i=1\sim L$，并以这些中心频率构成一个三角滤波器组，如图2.2.2所示。

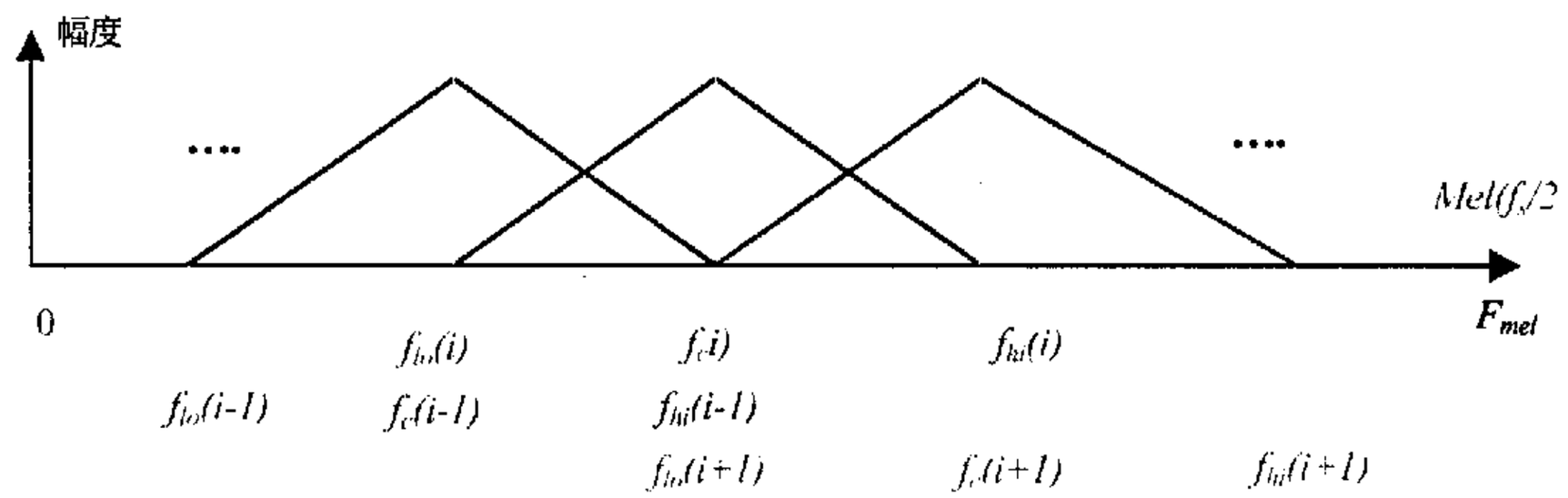


图2.2　Mel频率域的等间隔分布三角滤波器组

其中，f_s 是采样频率，$f_c(i)$ 表示第i个滤波器的中心频率，$f_{lo}(i)$ 表示第i个滤波器的低频边界，$f_{hi}(i)$ 表示第i个滤波器的高频边界。相邻滤波器的中心频率与边界之间存在如下关系：

$$\begin{aligned} &f_c(i) = \frac{Mel(f_s/2)}{L+1}i \\ &f_c(i) = f_{lo}(i+1) = f_{hi}(i-1) \\ &f_{lo}(1) = 0 \\ &f_{hi}(L) = Mel(f_s/2) \end{aligned} \tag{2.9}$$

（2） 求语音信号的离散傅立叶变换 $S(k)$　$k=0\sim N/2-1$，并将其映射到Mel频率域，得到一个如下所示的Mel频率的DFT 序列：

$$S'(f_k) = S'(Mel(\frac{f_s}{N}k)) = S(k) \quad k=1\sim N/2-1 \tag{2.10}$$

（3） 计算Mel频率域各滤波器的输出值，计算公式如下：

$$m(i)=\sum_{k=lo}^{hi}W(k,i)|S'(f_k)| \qquad i=1\sim L$$

$$W(k,i)=\begin{cases}\dfrac{f_k-f_{lo}(i)}{f_c-f_{lo}(i)} & f_{lo}(i)\le f_k\le f_c(i)\\ \dfrac{f_{hi}(i)-f_k}{f_{hi}(i)-f_c(i)} & f_c(i)\le f_k\le f_{hi}(i)\end{cases} \tag{2.11}$$

即第i个滤波器的输出是其通带内所有DFT的和。

（4） 对 $m(i)$ 进一步求离散余弦变换DCT（Discrete Cosine Transform）得到MFCC系数。计算公式如下：

$$C_{mfcc}(n)=\sqrt{\frac{2}{N}}\sum_{i=1}^{L}\log m(i)\cos[(i-\frac{1}{2})\frac{n\pi}{L}] \tag{2.12}$$

通过将Mel频率域的三角滤波器组映射到线性频率域同样也可以求MFCC。显然，n的值可以大于L，但在实际应用中一般取小于L的值，这是因为低频部分表示声道频谱包络，而倒谱计算的目的就是要求声道的频谱包络。

2.2 说话人模型与匹配

说话人模型主要有参数（统计）型模型和非参数（模板）模型两大类[4]。在参数模型中，通过训练数据得到相应的统计分布，而该统计分布的参数是基于某种最大化准则估计计算得到，如高斯混合模型GMM（Gaussian Mixture Model）[59]。非参数模型则是基于最小化准则获得，如基于矢量量化VQ（Vector Quantization）的码本模型CBM（Code-Book Model）[77]。

模式匹配部分的功能是计算未知说话人的输入语音特征矢量和每个模型之间距离或似然度。参数模型一般通过计算后验概率表示似然度，而非参数模型则往往直接计算距离。图2.3描述了对同一组数据以两种不同的模型尺寸参数训练得到的CBM和GMM说话人模型。

2.2.1 基于 VQ 的码本模型 CBM

在CBM码本模型设计方法中，首先采用LBG等聚类算法[78]对原始训练样本的特征矢量进行聚类，形成K个互不相干的集群。每个集群由一个中心特征矢量表示，该中心特征矢量是对应集群所有特征矢量的平均值，称为码字，所有的码字集合 $\{C_1,C_2,\ldots,\ C_k\}$ 构成说话人码本模型。与训练样本集中特征矢量数目相比，码本的大小（码字数目）要小得多。显然，码字的统计分布与原始训练样本特征矢量的分布是一样的[79]。因此，在保持原有分布基本信息的基础上，码本大大减少了所需要处理的数据量。

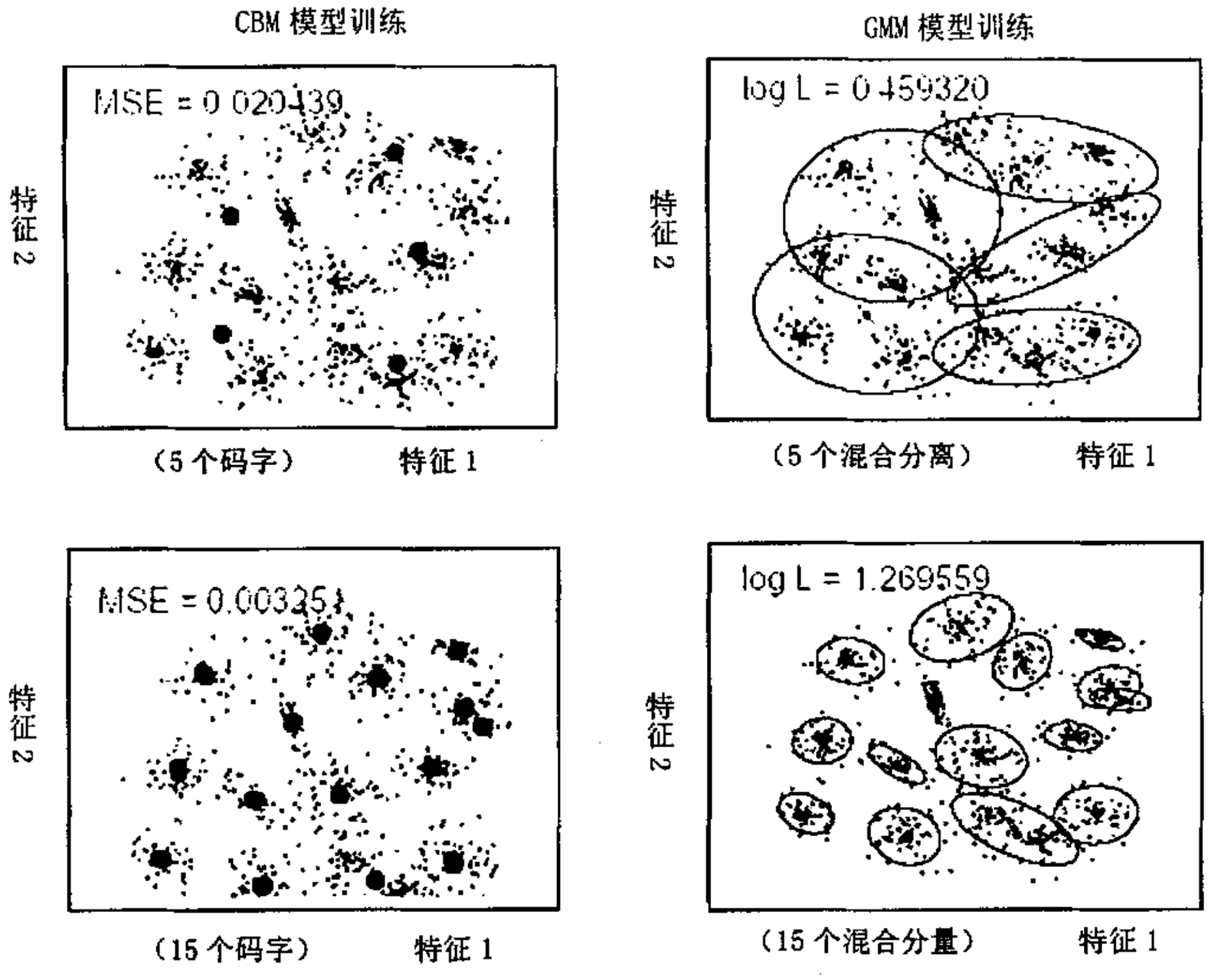


图2.3 说话人识别的CBM与GMM模型训练示例

码本模型设计涉及两个基本问题：（1）生成码本的方法；（2）码本的大小。关于码本大小的问题，一般认为增加码本尺寸可以减少误识率[60,77,80]。但是，如果码本尺寸太大，则会产生过学习问题，即码本过多地依懒于训练样本数据，而不能反映数据的一般统计分布，并且，码字之间的相关性增强。

码本生成方法有两大类，即无监督学习算法与监督学习算法。在无监督学习算法中，每一个说话人模型的训练是相互独立的，而在有监督学习训练算法中，考虑了码本之间的相关性，并使这种相关性最小化。通常会采用无监督学习训练模型，因为需要人为干预的内容较少。He & Liu [80] 提出了一种监督学习算法——群矢量量化的方法GVQ（Group Vector Quantization）。 这一算法的思路是首先独立地训练每一个模型，随后对它们进行优化，使相互之间的差异增强。 无监督学习训练算法中最流行的是LBG（Linde-Buzo-Gray）算法[78]。该算法需要预先设置码本的大小K，然后从原始训练样本数据中选择K个矢量作为初始值，不断地叠代优化，直到码本中各码字保持不变。

Kinnunen,等对矢量量化的码本设计方法进行了分析[81]，发现不同的算法对识别性能的影响并不大。一个解释是，从相互交叉的语音信号帧中提取的特征矢量也许并不存在聚类结构，

但存在一个连续的统计分布[82]。所以，码本的训练更象是从原始样本数据中进行的取样，而不是去找到一个聚类结构。所以，具体的聚类算法的选择不是至关重要的。

基于VQ码本模型CBM的说话人识别中，模式匹配的一个典型描述是两个矢量集合 $X=\{X_1,X_2,\quad X_T\}$，$C=\{C_1,C_2,\quad C_K\}$之间的量化失真。说话人语音的一个特征矢量 X_i 与码本 $C=\{C_1,C_2,\quad C_K\}$ 之间的量化失真 $D(X_i;\ C)$ 定义如下：

$$D(X_i;\ C)=\min_{C_j\in C}\ d(X_i;\ C_j) \tag{2.13}$$

这里，$d(X_i;\ C_j)$ 表示输入语音信号的特征矢量 X_i 与码字 C_j 之间的失真度，量化失真 $D(X_i;\ C)$ 取其所有 $d(X_i;\ C_j)$，$\forall j$ 中的最小值。通常采用Eculidean距离表示，不仅因为其距离空间的直观性，而且对于倒谱系数LPCC作为特征参数的情况，Eculidean距离反映了语音信号短时功率谱的差，能够很好地表达频谱失真度。其它的失真测度，例如，Itakura 和 Mahalanobis 距离测度[74]，作者提出的互信息测度[83]等也是可以应用的失真测度。

输入语音信号特征矢量序列与CBM模型之间的总体失真由平均量化失真测度表示，其定义如下：

$$D(X;\ C)=\frac{1}{T}\sum_{i=1}^{T}D(X_i;\ C) \tag{2.14}$$

显然，如果 $X\subseteq C$，则 $D(X;\ C)=0$。两者匹配得越好，相应的平均量化失真度越小。平均量化失真的计算如图2.4所示，可以看出这一测度不是对称的，即 $D(X;\ C)\neq D(C;\ X)$。另外，以上测度计算中有关T的除法可以忽略，因为在具体判决中该项无意义。

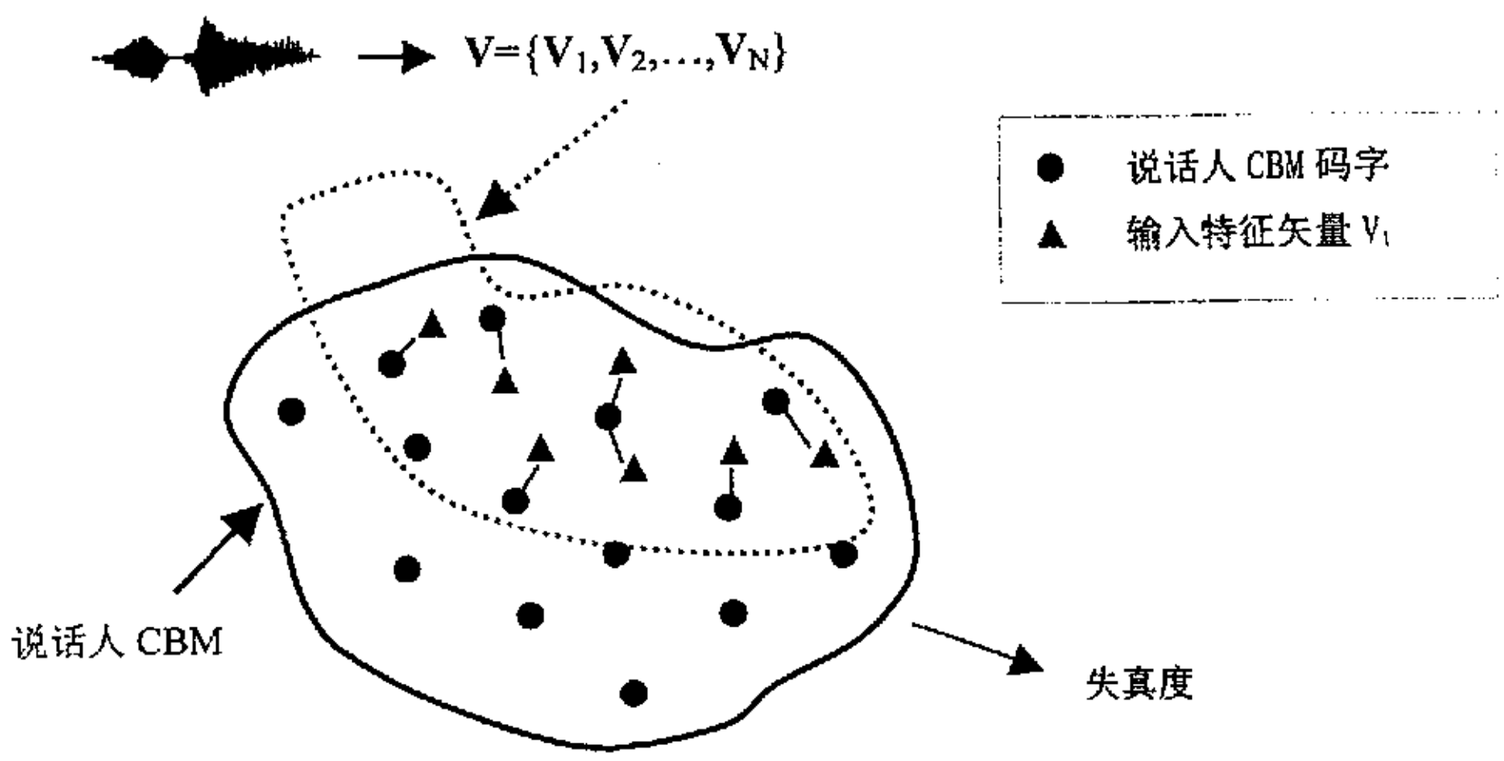


图2.4 基于CBM 码本模型的模式匹配失真度计算

许多研究人员在基本VQ算法的基础上根据实际应用情况作了修改。例如，Kinnunen[81,82]等通过对码本中各码字赋予一个分离度权值的方法，使拥有相近码字的不同说话人之间匹配时，这些码字对失真度的贡献很小。在这些不同的算法中，分离性训练和局部归一化距离测度[84]是较有效的方法，而且，两种方法可以组合在一起使用。

2.2 2 基于 GMM 的说话人识别

高斯混合模型GMM的实质是用若干高斯概率分布的混合加权组合去拟合语音信号的实际概率分布，然后运用 Bayes 最大似然准则进行识别判决[59]，因此该模型纯粹是一种统计参数模型。GMM 模型对应的混合高斯分布如下：

$$p(X \mid \lambda) = \sum_{i=1}^{M} p_i b_i(X) \tag{2.15}$$

其中 p_i 是混合系数，而且满足条件 $\sum_{i=1}^{M} p_i = 1$；$b_i(X)$ 是高斯概率分布，其相应的均值为 μ_i，协方差为 W_i；M 是混合成份数；λ表示 GMM 模型的系数集合，由混合系数 p_i，高斯概率分布均值 μ_i 和协方差矩阵 W_i 构成，共有 M 组。

设有 N 个说话人的 GMM 模型,采用 EM（Expectation-Maximum）算法进行训练[59]得到模型系数集 λ_k， k = 1,2,...,N，在识别时运用 Bayes 最大似然准则对输入语音信号 X 进行识别，判决准则如下：

$$\lambda^* = Arg \max_k \quad p(X \mid \lambda_k) \tag{2.16}$$

即，如果说话人的 GMM 模型(λ^*)具有最大概率值，则识别结果为对应λ^*的说话人。

给定一个说话人的训练语音，就可以采用 EM 算法估计 GMM 的参数。设某说话人的训练特征矢量序列为 $X = \{X_t, t = 1,2,...,T\}$，它对于模型λ的似然度可表示为：

$$p(X \mid \lambda) = \prod_{t=1}^{T} p(X_t \mid \lambda) \tag{2.17}$$

训练的目的就是找到一组参数λ，使 $p(X \mid \lambda)$ 最大，即：

$$\lambda = \arg \max_{\lambda} \ p(X \mid \lambda) \tag{2.18}$$

这种最大参数估计可利用 EM 算法的一种特殊形式，通过迭代得到。

(1) 混合分量：

$$p_i = \frac{1}{T} \sum_{t=1}^{T} p(i \mid X_t, \lambda) \tag{2.19}$$

(2) 均值：

$$\mu_i = \frac{\sum_{t=1}^{T} p(i \mid X_t, \lambda) X_t}{\sum_{t=1}^{T} p(i \mid X_t, \lambda)} \tag{2.20}$$

(3) 方差：

$$W_i = \frac{\sum_{t=1}^{T} p(i \mid X_t, \lambda) X_t X_t^T}{\sum_{t=1}^{T} p(i \mid X_t, \lambda)} - \mu_i \mu_i^T \tag{2.21}$$

(4) 训练数据落在假定的隐状态 i 的概率：

$$p(i \mid X_t, \lambda) = \frac{p_i b_i(X_t)}{\sum_{k=1}^{M} p_k b_k(X_t)} \tag{2.22}$$

(5) 训练中高斯分量的分布为：

$$b_i(X \mid \lambda) = \frac{1}{(2\pi)^{U/2} |W_i|^{1/2}} \exp\{\frac{(X-\mu_i)^T W_i^{-1} (X-\mu_i)}{2}\} \tag{2.23}$$

在GMM的概率计算中，协方差矩阵 W_i 一般假设为对角矩阵，或按对角矩阵的方式计算行列值，以简化运算量。实验证明这种假设对实际识别性能的影响很小，特别是在采用MFCC参数的情况下[48]，由于该参数经过了DCT正交变换，各参数之间的相关性很小，可以认为相应的协方差矩阵 W_i 是对角矩阵。

2.2.3 其它的说话人模型

虽然GMM模型能够较好地描述信号的统计分布，但EM算法比较复杂。例如，必须设置模型参数的极小值以避免出现数据过小的问题[59]。如果特征空间的维数很高，高斯分布概率密度的计算就容易产生溢出。显然，在实际应用中存在一个特征矢量维数极大值的限制。矢量量化VQ方法没有这些问题，但其一个基本的缺陷就是聚类不能有覆盖，因此，码字所表示的概率密度函数不是连续的。值得注意的是，GMM是VQ方法的一个扩展，并且，各个聚类是可以覆盖的。

当然，VQ与GMM相结合的方法也是可行的。首先，LBG算法将特征空间分解为K个互不相连的聚类。然后，根据聚类中包含的特征矢量和码字计算协方差矩阵。混合权值可以依据训练矢量中属于该聚类的特征矢量数决定。实验证明[85,86,87]，以上这种方法具有与GMM相

似的说话人识别性能，但实现起来要容易的多。

已经提出了一些应用于文本无关说话人识别的其它说话人模型，例如，神经网络、单高斯（monogaussian）模型、支持向量机SVM（support vector machine）、决策树等。一些模型方面的实验评价在文献[60,67]中。另外，群分类器（classifier ensembles）也是目前流行的方法之一，其基本思想是使模型能够很贴切地与训练数据一致，并将每一个分类器的局部得分与最终得分结合起来。本文提出多模板模型MTM和全特征矢量集模型CFC分别应用在基于文本和文本无关的说话人识别中，实验中显示出了较好的识别性能。

2.3 决策与判决

说话人识别系统的最后一个处理模块是判别决策。特征提取和模式匹配对不同的识别任务而言都是一样的，而判别决策的方法根据不同的任务而有所不同。设说话人k对应的模型为 S_k，则系统的说话人模型库为 $S=\{S_1,S_2,...,S_N\}$。如果未知说话人的输入语音 X 与模型 S_k 的匹配得分（如似然度）为 $Score(X;S_k)$，那么可以认为该得分越高，说明相互之间的匹配越好。当然，对于基于距离空间的分类器，该得分应为距离的倒数。

面向闭集的说话人辨认，判别决策的结果为在所有说话人模型中具有最大得分的第 i^* 个说话人，其可以由下式给出：

$$i^* = \arg\max_i \quad Score(X;S_i) \tag{2.24}$$

在说话人确认的情况下，系统需要作出的判别决策是是否拒绝或接受申请者，此时，系统的判别决策规则如下：

$$Score(X;S_i) \begin{cases} \geq \theta & \text{接受说话人} \\ < \theta & \text{拒绝说话人} \end{cases} \tag{2.25}$$

这里，θ 表示判决门限。判决门限可以设定为对所有说话人一样，也可以设定为不同的说话人模型采用特定的门限值。在实际的说话人确认系统的设计中应该调节判决门限使系统的错误接受FA（false acceptances）概率和错误拒绝FR（false rejection）概率一致。FA是指错误地接受了一个不该接受的说话人，而FR是指错误地拒绝了一个应该接受的说话人。门限 θ 的选择与FA和FR是有关系的。当门限提高时，FR的概率就升高，而FA的概率将降低；反之，当门限降低时，FR的概率将降低，但FA的概率将升高。具体的判决门限调节方法取决于具体的识别任务[88]

在面向开集的说话人辨认中，未知说话人有可能不属于系统的说话人数据库，因此，相应的判决准则如下：

$$\text{判决结果} = \begin{cases} \text{i}^* & i^* = \arg\max_i \quad Score(X;S_i) \quad and \quad Score(X;S_{i^*}) \geq \theta \\ \text{No} & i^* = \arg\max_i \quad Score(X;S_i) \quad and \quad Score(X;S_{i^*}) < \theta \end{cases} \tag{2.26}$$

如果最大匹配得分超过设定的判决门限，则相应的识别结果为该模型对应的说话人。反之，系统将不输出任何识别结果。

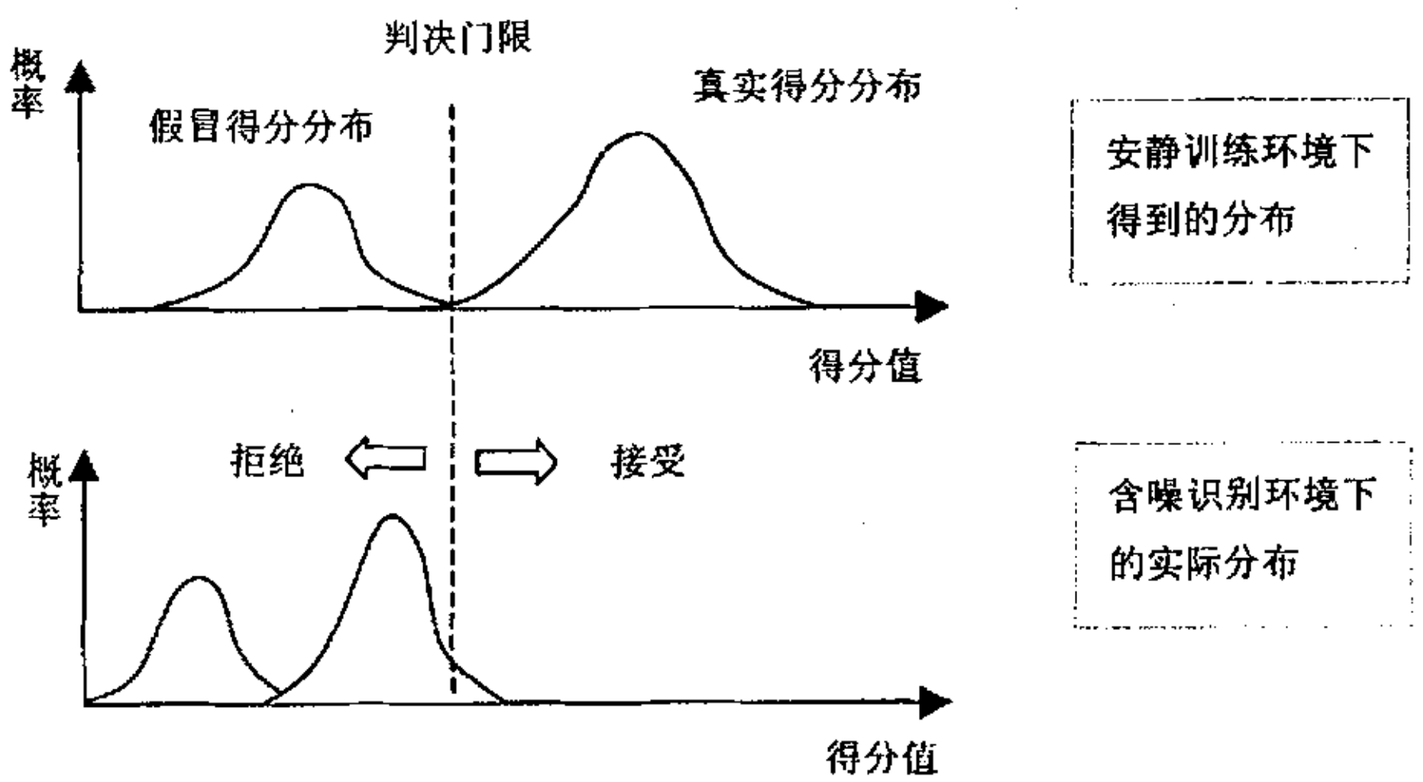


图2.5　说话人确认中没有归一化而出现的问题

图2.5 显示了两个说话人的语音特征分布在训练和识别环境的变化情况，这种变化往往是由噪声引起的，并且会使得匹配得分发生变化，从而引起识别性能大幅下降。因此，实际的说话人识别系统中匹配得分 $Score(X;S_k)$ 应该归一化 [88]。在归一化过程中，一方面要考虑同一说话人不同时刻由各种噪声引起的发音变化，另一方面要考虑不同说话人之间的发音变化。对于说话人辨认系统，由于匹配得分计算中对所有模型都是针对相同的输入语音，因此，仅仅需要进行针对模型进行归一化处理。例如，在基于距离空间匹配时采用平均帧距离计算得分。在说话人确认系统中，当所有模型采用相同判决门限时，归一化是必须的，一个主要方法是将说话人声明的模型的匹配得分除以其他说话人模型的平均匹配得分，即

$$NormScore(X;S_i) = \frac{Score(X;S_i)}{\frac{1}{E}\sum_j Score(X;S_j)} \qquad j \neq i,\ E\text{为模型数} \tag{2.27}$$

归一化处理的方法有很多，不同的方法对识别性能带来的影响也不同，一般应该根据具体的识别任务选择合适的方法。

第三章　语音信号处理与互信息理论基础

本章提要：

- 语音的声学感知特性分析：时频分析，语谱图，Mel、Bark 与 ERB 频率尺度
- 语音的短时谱分析：Fourier 变换，频谱的动态特性，小波变换多分辨率分析
- 说话人的个性特征：分析说话人之间的固有差异
- 互信息理论：描述随机信号之间相互携带的信息量

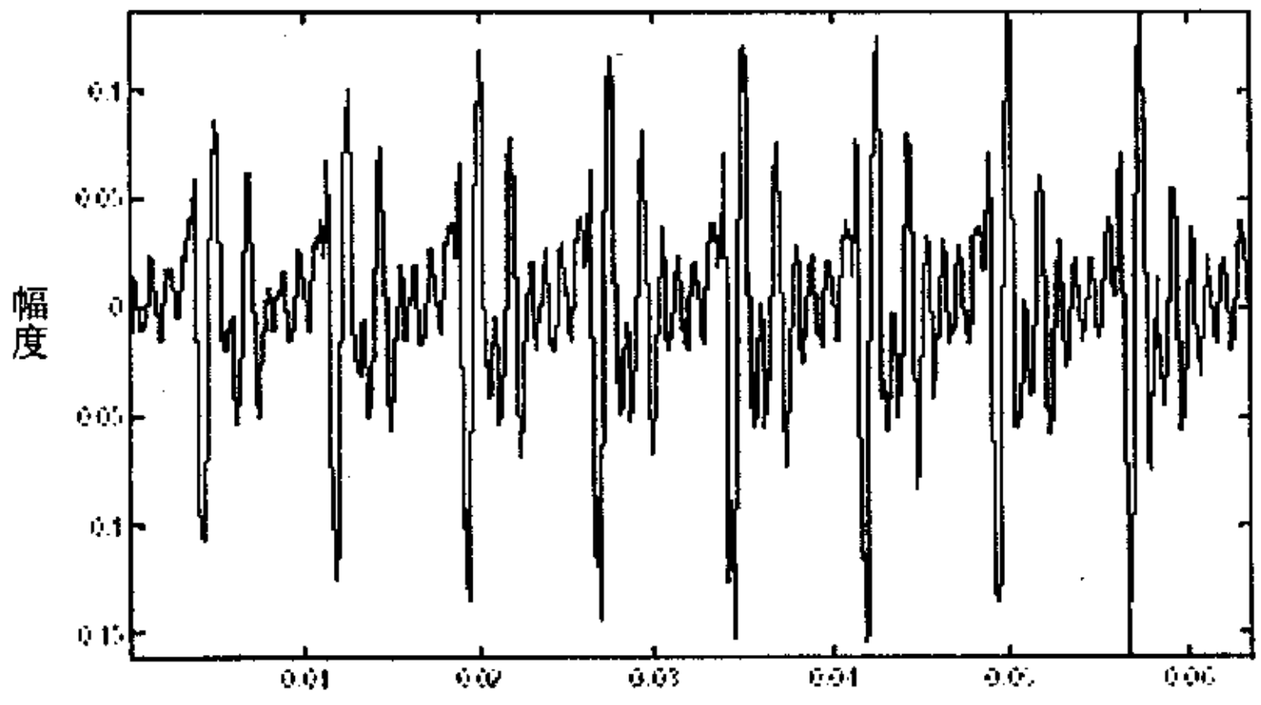
语音信号是一种非平稳的随机信号，但可以认为是短时平稳的。一般认为，在15ms~30ms较短时间内，语音信号的统计特性一致，因此，短时分析是一种最基本的处理手段。

互信息反映了随机变量相互之间携带的信息量，其值越大，说明两个信号越相似，反之则差异越大。互信息的计算必须依据统计特性进行。

说话人识别的研究必须充分了解说话人的个性特征表现，运用声学感知分析、信号处理、模式识别等多种手段分析解决。

3.1 语音的声学感知特性分析

语音信号可以通过时域和频域两个途径进行分析，图3.1表示了一个时域信号波形以及相对应的短时频谱。



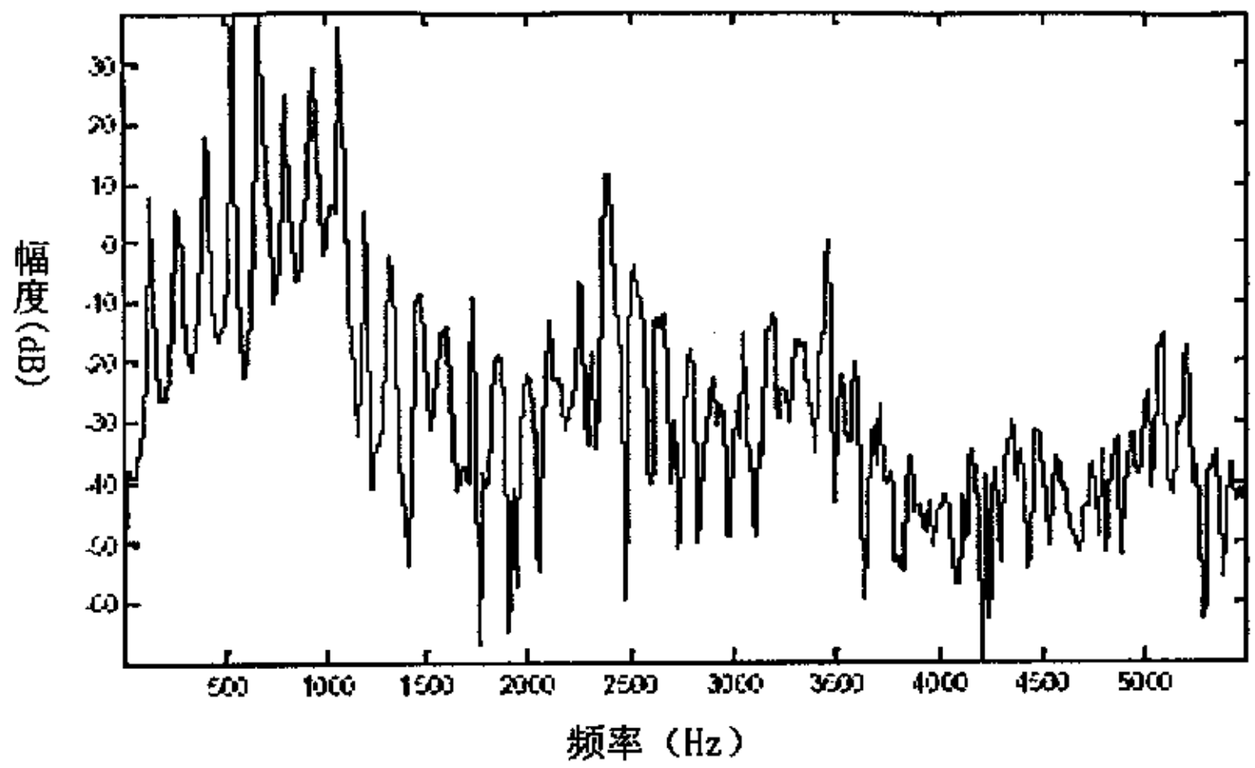


图3.1　一个男性说话人所发元音[a]的时域信号波形和频谱

3.1.1 语谱图分析

语音信号随时间变化的频谱，即时-频分布曲线就是语谱图，它是分析语音信号声学语音特征的有效方法。图3.2是语音信号“上海”的波形以及相应的语谱图。

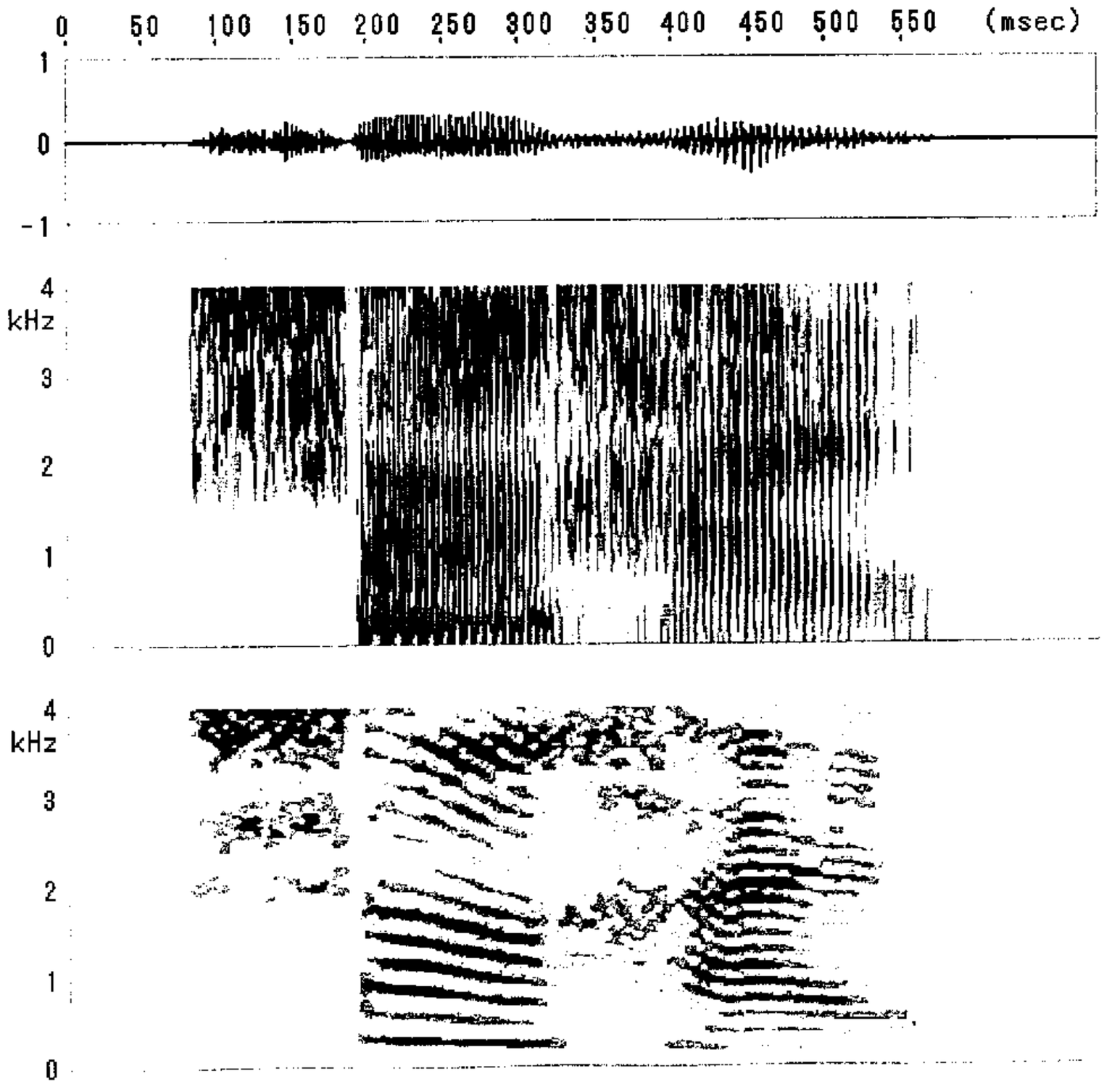


图3.2　女声“上海”的语音信号波形与语谱图

语音信号波形表示发音过程的气压变化，而语谱图反映信号频谱随时间的幅度变化情况。语谱图不仅可以提供语义信息，而且可以揭示说话人的个性特征。例如，从图3.2中能够分辨出元音[a]的区域，根据基音频率可以判断说话人为女性。

语谱图的时间分辨率和频率分辨率是一对矛盾。如果提高时间分辨率，即使用较短的分析窗，则由于短时窗频谱带宽增加而降低了频率分辨率，反之亦然。两者基本呈现线性反比的关系，例如，一个20ms的时间分辨率对应的频谱分辨率大约是50Hz。在信号处理和物理学理论中，时间-频率分辨率的关系由不确定定理来概括[89,90]。

语谱图有宽带语谱图（图3.2中）和窄带语谱图（图3.2下）之分。宽带语谱图的频率分辨率为300Hz左右，时间分辨率为1/300s=3.33ms。窄带语谱图的频率分辨率为50Hz，时间分辨率为1/50s=20ms[91]。宽带语谱图适合分析共振峰轨迹，而窄带语谱图主要应用在基频F0估计中。

3.1.2 韵律特征分析

韵律控制语音声碉、语调、重音以及语音节律[92,93]。典型的韵律特征信息为基音频率F0轨迹以及短时能量强度，如图3.3所示。

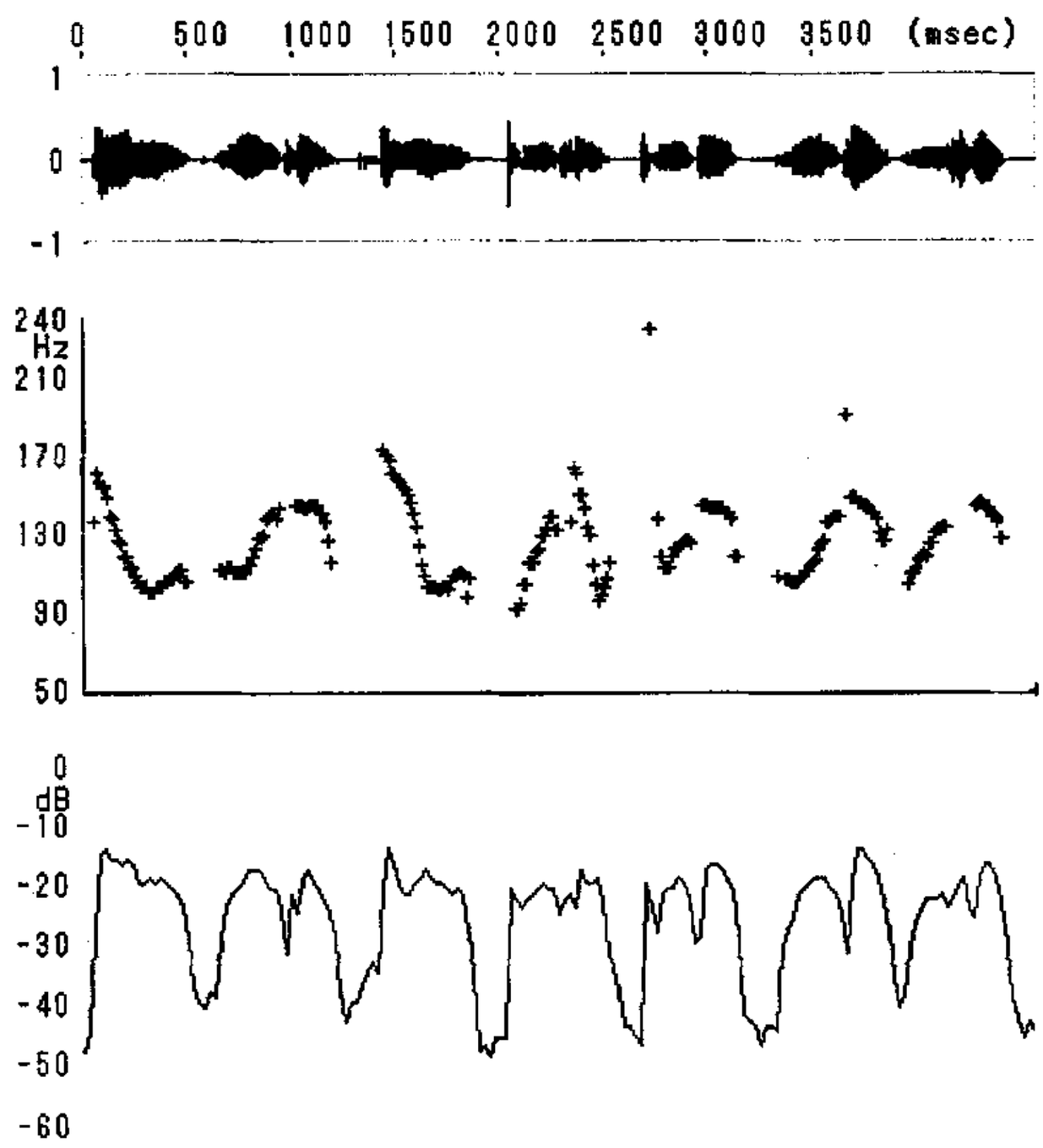


图3.3　韵律参数的声学分析（信号波形、基音轨迹、短时能量）

韵律特征反映了说话的态度以及对听众的情绪，还提供了说话口音、社会地位和所使用的语言属性等。另外，在汉语这样的声调语言中，基音轨迹在反映语言信息上起相当的作用[92]，同样可以反映说话人本身的信息。但是，韵律信息并非都能反映说话人的信息。例如，一个以色列的模仿者可以模仿不同的政治人物[94]，对这位模仿者的三个韵律参数（基音、共振峰以及由声道闭合面积）进行的研究与比较发现，模仿者的基音轨迹与被模仿者的基音轨迹非常接近，并且，第二和第三共振峰也非常接近一致。但是，第一共振峰存在一定的差异，声道的闭合面积情况也不象基音轨迹那样拟合得好。

在说话人识别中运用韵律信息的主要好处是其鲁棒性，对传输信道的噪声具有很好的抗干扰能力[45]，因此，韵律参数对于电话信道的说话人识别系统有特别意义。

3.1.3 感知特性分析

声音的大小并不是与所测量得到的声音强度存在线性关系。例如，声音的强度扩大一倍，但听觉上并没有放大一倍的感觉。所以，通常使用dB（decibel scale）表示两者的关系[38]，其实际含义为两个声音的比较。

$$10\log_{10}(\frac{I}{I_o}) \tag{3.1}$$

这里，I_0是参考声音强度。如果声音的强度I 为两倍I_0那么大，则声音的感知值大约是+3dB。

基音频率F0是浊音发音时的声带振动频率，基音频率的倒数则称基音周期。即使语音信号的基音频率区域被滤波器滤掉，但人类仍然能够感知到其存在[48]。人类的听觉对声音的感知是基于对数尺度，而非线性尺度。已经发现，在高频区域，F0必须有更大的变化才能使人类听觉系统感觉到两个音调之间的差异。Mel是一个基本的听觉感知频率单位，最初是通过听觉测试确定的，并已提出了若干种分析模型来逼近Mel尺度。例如，Fant提出了映射公式(3.3)，这与公式(2.8)有一定区别。

$$F_{mel} = 1000\log_2(1+\frac{F_{Hz}}{1000}) \tag{3.2}$$

不同频率的相对幅度决定了整个频谱形状。如果基音频率保持不变，而高频谐波的相对幅度发生变化，则声音听起来就有不同的音色。所以，音色是频谱形状的一种感知特性，这一特性在说话人识别中是一个重要特征。例如，广泛使用的Mel倒谱特征反映了感知频谱包络形状。

人类听觉机制的研究发现，在人类听觉系统的开始部分，输入的声音被分配到若干频带，

频带内的两个频率是不可区分的。这些频带称为临界频带，人耳对每一临界频带内的能量进行平均，从而形成一个原始信号的压缩表示。这一发现给出了如何在语音识别或说话人识别系统中进行预处理的思路。

有许多临界频带的逼近方法已经被提出，其中一个知名的映射方法是Bark尺度[91]，其Bark尺度分析公式之一是由Zwicker和Terhardt[95]提出的，如下：

$$F_{bark} = 13\tan^{-1}(\frac{0.76F_{Hz}}{1000}) + 3.5\tan^{-1}(\frac{F_{Hz}}{7500})^2 \tag{3.3}$$

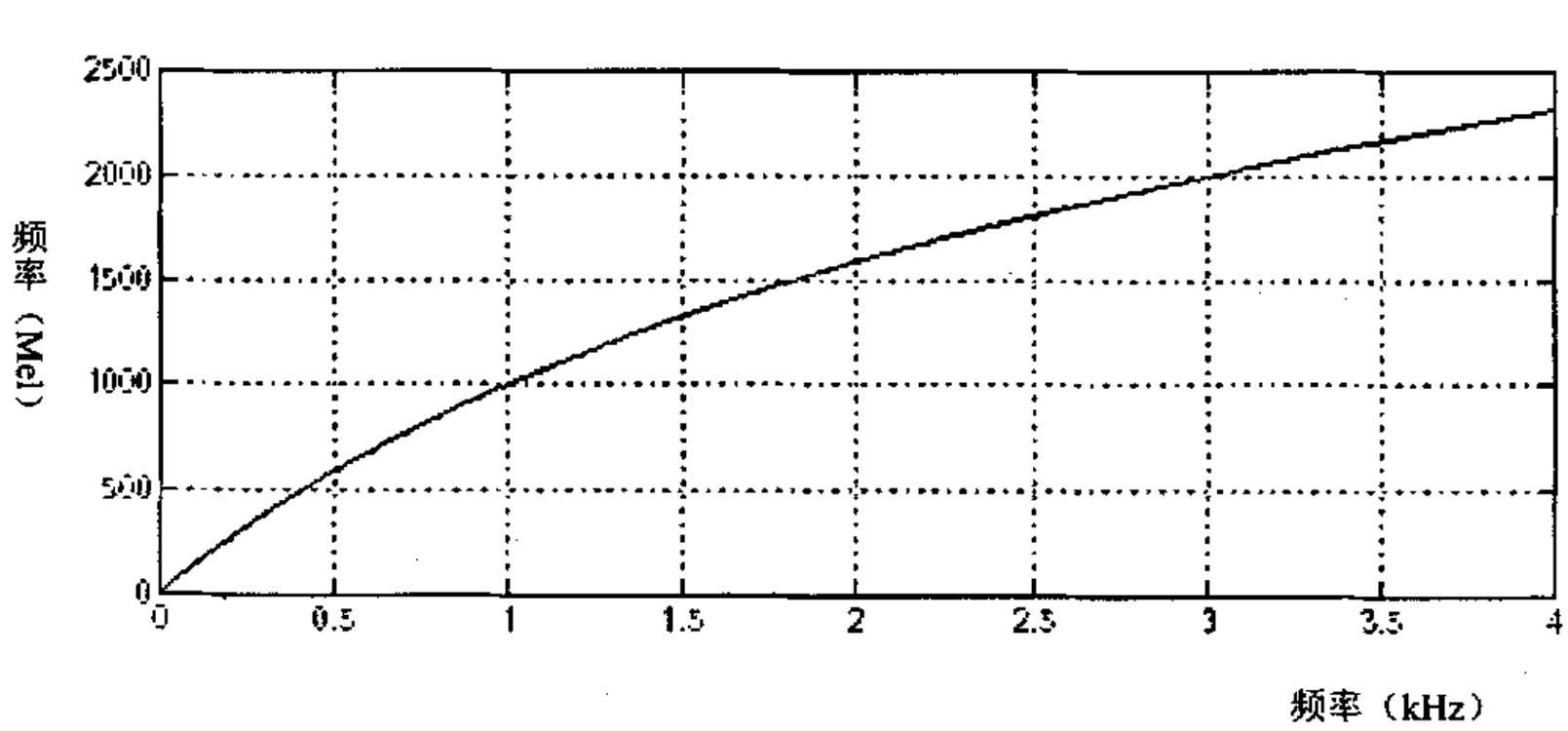
Bark尺度的另一个逼近公式如下：

$$F_{bark} = 6\sinh^{-1}(\frac{F_{Hz}}{600}) \tag{3.4}$$

另一个临界频带的逼近式是ERB（Equivalent Rectangular Bandwidth of the auditory filter）尺度[91]，其定义如下：

$$ERB = 21.4\log_{10}(1 + \frac{4.37F_{Hz}}{1000}) \tag{3.5}$$

有关Mel、Bark以及ERB尺度的解释与比较如图3.4所示，采用的公式分别是（3.2），（3.4），（3.5）。虽然曲线的形状是不一样的，但表示的信息是相同的。在高频部分，两个不同的信号必须有较大的差异才能使人类的听觉对其区分。在低频区，人耳的频谱分辨能力较高。



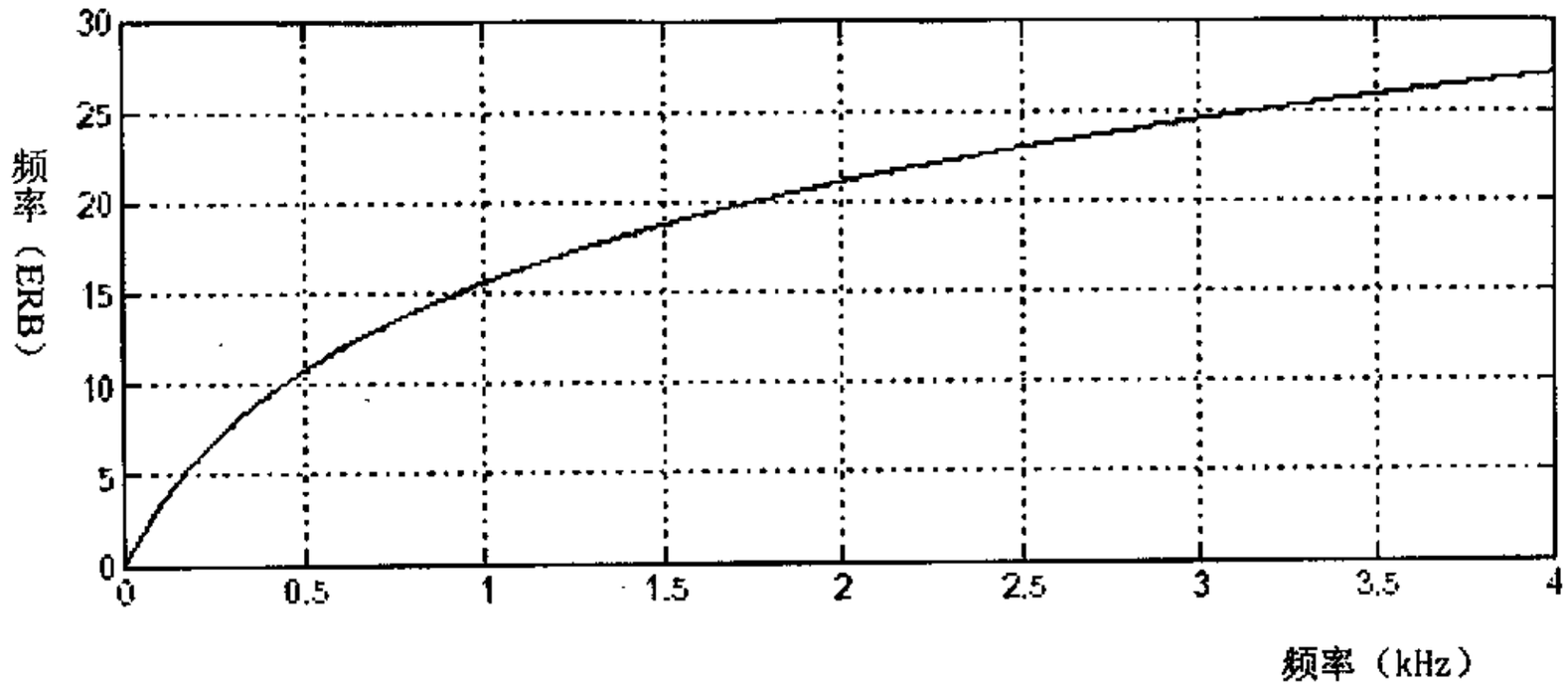


图3.4 Mel , Bark 以及 ERB 尺度

给定一个临界频带的中心频率，该频带的带宽可以由下式计算[95]：

$$BW = 25 + 75(1 + 1.4(\frac{F_{Hz}}{1000})^2) \tag{3.6}$$

基于感知特性的频谱表示方法已经在语音识别中得到了成功的应用，在说话人识别中同样如此，尽管两个任务的目的完全不同。

3.1.4 说话人个性特征分析

说话人的个性特征是一个比较复杂的现象，既与说话人发音器官等先天因素有关，也与其生活、学习等成长经历有关。一直以来，对于将个性特征划分为生理特征和学习特征两大类[18,96]是否充分的问题都有不同的意见。根据Nolan[96]的看法，没有一个声学特征能够保持恒定，因为声道形状本身就是变化的。当然，声道的变化是有限制的，仅仅在一定的限制界限内发生变化。因此，声音是一种比指纹具有更大变化特性的生物特征，它随时间的变化而变化。

（1）音源

喉部具有很大的个性化特征。一般，女性和儿童的声带相对男性而言较小，所以，其基音频率就较高。并且，即便是同一性别的人之间，基音频率也会发生变化，各自具有特定的统计分布。基音频率F0的对数统计分布，特别是均值[97,98]带有很重要的说话人信息。

基音频率是一种个性特征源。声带的紧张直接影响声门脉冲参数，例如，声门闭合的频率，张开的幅度等。对于某些说话人来说，发音时声门会全部闭合，但其他的说话人也许发音时声门从来没有完全地闭合，因此，某些发音听起来如同气声一般[92]。这种现象使得在说话人识别系统中可以检测语音的质量，但要做到可靠的语音质量测量是困难的。

声门脉冲的形状影响频谱斜度，即频谱包络的总体下降斜率。频谱斜度可以根据高频段功率对数谱与低频段功率对数谱的比进行估计计算得到[96]。

（2）声道

声道因人而异。首先，声道尺寸是不一样的[99]，在不同的性别组之间的差异尤其明显。总体上，男性、女性以及儿童呈现出越来越小的趋势。假设两个说话人的发音器官均相同，唯一的不同是声道的长度（从声门到嘴唇），那么，根据声学理论可以推出这样的结论：声音的共振峰频率与说话人的声道长度成反比。

此外，除了声道的总体尺寸之外，不同说话人之间的发音谐振腔也是不一样的[99]。声道的口腔和咽喉部分随不同的说话人而有所不同，并且，口腔部分谐振腔的长度变化相对咽喉部分谐振腔的长度变化要小。另外，由发音时关节收缩引起的前后共振腔与语音和说话人的特征是相关的。总之，声道的长度和形状都具有个性化特征。

（3）发音差异

Eatock和Mason[100]对125个不同说话人所发音素和音素组的差异性进行了研究，通过对语音波形进行手工标注，并由LPC倒谱系数表示各段语音。他们发现，鼻音和元音的区别最大，爆破音最小，这与其他研究人员的研究结果一致。唯一的不同是，他们的研究表明，摩擦音[s]与元音和鼻音一样具有可分性。

鼻音包含很有效的说话人个性特征[18,96,100,101]，原因是鼻腔不仅有相当的个性特征，而且更重要的是它的恒定性，它的形状和体积不会轻易被改变。因此，从鼻音提取的特征参数具有很好的稳定性，受环境影响很小。但是，由于咽喉以及口腔也是鼻音发音的声学谐振腔，会使得鼻音多少有些变化。当然，鼻音非常容易受感冒等影响[101]。

元音几乎在每个汉字音节中存在，元音的声学特性中，前两个共振峰F1和F2主要决定元音的语音特征，而第三个和更高的共振峰则更多地反映了说话人的特性。这些结论无论在声学研究还是感知研究中都得到了证实[99]。

同一个元音不同说话人的发音，相应的低频共振峰的绝对位置是不同的，但共振峰之间的相对位置变化比较小。所以，元音的低频共振峰同样带有说话人的信息，尽管相对高频共振峰来说，说话人之间的变化要小些。另外，共振峰的分类特性与元音的具体情况有关，例如，前元音的F2携带说话人信息。

研究表明，语音的中频和高频区域频谱对说话人识别来说很重要[99,102,103]。此外，频谱的低频区域F0附近携带有用的信息[102,104]。总体上，频谱的低频部分主要反映的是语音特征，但是，语音信息和说话人信息以复杂的形式包含在整个频谱中[105]，提取说话人信息并不是很简单地利用一个滤波器组就可以解决的。另外，即便是某些频段具有较好的分类特性，但也不是一概而论，而是与具体的音素有关的[102,104]。

（4） 韵律特性的差异

语音韵律特性具有很强的个性化特征，例如，语调、音调、重音、时长与节奏等都能够反映说话人的特征。韵律特性的一个特点是具有很强的鲁棒性，不象频谱那样容易受环境噪声以及传输线路噪声的影响。但是，这些特性却很容易受情绪和不同语音的影响，并且，这种特性也比较容易被模仿。因此，韵律信息不是一种可靠的分类信息。韵律信息的模型建立和检测也比较复杂，例如，基音轨迹的检测仍然没有一种很可靠的、普遍适用的方法。

韵律信息的研究在说话人识别研究的早期就开始了[106,107]，但最近几年人们又开始对它进行了大量的研究[98,108,109]，这主要是因为随着计算机速度的提高以及高速数字信号处理器的出现，说话人识别系统正在逐步从实验室向实际应用方向发展，而一个实际应用的说话人识别系统必须是稳定可靠以及鲁棒性高的系统。

3.2 傅立叶频谱分析

傅立叶（Fourier）变换提供了在频域分析信号频谱特性的手段。例如，可以利用傅立叶变换得到语音信号的语谱图。傅立叶分析的理论基础是将信号看成由一些不同频率的正弦信号的叠加，其变换式如下：

$$S(e^{j\omega}) = \sum_{n=-\infty}^{\infty} s(n)e^{-j\omega n}$$
$$s(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S(e^{j\omega})e^{j\omega n}d\omega \tag{3.7}$$

傅立叶变换是一种有力的分析工具，利用它可以分析一个信号的正弦信号成分（正变换），或者利用信号的频率表示恢复原始时域信号（反变换）。图3.5（a）显示了三个频率分别为15Hz、30Hz和50Hz的正弦信号，图3.5(b)是它们的叠加合成信号以及傅立叶变换得到的合成信号频谱幅度。

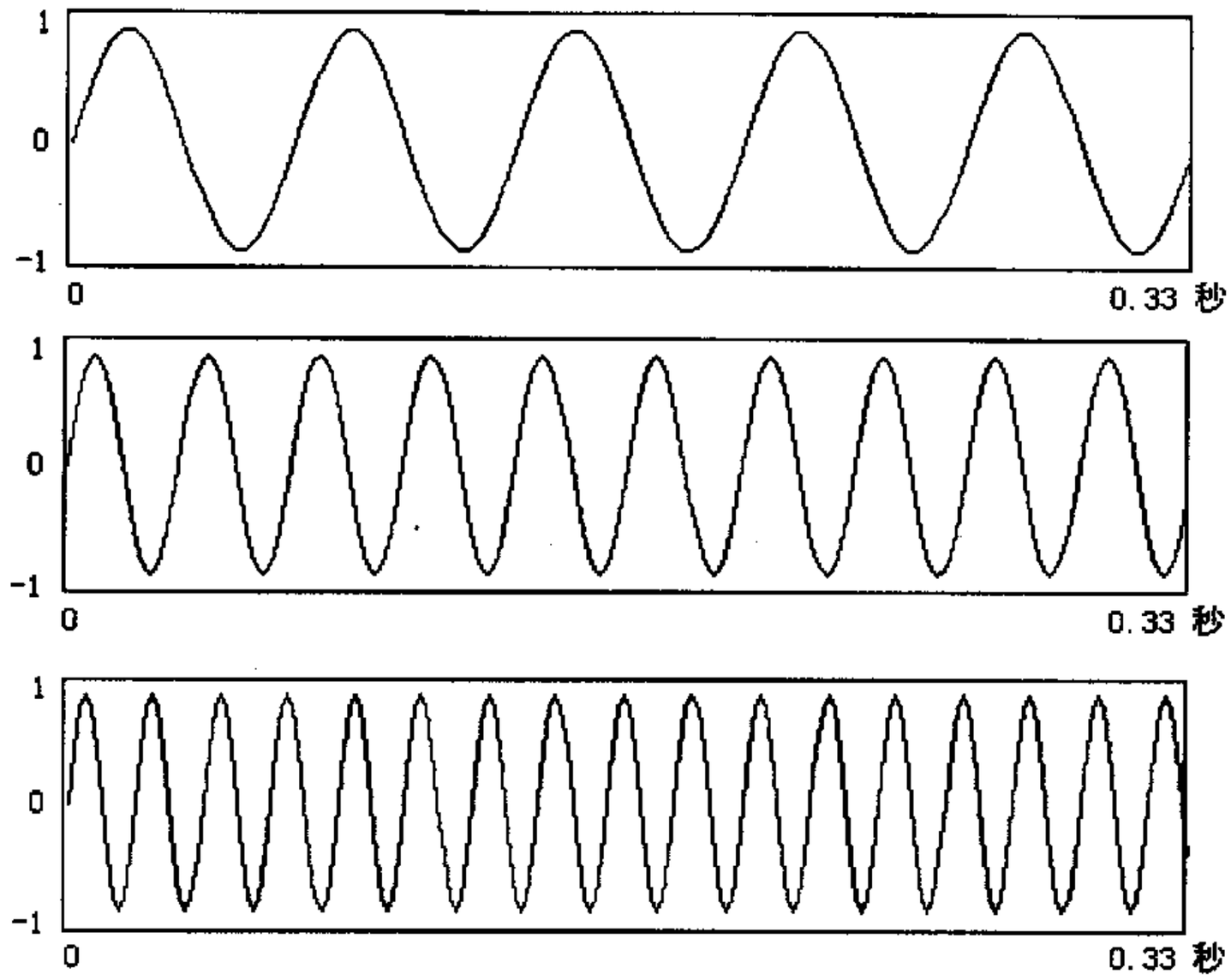


图3.5(a) 三个正弦信号

图3.5(b) 复合信号的频谱幅度

由图看出，根据傅立叶频谱基本上能够正确地分析信号包含的正弦成分，具体应用中，频谱峰值的检测精度还与采样频率和分析窗长度有关。傅立叶频谱是以2π为周期的周期信号，并且，幅度谱在一个周期内对称分布，而相位谱反对称分布。显然，信号的频谱仅仅只需要0~π部分。

对于语音信号来说，传统的频谱分析方法一贯重视幅度谱，对相位谱的利用几乎没有，原因是认为人耳对相位是不敏感的[29,110]。但是，最近的有关研究却认为相位谱对于语音感

知是重要的，例如，Paliwal 和 Alsteris 的研究发现在某些特定条件下相位谱对语音感知而言比幅度谱更有意义[111]。

实际应用中的信号往往是非平稳的随机信号，因此，一般以短时离散傅立叶变换（STDFT: Short-Time DFT）进行动态频谱分析，这样得到的频谱是一个随时间变化的频谱序列，对于语音信号而言就是语谱图（Spectrogram）。

3.3 语音信号短时频谱分析

语音信号呈现时变非平稳的特征。因此，不能简单地对整个语音信号求频谱，否则，信号频谱的时变特征就得不到反映，语音中各音素的频谱表现无法观察和分析。所以，必须将语音信号划分为一系列较短的区段，使得这些区段的信号表现为准平稳信号，一般取15-30ms作为短时分析段，或称短时帧[112]，如图3.6所示。

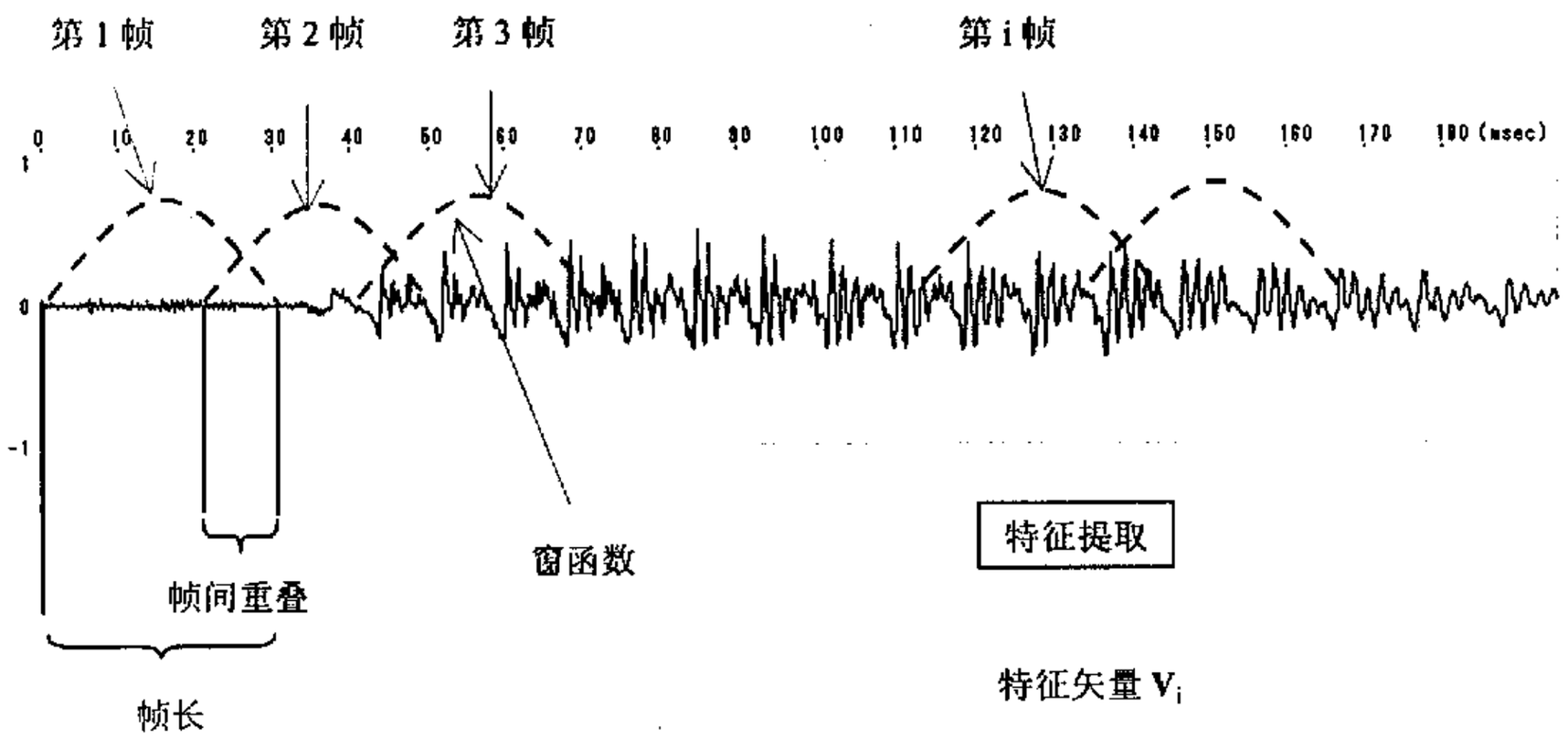


图3.6 语音信号的短时处理分析

具体的短时帧可以通过加窗来实现，即将语音信号 $s(n)$ 乘以短时窗 $w(n-rL)$　　$r=0,1,2,\ldots$。这里，L是短时窗的移动步长，一般为窗长为N的30%~50%。

3.3.1 短时窗的功能

短时窗的作用是实现分段，窗函数 $w(n)$ 与原始信号 $s(n)$ 逐点相乘得到短时帧 $x(n)$。显然，窗函数会对短时帧频谱产生影响，短时谱将是原始信号频谱与短时窗频谱的卷积：

$$X(e^{j\omega})=S(e^{j\omega})*W(e^{j\omega}) \tag{3.8}$$

根据卷积理论，短时窗的频谱如果表现为脉冲形式，则其频谱对短时谱的影响最小。一般而言，希望短时窗频谱具有窄的主瓣和小的旁瓣，但实际上两者不可完全兼顾。主瓣窄化的同时旁瓣也提升了。总体来说，一个窗函数在边缘处最好逐步衰减，以减小信号分段而引起的不连续影响。

最简单的短时窗函数为矩形窗，其表达式如下：

$$w_{\text{Rectangle}}(n)=\begin{cases}1 & 0\le n\le N-1\\ 0 & \text{其它}\end{cases} \tag{3.9}$$

矩形窗虽然在时域上不会修改原始信号的值，但其频谱中包含的较大旁瓣成分会引起短时谱的泄漏现象。因此，实际应用中往往由哈明窗（Hamming Window）来替代，其表达式如下：

$$w_{Hamming}(n)=\begin{cases}0.54-0.46\cos(\dfrac{2\pi n}{N}) & 0\le n\le N-1\\ 0 & \text{其它}\end{cases} \tag{3.10}$$

哈明窗具有较宽的主瓣，但有较低的旁瓣。图3.7显示了两中短时窗的时域波形与相应的频谱。可以看到，哈明窗使得短时帧的边缘被逐步减小了。

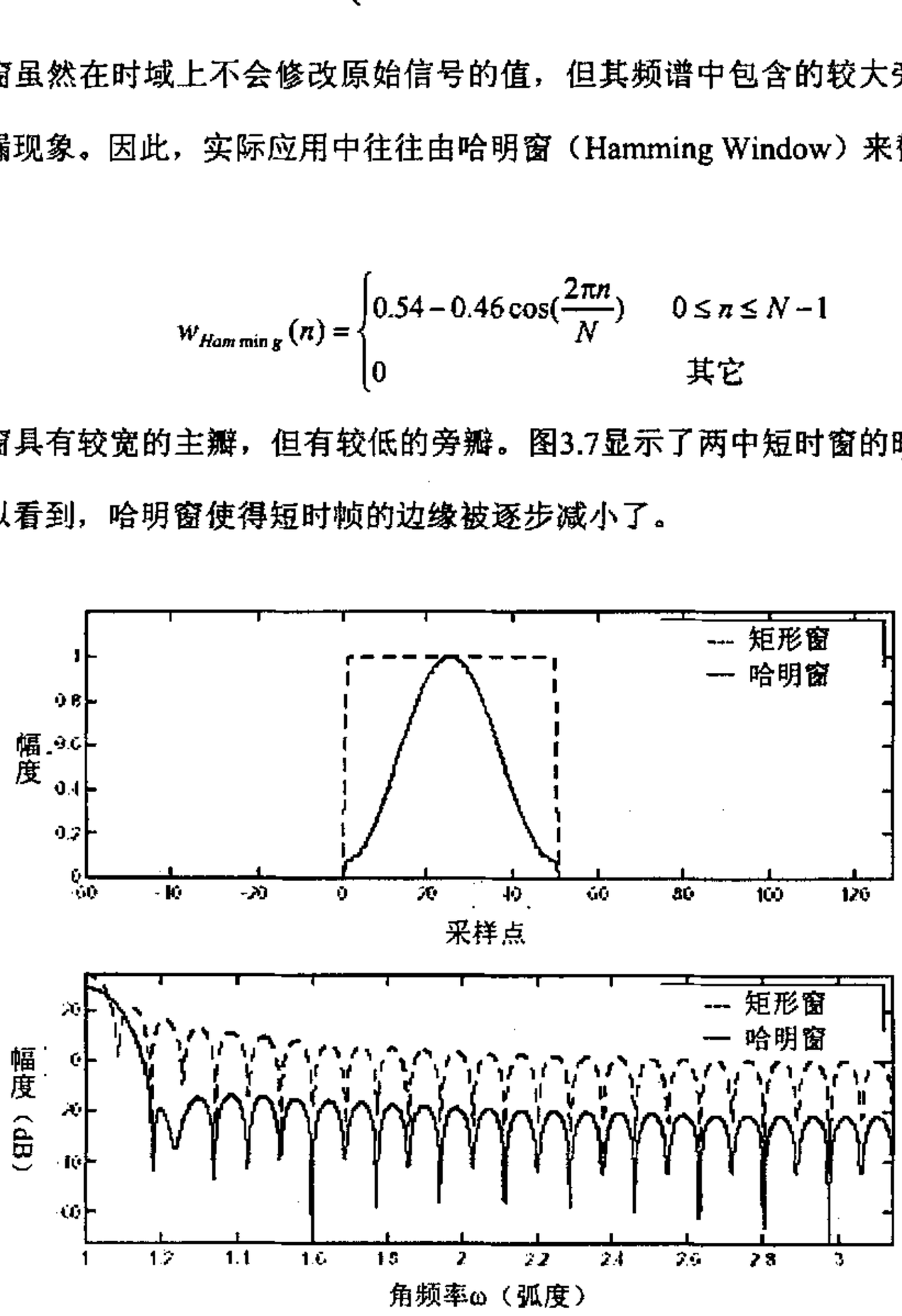


图3.7 矩形窗与哈明窗的时域波形与幅度谱

图3.8和图3.9表示了对同一说话人的浊音和清音分别用两种短时窗分段形成的短时帧和短时谱，可以看到哈明窗的泄漏现象要小很多。矩形窗使得基音F0的谐波能量向临近谐波泄漏，因而谐波被平滑了。哈明窗几乎没有这种情况，因此谐波成分较明显。清音没有基音，

当然也就没有谐波存在，但泄露现象还是可以看到。

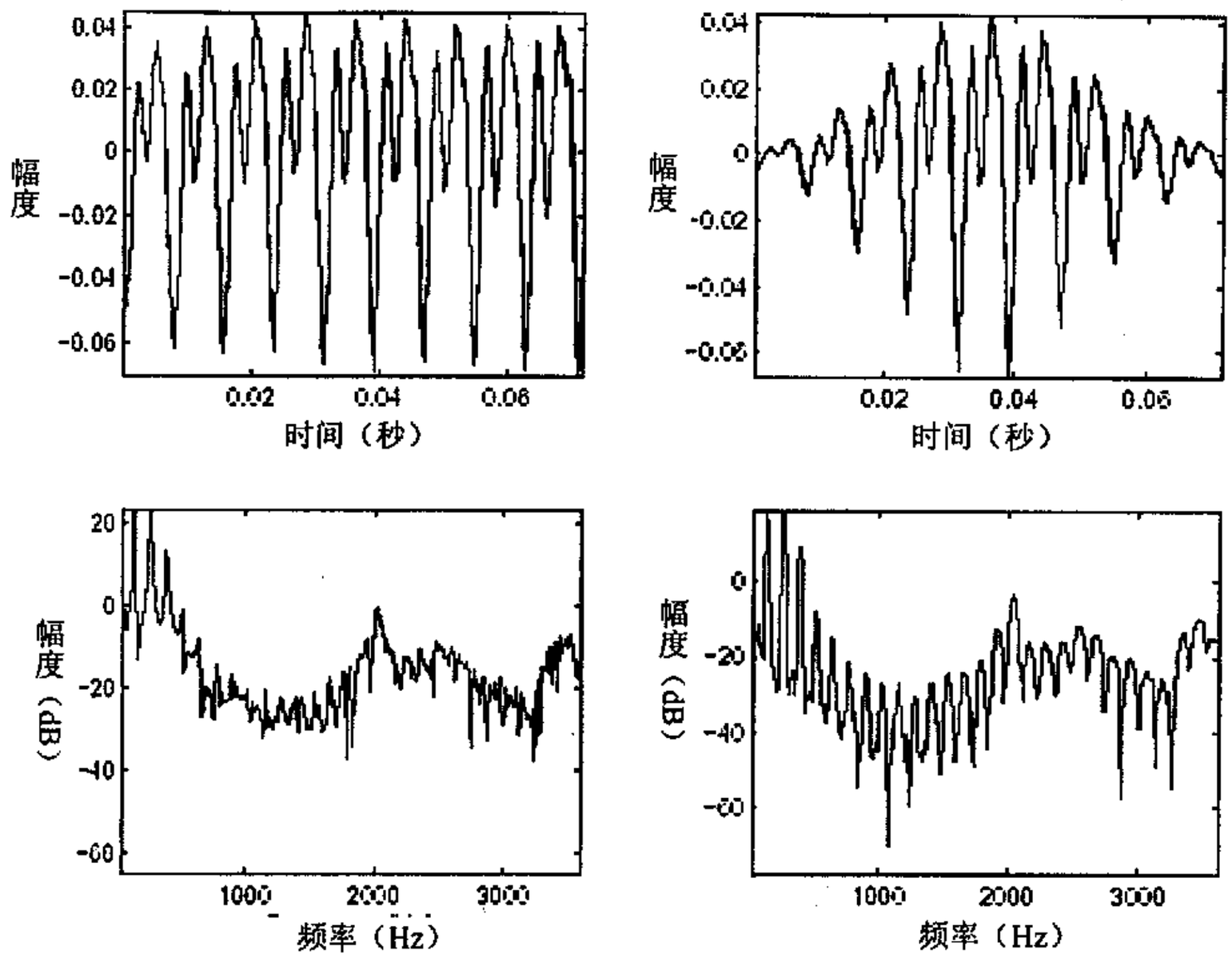


图3.8　浊音信号分别加矩形和哈明窗的时域和频谱波形

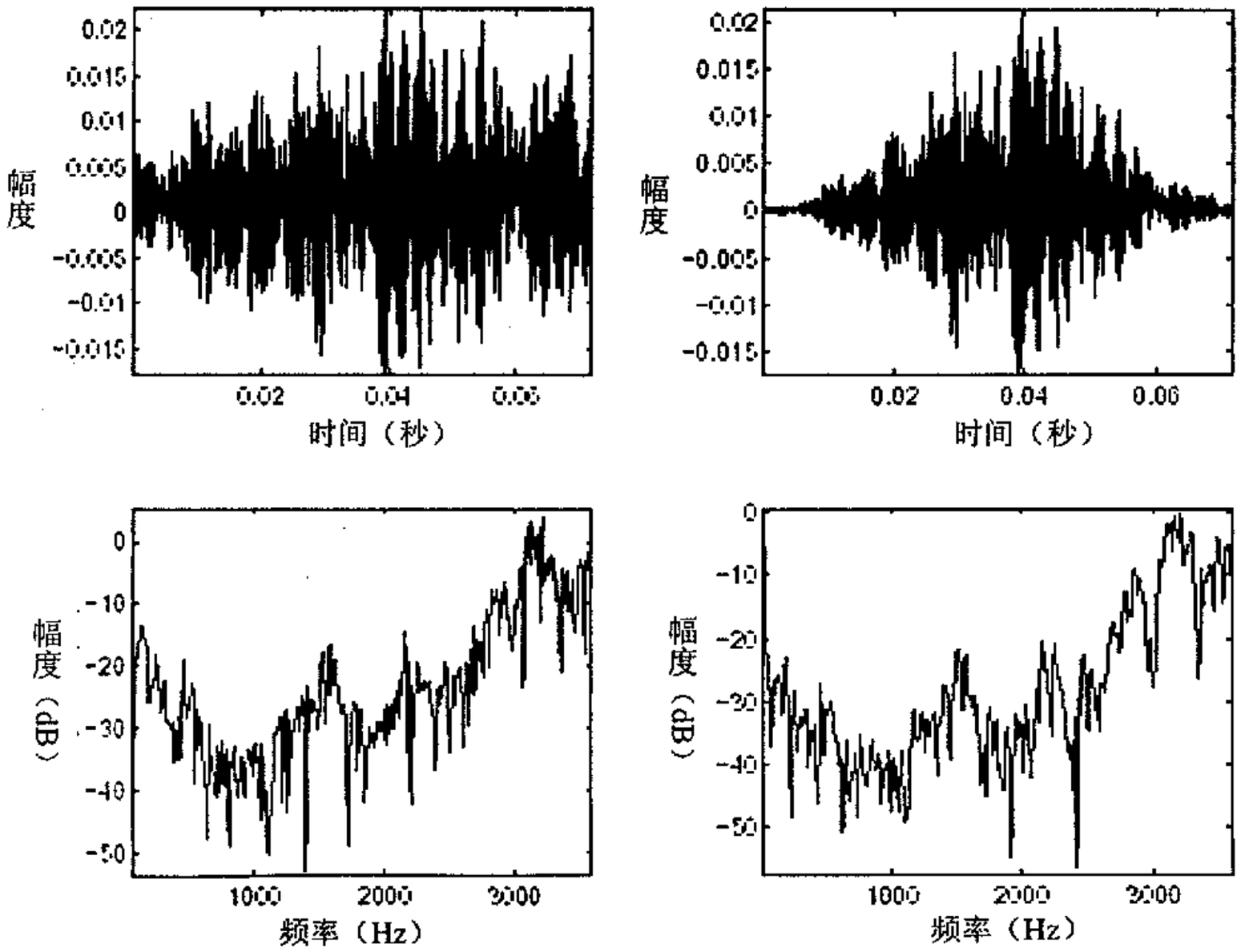


图3.9　清音信号分别加矩形和哈明窗的时域和频谱波形

3.3.2 短时帧长度与移动步长

基于短时帧的短时分析，特别是谱分析，对短时帧的长度有严格的要求，在具体的处理过程中必须具体选择。其中主要的因素是要考虑短时帧的时间分辨率与频率分辨率之间的协调。为了具有足够高的频率分辨率来捕捉频域的细微变化，短时帧的长度需要足够长。但是，为了能够获取时域的局部变化引起的频谱特征变化，需要较短的帧长。因此，频率分辨率与时间分辨率不能同时最大化，应该根据实际语音信号的特征来适当选取，一般为15~30ms。如果说话人是女性或儿童，由于其基音频率较高，所以短时帧长可选短些。而对于男性说话人类说，其基音频率较低，因此短时帧可选长些。显然，在说话人识别这样的实际应用中一般并不知道输入的语音是男声还是女声，因此只能选择一个合适的长度。

有一种称为基音同步分析（PSA: Pitch-Synchronous Analysis）的方法[38]，它随基音的变化而改变短时帧的长度。另一种帧长自适应方法为可变帧率（VFR: Variable Frame Rate）方法，这一方法基于短时平稳分析原理，通过相邻帧的距离计算判断是否属于同一分析段，并由此延长或终止短时帧的长度。目前一种更好的方法是基于小波变换的多分辨率分析[113]，通过对输入信号以不同的时频宽度进行分析，从而可以捕捉信号的全面特征。

短时分析时，帧移的长度一般是帧长的50~70%，这样可以得到较平滑的短时谱序列。

3.3.3 语音信号的预增强

所谓语音信号的预增强实际上就是一种提升高频频谱的处理。浊音语音信号频谱的高频区域会出现快速下降的情况，而这种下降也降低了频谱高频成分的作用。为了使信号频谱的整个频域特征能够均匀化地得到处理，提升高频成分就显得比较重要。所以，一般语音信号在预处理部分都需要有一个预增强的处理。

浊音语音的声门激励信号频谱有 -12dB/octave 的下降[114]，但当声音发出时又有一个 +6dB/octave 的频谱上升。因此，语音信号的录制输入中，信号频谱上出现了大约 –6dB/octave 的下降效应。所以，预增强的结果将使得语音信号的频谱中声门效应尽量消除，而尽可能地保持声道的频谱特性。对于清音，其信号频谱本身已经很平坦，所以没有必要进行预增强处理。

预增强对于线性预测也很重要。线性预测系数的稳定性与其对应的频域中频谱的动态范围成反比，动态变化越小，稳定性越高。因此，在求线性预测系数之前进行预增强可以尽量

避免系数的不稳定性。

常用的预增强滤波器系统函数如下：

$$H(Z)=1-\alpha Z^{-1} \tag{3.11}$$

其中，$\alpha>0$ 控制滤波器频率响应的斜率。可以看到，预增强滤波器是一个FIR，其输入与输出信号的时域关系以及频率响应如下：

$$\begin{aligned} y(n) &= s(n)-\alpha s(n-1) \\ H(e^{j\omega}) &= 1-\alpha e^{-j\omega} = 1-(\cos\omega - j\sin\omega) \\ |H(e^{j\omega})| &= (1-\alpha\cos\omega)^2+\alpha^2(\sin\omega)^2 \\ &= 1-2\alpha\cos\omega+\alpha^2 \end{aligned} \tag{3.12}$$

实际上，预增强在时域的处理就是将原信号变成一阶差分信号（$\alpha=1$）。图3.10显示了预增强滤波器幅度谱。

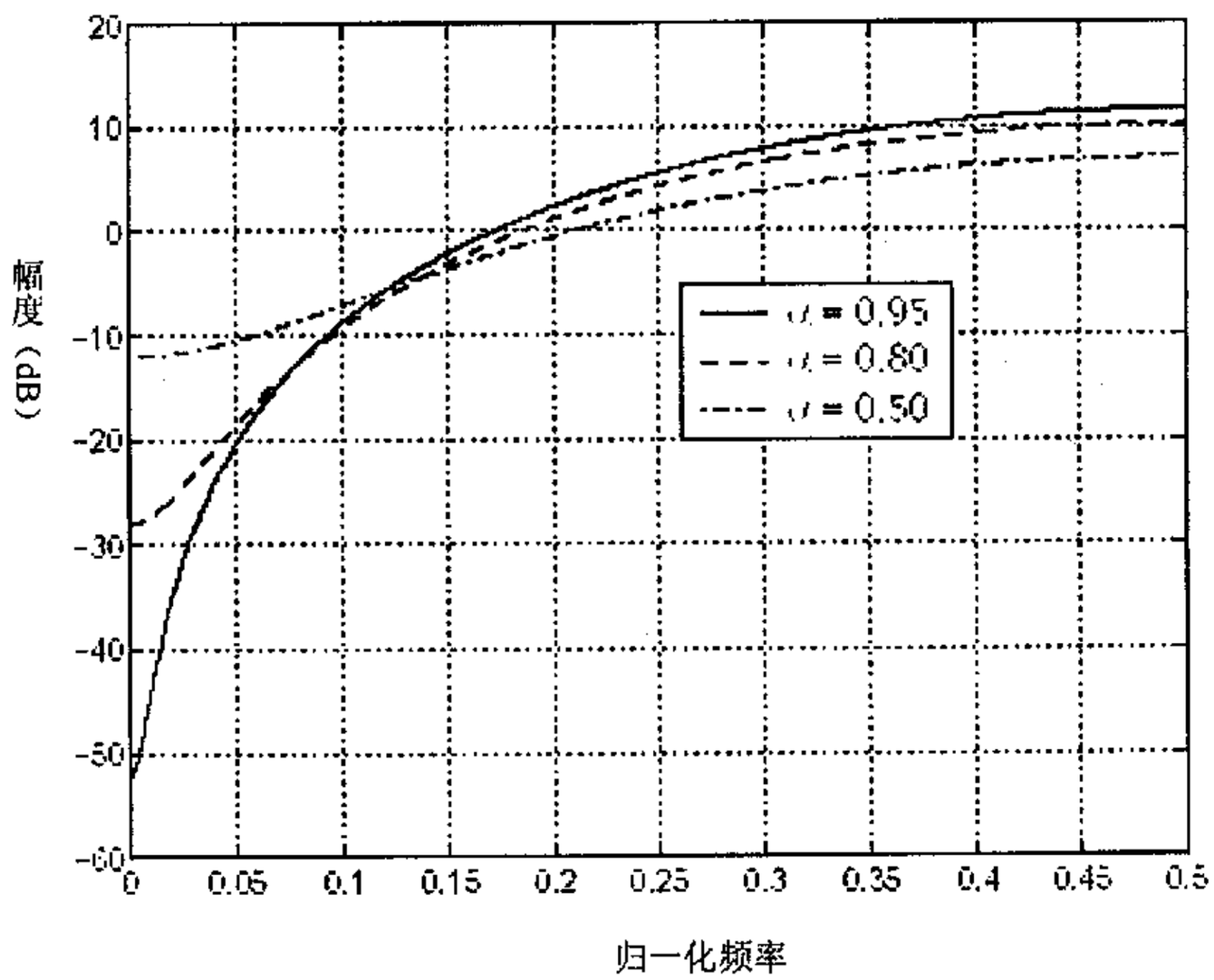


图3.10 预增强滤波器的幅度谱

另外，图3.11显示了一个语音信号预增强前后的时域以及频域效果，可以看到，预增强使得基音频率F0的高频谐波更加清晰，整个频域的频谱变得平坦。

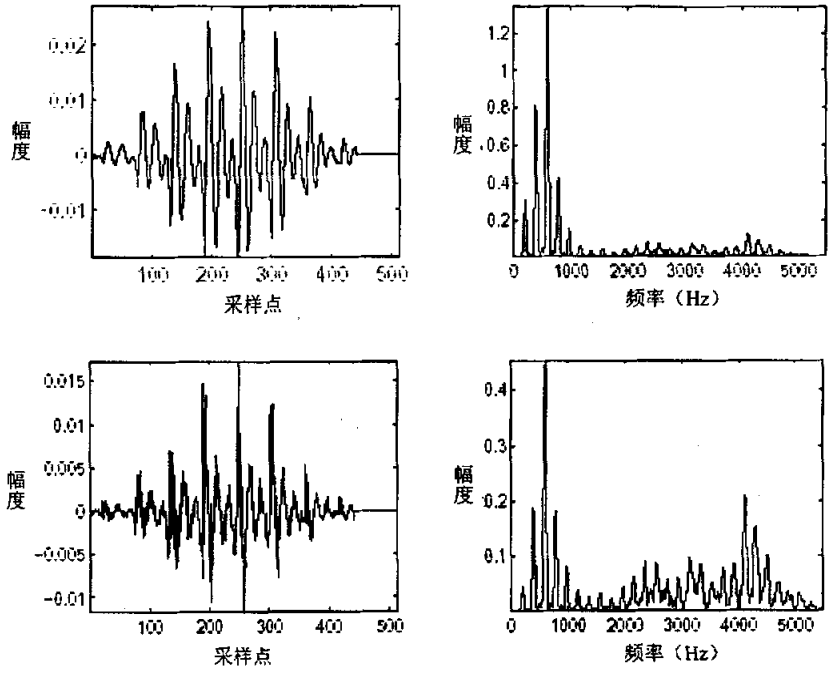


图3.11 预增强前后的时域与频域效果

3.4 小波变换分析

虽然短时傅立叶变换可以在一定程度上解决非平稳信号的频谱分析问题，但无法解决频率分辨率与时间分辨率的矛盾。小波变换[113]分析通过多尺度描述分析信号，能够全面地捕捉信号的时域与频域特征。

所谓小波，是由满足 $\int h(t)dt = 0$ 的母小波h(t)经过压缩或扩展形成：

$$h_{a,b}(t) = |a|^{-1/2} h(\frac{t-b}{a}) \quad a,b \in R,\ a \neq 0 \tag{3.13}$$

其中，a是压扩因子，b是位移因子。可以看出，小波函数随着a的增大波形会趋向平坦，而幅度降低，反之亦然，但总的信号能量不变，见图3.12。

小波变换是对信号 $s(t) \in L(R^2)$ 的小波调制，并求调制信号的积分，即小波能量，其定义如下：

$$WT_s(b,a) = |a|^{-1/2} \int_{-\infty}^{\infty} s(t)h(\frac{t-b}{a})dt \tag{3.14}$$

当尺度因子a取离散值，由若干比特位表示时，相应的小波变成离散小波，小波变换成为离散小波，如下：

$$DWT_s(b,2^j)=\frac{1}{2^{j/2}}\int_{-\infty}^{\infty}s(t)h(\frac{t-b}{2^j})dt \tag{3.15}$$

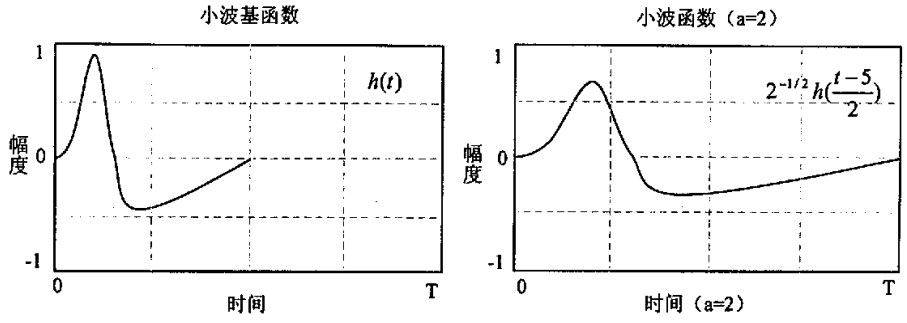


图3.12　小波函数是小波基函数的压扩

对小波变换公式的进一步分析可以发现，小波变换实际上是输入信号$s(t)$与具有压扩因子无移位的小波函数$h_a(t)$的线性卷积，如下：

$$\begin{aligned}&h_a(t)=|a|^{-1/2}\,h(\frac{t}{a})\\&WT_s(b,a)=s(t)*h_a(-t)|_{t=b}\end{aligned} \tag{3.16}$$

因此，如果将小波函数$h_a(t)$看作是滤波器的脉冲响应，则小波变换就是一个多尺度滤波器组的输出信号，其中各滤波器的中心频率和带宽由压扩因子a决定。小波变换输出与输入信号之间的频域关系如下：

$$\begin{aligned}&H_a(j\Omega)=|a|^{1/2}\,H(ja\Omega)\\&WT_s(j\Omega,a)=S(j\Omega)H_a^*(j\Omega)\\&|WT_s(j\Omega,a)|=|S(j\Omega)||H_a^*(j\Omega)|=|a|^{1/2}|S(j\Omega)||H(ja\Omega)|\end{aligned} \tag{3.17}$$

可以看到，对应小波函数$h_a(t)$的滤波器之幅度谱$|H_a(j\Omega)|$是小波基函数幅度谱的压扩，并且，压扩因子a增加，相应的中心频率和带宽就变小，如图3.13所示。

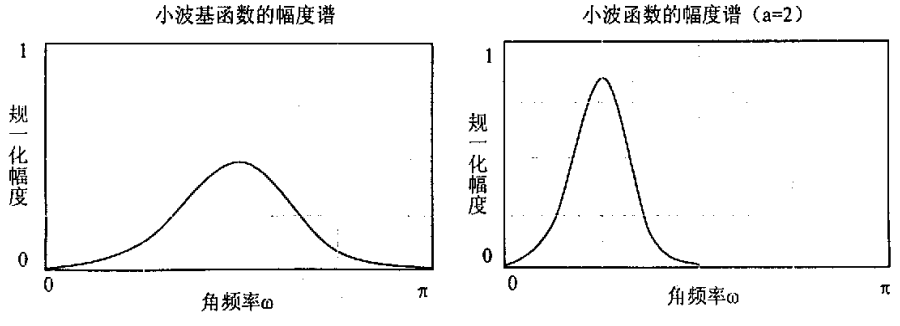


图3.13　小波函数频谱的压扩

显然，从滤波器组的角度来看，小波基函数应该具有频率局域化特征才能满足多分辨率分析的要求，使得通过压扩因子a的调节实现小波函数（滤波器）的时频分布对输入信号特定区域进行分析。一般地，对于信号中变化缓慢的低频成分可以使用较大的压扩因子实现较低时域分辨率和较高的频率分辨率分析信号。对于信号中快速变化的成分则采用较小的压扩因子实现较高的时域分辨率和较低的频率分辨率。小波变换分析的一般示意如图3.14所示。

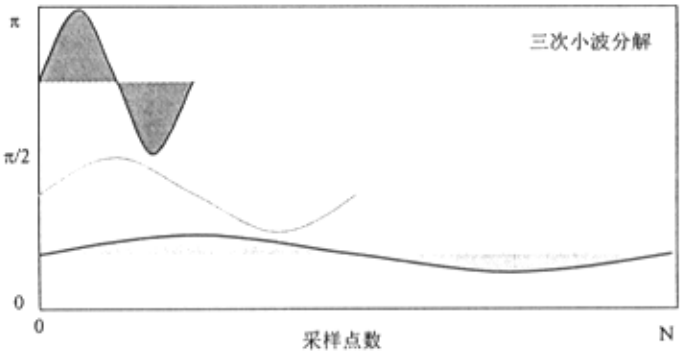


图3.14 基于小波变换的多分辨率分析

小波变换可以应用在语音特征提取方面，一方面通过小波变换多分辨率分析可以提取不同尺度下的特征，另一方面提供了一种更加可靠的特征提取方法。说话人识别中小波变换的应用不仅可以提高系统的鲁棒性，而且可以提高识别性能[72,115]。

3.5 互信息理论基础

互信息反映某一随机变量所带另一随机变量的信息。在通信中用来表示源信号在传输过程中的信息损失或接收端信号的失真程度。

设有二个随机变量X、Y，其概率密度分别为P（X）与P（Y），则X与Y之间的互信息定义如下：

$$\begin{aligned} &I(X;Y) = H(X) - H(X \mid Y) \\ &H(X) = -\int_X P(X)\log P(X)dX \\ &H(X \mid Y) = -\int_X \int_Y P(Y)P(X \mid Y)\log P(X \mid Y)dXdY \end{aligned} \tag{3.18}$$

这里，H（X）为随机变量X的熵，H（X|Y）为条件熵。I（X；Y）的含义为Y所携带的关于X的信息，称之为Y关于X的互信息。互信息具有以下特点：

（1） I(X；Y)= I(Y；X)。

（2） I(X;Y)>=0

（3） I(X; Y) >= I(X; Z)+I(Z; Y)

（4） 当 X 与 Y 完全相关时互信息最大。

（5） 当 X 与 Y 不相关时互信息最小。

以上随机变量可以是一维随机变量，也可能是多维随机变量，并具有一定的统计分布特征，该特征由相应的概率分布函数或密度函数表示。

互信息可以作为一种准则应用在 HMM 的参数估计中[62,64]，也可以作为模式之间的失真测度应用在语音识别中[83,116,117]。互信息计算中可以充分运用语音信号的统计分布特征，并且对统计分布的具体形式没有限制，因此越来越受到研究人员的重视。

第四章　语音信号互信息的计算

本章提要：

- 语音信号之间互信息的计算分析：随机干扰信号描述语音信号的失真，最大似然估计
- 互信息计算方法一：线性映射匹配算法 LPM
- 互信息计算方法二：非线性搜索匹配算法 NLM
- 互信息测度的聚类特性分析：类内凝聚度，类间耦合度，类间重叠
- 语音识别中互信息匹配的应用：数字识别，语音浏览器 VoiceIE，语音对话系统

语音信号包含多个层次的信息，主要的信息为语义和说话人个性特征信息。根据互信息理论，互信息表达随机变量之间相互携带对方的信息量或两者之间的失真度。

语音信号可以看作是随机信号源输出的非平稳随机信号，并具有短时平稳特征。如果将同一说话人的所有语音信号看作是同一随机信号源的输出，不同说话人所对应的语音信号看作是不同随机信号源的输出，那么，语音信号的随机特征表明，即便是属于同一随机信号源的输出语音之间也存在差异。对于同一随机信号源的输出语音信号来说，差异主要表现为语义特征的区别，说话人个性特征是一样的，但是，不同随机信号源输出的语音信号之间不仅存在语义特征的不同，而且还存在说话人个性特征的差异，见图4.1。

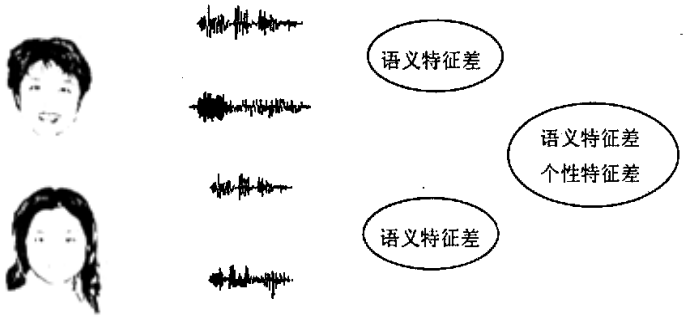


图4.1　语音信号之间的特征差

语音信号的互信息是两个语音信号之间相互携带对方信息量的定量描述。如果两个语音

信号的特征差异较小，一个语音信号的特征或信息就能够较多地从另一个语音信号中获取，即相互之间具有较大的互信息；反之，如果两个语音信号之间的特征差异较大，其互信息就小。设说话人 SPK_1~SPK_N 的对应语音信号集合为 SD_1~SD_N，每一个集合 SD_i 中包含若干与 SPK_i 对应的语音信号 $S_j^k \quad \forall k$。SD_i 中的语音信号对应同一说话人，是同一随机信号源的输出，其相互之间的特征差异相对较小，即相互之间有较大的相关性，因此，互信息较大。而对于 SD_i 中的任一语音信号 S_i^k 和 SD_j 中的任一语音信号 S_j^p，由于对应不同的说话人，属于不同的随机信号源的输出，所以相互之间的特征差异相对较大，相关性较小或没有，因此，互信息也较小。不同语音信号之间的互信息关系可以由下式表示：

$$\frac{1}{KK}\sum_{k=1}^{K}\sum_{q=1}^{K} I(S_i^k;S_i^q) > \frac{1}{KP}\sum_{k=1}^{K}\sum_{p=1}^{P} I(S_i^k;S_j^p) \qquad i \neq j \tag{4.1}$$

其中 I(.)表示互信息。上式给出的启示是，如果能够找到一种有效的说话人模型描述说话人的语音特征，并且能够找到可实际应用的语音信号之间或语音信号与说话人模型之间的互信息计算方法，那么，互信息理论就可以应用于说话人识别。但迄今为止，互信息理论在语音识别中主要应用在参数模型的训练、多频带的分配与组合等[62,63,64,65,66]，在说话人识别中的具体应用还没有成熟的研究成果。主要的原因是语音信号互信息的计算需要统计特征信息，而获得语音信号之间的条件概率分布特性是一个困难的问题。

这一章通过分析语音信号和特征矢量的统计分布特性，提出随机干扰信号描述语音信号之间的失真，并运用统计分析和估计理论解决语音信号互信息的计算问题。

4.1 语音信号互信息的计算分析

为了消除语音信号中的冗余信息，反映语音信号的非平稳时变分布特征，语音信号一般由随时间分布的特征矢量序列来表示，即语音模式。设语音信号 S_i 和 S_j 相应的特征矢量序列语音模式 $\boldsymbol{VT}$ 和 $\boldsymbol{VR}$ 分别如下：

$$\begin{aligned} S_i &\Rightarrow \boldsymbol{VT}:\{\boldsymbol{VT}_1, \boldsymbol{VT}_2, \quad \ldots \quad \boldsymbol{VT}_N\} \\ S_j &\Rightarrow \boldsymbol{VR}:\{\boldsymbol{VR}_1, \boldsymbol{VR}_2, \quad \ldots \quad \boldsymbol{VR}_M\} \end{aligned} \tag{4.2}$$

$\boldsymbol{VT}$ 和 $\boldsymbol{VR}$ 的特征矢量有相同的参数类型和个数，如果采用 LPC 参数，则序列中每一特征矢量由 p 个线性预测系数表示，形式如下：

$$(a_1, a_2, \ldots, a_p)^T \tag{4.3}$$

根据 L.Rabiner 等人的研究结果，LPC 以及 LPCC 倒谱等特征参数具有近似的正态统计分布特征[118]。因此，如果采用 LPC 等特征参数，则上述语音模式 $\boldsymbol{VT}$ 和 $\boldsymbol{VR}$ 可以分别表示为

具有概率密度函数 $N(\mathbf{m}_T,\mathbf{C}_T)$和 N（$\mathbf{m}_R,\mathbf{C}_R$）的 p 维特征矢量序列。

$$p(VT)=\frac{1}{(2\pi)^{p/2}|C_T|^{1/2}}\exp\{\frac{(VT-m_T)^T C_T^{-1}(VT-m_T)}{2}\} \tag{4.4}$$

$$p(VR)=\frac{1}{(2\pi)^{p/2}|C_R^i|^{1/2}}\exp\{\frac{(VR-m_R)^T (C_R^i)^{-1}(VR-m_R)}{2}\} \tag{4.5}$$

语音信号 S_i 和 S_j 之间的互信息也就是语音模式 **VT** 与 **VR** 之间的互信息，其计算公式如下：

$$\begin{aligned}
&I(S_i;S_j)=I(VT;VR)=H(VT)-H(VT|VR)\\
&I(S_j;S_i)=I(VR;VT)=H(VR)-H(VR|VT)\\
&I(S_i;S_j)=I(S_j;S_i)\\
&H(VT)=-\int_{VT} p(VT)\log p(VT)dVT\\
&H(VR)=-\int_{VR} p(VR)\log p(VR)dVR\\
&H(VT|VR)=-\int\int_{VTVR} p(VR)p(VT|VR)\log p(VT|VR)dVRdVT\\
&H(VR|VT)=-\int\int_{VRVT} p(VT)p(VR|VT)\log p(VR|VT)dVTdVR
\end{aligned} \tag{4.6}$$

从理论上分析，由于不可能得到严格的语音信号特征矢量的概率分布密度函数以及语音模式的条件概率，语音模式之间的互信息计算只能通过合理的估计进行。对于正态分布的特征矢量，语音模式 **VT** 和 **VR** 的熵如下式所示：

$$\begin{aligned}
H(VT)&=\frac{p}{2}log(2\pi e)+\frac{1}{2}log|C_T|\\
H(VR)&=\frac{p}{2}log(2\pi e)+\frac{1}{2}log|C_R|
\end{aligned} \tag{4.7}$$

语音信号中语义和个性特征的差异最终通过声学特性表现出来。声学特性的差异不仅表现在时变分布特征方面，而且表现在统计分布特征方面，并可以用一种随机干扰信号来描述，即信号 **VT** 和 **VR** 之间的差异由随机干扰信号 **XP** 引起，或者，**VT** 是 **VR** 和 **XP** 的函数，$VT=f(VR,XP)$。

根据概率统计理论，如果随机变量 X 和 Y 是相互独立的具有高斯统计分布特性的随机变量，则它们的和 Z 也是一个高斯分布随机变量。因此，从统计分布特征角度分析，如果使用 LPC 等具有正态分布的特征参数，则由于 **VT** 和 **VR** 都是具有高斯统计分布特征的随机特征矢量，这一随机干扰信号 **XP** 可以看作是一个与 **VR** 独立的具有相同高斯统计分布特征的叠加性噪声。当然，同一随机信号源输出的语音信号，其声学特性差异主要由语义差引起，不

同随机信号源输出的语音信号之间的声学特性差异不仅由语义差，而且由个性特征差引起。因此，同一随机信号源输出的语音信号所对应的随机干扰信号具有较小的均值漂移和方差，而不同随机信号源输出的语音信号所对应的随机干扰信号具有较大的均值漂移和方差。

通过以上分析可以得出，语音模式之间存在关系 **VT = VR+XP**，其中 **XP** 为独立的随机干扰信号模式。即便从实际应用角度看，由于随机干扰信号描述两个语音信号的特征差异，而语音信号本身是随机信号，因此，随机干扰信号与语音信号的相关性较小，也可以近似地将随机干扰信号看成是与语音信号独立的一个随机信号，与上面的理论分析一致。

通过以上分析，模式之间的互信息 I(**VT**；**VR**)计算如下：

$$\begin{aligned} &P(VT \mid VR) = P(VR + XP \mid VR) = P(XP) = P(VT - VR) \\ &H(XP) = \frac{p}{2} log(2\pi e) + \frac{1}{2} log \mid C_X \mid \\ &I(VT;VR) = H(VT) - H(VT \mid \hat{VR}) = H(VT) - H(XP) = \frac{1}{2} log \frac{\mid C_T \mid}{\mid C_X \mid} \end{aligned} \tag{4.8}$$

上式表示，语音信号之间的互信息与语音信号模式和随机干扰信号模式的自协方差矩阵值有关，其计算的关键是后者。在说话人识别的模式匹配过程中，输入语音信号模式对所有类别都一样，因此，模式匹配只需计算随机干扰模式的自协方差矩阵值$|C_x|$，并且，$|C_x|$的值越小，互信息越大。

以上分析主要针对语音信号的统计分布特征，而在实际的互信息计算中同样必须考虑语音信号的时变分布特征。以下 42 节和 4.3 节介绍互信息计算的具体实现方法。

4.2　互信息估计的线性映射匹配算法 LPM

互信息的计算最终归结到随机干扰信号和语音信号模式的自协方差矩阵值的计算。自协方差矩阵是一个对称正定矩阵，根据最大似然统计估值理论，自协方差矩阵可以由信号样本数据来估计得到。各信号模式对应的自协方差矩阵定义如下：

$$\begin{aligned} &m_T = E[VT]; \quad m_R = E[VR] \\ &m_X = E[VT - VR] = E[VT] - E[VR] = m_T - m_R \\ &C_T = E[(VT - m_T)(VT - m_T)^T] \\ &C_X = E[(VT - VR - m_X)(VT - VR - m_X)^T] \end{aligned} \tag{4.9}$$

期望值的计算需要随机变量的概率密度，一般都无法直接得到，但可以根据以下公式使用样本数据进行最大似然估值：

$$
\begin{aligned}
&m_T = \frac{1}{N}\sum_{i=1}^{N} VT_i ; \quad m_R = \frac{1}{M}\sum_{j=1}^{M} VR_j \\
&m_X = m_T - m_R \\
&C_T = \frac{1}{N}\sum_{i=1}^{N} (VT_i - m_T)(VT_i - m_T)^T \\
&C_X = \frac{1}{M}\sum_{j=1}^{M} (VT_{f(j)} - VR_j - m_T + m_R)(VT_{f(j)} - VR_j - m_T + m_R)^T
\end{aligned}
\tag{4.10}
$$

上式随机干扰信号模式 **XP** 的自协方差矩阵估计同时需要 **VT** 和 **VR** 两个语音信号模式的样本数据，而由于两个信号的时长不一致，特征矢量差的时序对准必须在进行时长归一化的基础上进行。通常的归一化方法是线性压扩或线性映射（Linear Projection），例如，以 **VR** 为基准的线性压扩映射公式如下：

$$
f(j) = \frac{N-1}{M-1}(j-1)+1 \tag{4.11}
$$

即对于 **VR** 的特征矢量序号 j，对应的 **VT** 特征矢量序号是 f(j)，映射关系如图 4.2 所示。

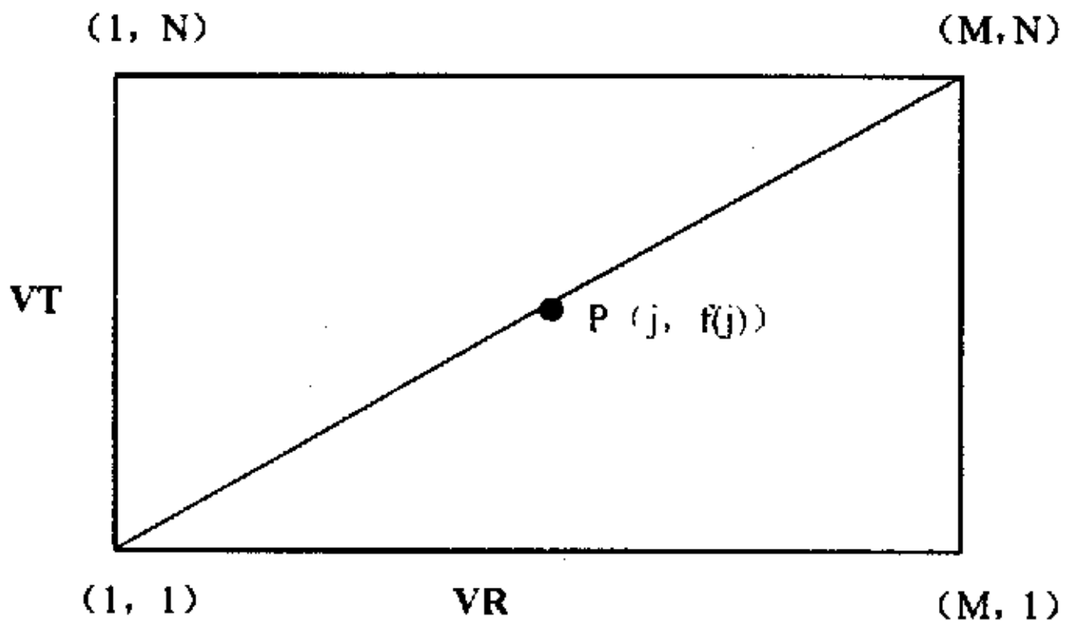


图 4.2　互信息计算的线性映射匹配估计方法

4.3　互信息估计的非线性搜索匹配算法 NLM

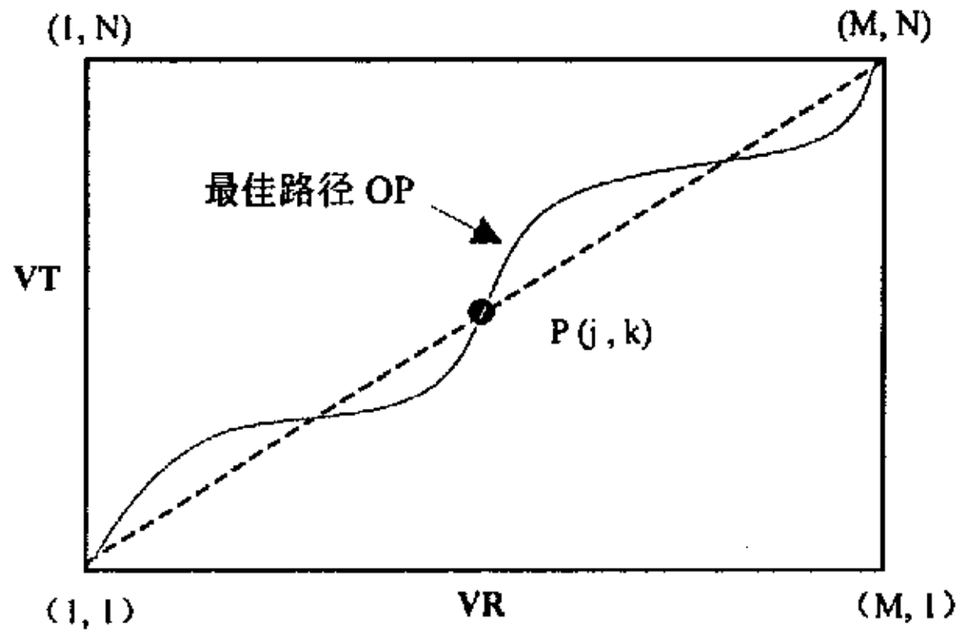
同一说话人所对应的不同语音信号之间，不仅存在由线性波动引起的时变分布特征差异而且存在非线性波动引起的时变分布特征差异，在语音模式匹配中应该充分考虑对线性和非线性波动造成的时变分布特征差异进行处理，消除其对匹配精度的影响。因此，互信息的计算应该允许沿 **VT-VR** 模式空间中的非线性路径进行，而不是简单地沿线性路径进行计算，如图 4.3 所示。

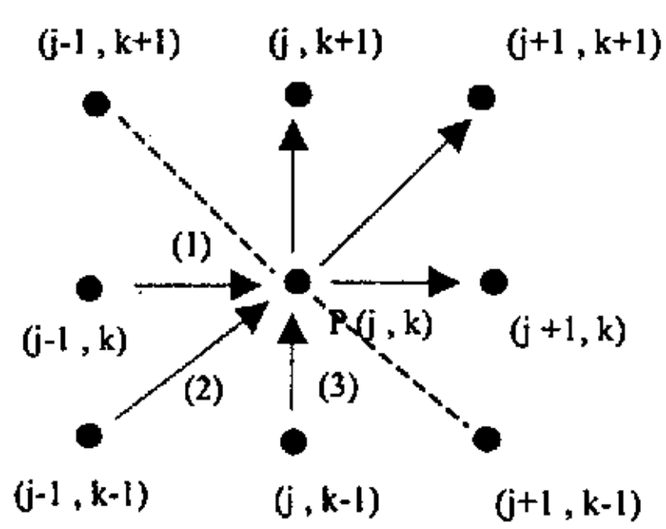
考虑以上因素，公式（4.10）中关于自协方差矩阵值$|C_x|$的计算应该采用非线性方式进行

匹配，具体计算公式如下：

$$C_X = \frac{1}{L}\sum_{l=1}^{L}(VT_{k(l)} - VR_{j(l)} - m_T + m_R)(VT_{k(l)} - VR_{j(l)} - m_T + m_R)^T \tag{4.12}$$

上式中 k(l)和 j(l)分别为 **VT-VR** 模式空间中模式匹配路径与测试模式和参考模式坐标之间的非线性映射函数，L 为匹配路径上包含的匹配点数目。如果在所有可能的路径中，沿路径 OP 的计算得到最大互信息，则称 OP 为最佳路径，并将相应的互信息值作为模式匹配的结果 **I(VT；VR)**。





P 点局部匹配选择

图 4.3　互信息估计的非线性搜索匹配算法

最佳路径 OP 可以采用非线性搜索方法得到。可以认为，语音信号的线性与非线性波动随时间呈现单调性变化特征，即图 4.3 中最佳路径上的点 P 仅仅可能与三个前向点之一派生连接，其选择前向连接点的规则是使 P 点的累积互信息最大或累积自协方差矩阵值$|C_X(j,k)|$最小。具体的非线性搜索匹配算法如下：

（1）初始化：$P(1,k)=Z$ ，$P(j,1)=Z$ ，$C_X(1,k)=Z$ ，$C_X(j,1)=Z$ ，$L_X(1,k)=0$ ，$L_X(j,1)=0$

$k=1\sim N$，　$j=1\sim M$，其中 Z 为零矩阵。　　(4.13)

（2）计算：$P(j,k) = (VT_k - VR_j - m_T + m_R)(VT_k - VR_j - m_T + m_R)^T$　　(4.14)

（3）计算：
$$\begin{aligned} C_1(j,k) &= (C_X(j-1,k)L_X(j-1,k) + P(j,k))/(L_X(j-1,k)+1) \\ C_2(j,k) &= (C_X(j-1,k-1)L_X(j-1,k-1) + P(j,k))/(L_X(j-1,k-1)+1) \\ C_3(j,k) &= (C_X(j,k-1)L_X(j,k-1) + P(j,k))/(L_X(j,k-1)+1) \end{aligned} \tag{4.15}$$

（4）$C_X(j,k) = C_{min}(j,k)$ ；　　if $|C_{min}(j,k)| = Min(|C_1(j,k)|,|C_2(j,k)|,|C_3(j,k)|)$　　(4.16)

$$(5)\quad L_X(j,k)=\begin{cases}L_X(j-1,k)+1; & if \quad |C_X(j,k)|=|C_1(j,k)|\\ L_X(j-1,k-1)+1; & if \quad |C_X(j,k)|=|C_2(j,k)|\\ L_X(j,k-1)+1; & if \quad |C_X(j,k)|=|C_3(j,k)|\end{cases} \tag{4.17}$$

通过对步骤（2）~（5）进行 j=1~M，k=1~N 的迭代计算，可以得到自协方差矩阵沿最佳路径的估计值$|\mathbf{C}_x|$，即

$$|C_X| = |C_X(M,N)| \tag{4.18}$$

最佳路径匹配点数目 L 可以在迭代计算的同时统计并记录，其值为 $L_X(M,N)$。

从以上两小节互信息估计和非线性搜索算法的计算公式可以看出，VT-VR 描述模式之间的时变分布特征差异，互信息随着这差异的增加而减少，反之亦然。$\mathbf{m_T}$-$\mathbf{m_R}$ 和 $\mathbf{C_X}$ 描述模式之间的统计分布特征差异，特别是 $\mathbf{C_X}$，它的 Fourier 变换反映随机干扰信号 XP 的平均功率，并随 $\mathbf{C_X}$ 的增大而增大，因此，$\mathbf{C_X}$ 的值越大，表示干扰信号越强，即模式之间的统计分布特征差异越大，相应的互信息就越小。另外，非线性搜索算法使互信息的计算能够沿模式空间的非线性路径进行，这样，同类模式匹配时时变分布特征的线性和非线性波动都可以得到有效的处理，从而减少其对互信息匹配精度的影响。

综上所述，互信息估计的非线性搜索算法不仅能够有效地揭示语音信号之间的时变分布特征和统计分布特征差异，并且能够很好地处理时变分布特征的线性和非线性波动。

4.4　互信息测度的聚类特性分析

模式识别中都需要运用距离测度来衡量模式之间的差异或相似程度，因此其选择非常重要。在说话人识别中，一般采用 Euclidean 和 Itakura-Saito 等测度来衡量[119,120,121]。这些测度都是基于短时帧进行计算，并通过线性或非线性的方式进行累加得到模式之间的距离。并且，这些传统的测度一般被用来计算语音短时功率谱的差，无论是 DTW[19]这样基于模板的识别方法，还是 HMM[21]这样基于统计模型的识别方法，都运用了这些短时功率谱测度。另外，Mahalanobis 测度也是一种可选择的模式距离测度[121]。但作为一种应用于说话人识别的距离测度，它们本身都没有考虑语音信号的统计特征。Itakura-Saito 测度是由 LPC 预测残差推导得出，在采用 LPC 参数作为模式特征参数时具有较好的性能，但对于其他参数并不合适。以下是这些测度的计算公式。

(1) Euclidean　　$d(x,y)=(x-y)^T(x-y)$ (4.19)

(2) Mahalanobis　　$d(x,y)=(x-y)^T C_y^{-1}(x-y)$ (4.20)

(3) Itakura-Saito　　$d(x,y)=y^T R_x y / x^T R_x x-1$　　(4.21)

其中，x，y 分别是测试模式和参考模式的短时帧特征矢量。C_y 是参考模式特征矢量的自协方差矩阵，R_x 是测试模式特征矢量的自相关矩阵。

互信息反映模式之间的相似度，可以看作是另一种用于计算模式之间差异的测度。从原理来看，互信息测度依据互信息理论而提出[83]，是一种基于语音模式的距离测度，与以上传统的基于短时帧的测度有很大的区别。此外，互信息测度中考虑了语音信号的统计分布特征，在运用非线性搜索算法的情况下不仅能很好地处理模式间的时变分布差异，而且能够很好的处理模式之间的统计分布差异。在以下关于聚类特性的分析中，考虑到与传统距离测度的比较，采用与互信息相对应的随机干扰信号自协方差矩阵值的方法，该自协方差矩阵的计算公式如 4.2 和 4.3 节所示。

聚类特性分析的目的是依据具体的语音模式样本数据，对模式类内凝聚度和类间耦合度的聚类特性指标进行统计计算分析，并在此基础上对互信息测度与传统距离测度的特性进行比较。本文选择汉语中具有典型声学语音特征的数字集{0，1，...，9}作为基本数据，每一数字对应一个模式类别，这样共有 10 个模式类别 $\omega_0,\sim\omega_9$。每一个数字类别输入 12 个语音模式样本用于分析，共输入 120 个模式样本，形成模式样本数据库 $\{x_i^k\in\omega_k \quad i=1\sim12;\ k=0\sim9\}$。

在统计计算分析中，Euclidean(EU)，Mahalanobis(MA)，Itakura-Saito(I-S) 和互信息(MI)四个测度都基于相同的实验环境和数据，采用相同的 LPC 特征参数以及非线性搜索匹配机制，但特征参数的个数不同。考虑到 Itakura-Saito 仅适合于 LPC 参数，因此选择 8 阶 LPC 全极点自回归 AR 模型系数作为模式特征参数，并由 Durbin 算法计算。考虑到 MI 和 MA 测度计算中需要使矩阵求逆以及相应行列式求值与各种测度计算时间相对一致[116]，以上四种测度计算时分别采用以上模型系数的前 8、6、8 和 4 个 LPC 参数。

为了评价某一测度的聚类特性，亦即模式类内凝聚度以及类间耦合度，需要计算模式之间的距离。前面提到，除了互信息测度，其它测度都是基于短时帧进行计算的，但无论哪一种测度，模式匹配计算都是采用 4.3 节介绍的非线性搜索算法进行[122]。

4.4.1 类内凝聚度分析

类内凝聚度反映同类模式之间的紧凑程度，在本文分析中采用平均类内距离来衡量。平均类内距离越小，凝聚度越高，反之亦然。一些基本的分析参数定义如下：

(1) $d_{ij}^k = d(x_i^k; x_j^k)$ ：计算类内模式样本距离。

(2) $D_i^k = \frac{1}{N-1}\sum_{j=1}^{N} d_{ij}^k$，$j \neq i$ ：计算从模式样本 x_i^k 到类内其它模式样本的平均距离。

(3) $COV_i^k = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(d_{ij}^k - D_i^k)^2}$，$j \neq i$ ：计算从模式样本 x_i^k 到类内其它模式样本距离的均方差。

(4) $MAX_i^k = \underset{j}{MAX}\ d_{ij}^k$，$j \neq i$ ：计算从模式样本 x_i^k 到类内其它模式样本的最大距离。

(5) $MIN_i^k = \underset{j}{MIN}\ d_{ij}^k$，$j \neq i$ ：计算从模式样本 x_i^k 到类内其它模式样本的最小距离。

(6) $AV = \frac{1}{N}\sum_{i=1}^{N} D_i^k$：计算类内所有模式样本的平均距离。

(7) $COV = \frac{1}{N}\sum_{i=1}^{N} COV_i^k$：计算类内所有模式样本距离的均方差。

(8) $MAX = \frac{1}{N}\sum_{i=1}^{N} MAX_i^k$：计算类内平均最大模式样本距离。

(9) $MIN = \frac{1}{N}\sum_{i=1}^{N} MIN_i^k$：计算类内平均最小模式样本距离。

以上各参数计算式中，N表示类内模式样本个数，实际计算分析中为12；k是模式类的标号，分布在0–9；$d(\cdot)$表示距离测度函数，随不同的测度而改变。基于以上基本分析参数，定义以下两个归一化平均距离参数作为评价指标对类内模式凝聚度进行评价。

(1) NOR_AV0=AV/MAX

(2) NOR_AV1=AV/(AV+COV)

各种测度随模式类别（0~9）的指标值以及所有类别的平均（AV）指标值分析结果如图4.4和图4.5所示。总的说来，两种评价指标随各类分布变化的情况大体相似。由图可见，互信息测度的总体平均距离最小，并且在七个模式类处具有最小平均距离，因此可以认为具有最大的类内凝聚度。另外还看到，MI测度的平稳性较好，平均距离值随模式类的变化较小。图4.4，4.5显示EU和I-S测度的总体平均距离比MI的大一倍，并且所有模式类的平均距离都大于MI的平均距离，亦即各类的凝聚度较差。

类内凝聚度指标具有绝对性，能够反映同类模式之间的紧凑程度，但不能够反映不同类模式之间的分离程度。因此，仅仅依据类内凝聚度来判断或预测不同距离测度的实际识别性能并不可靠。

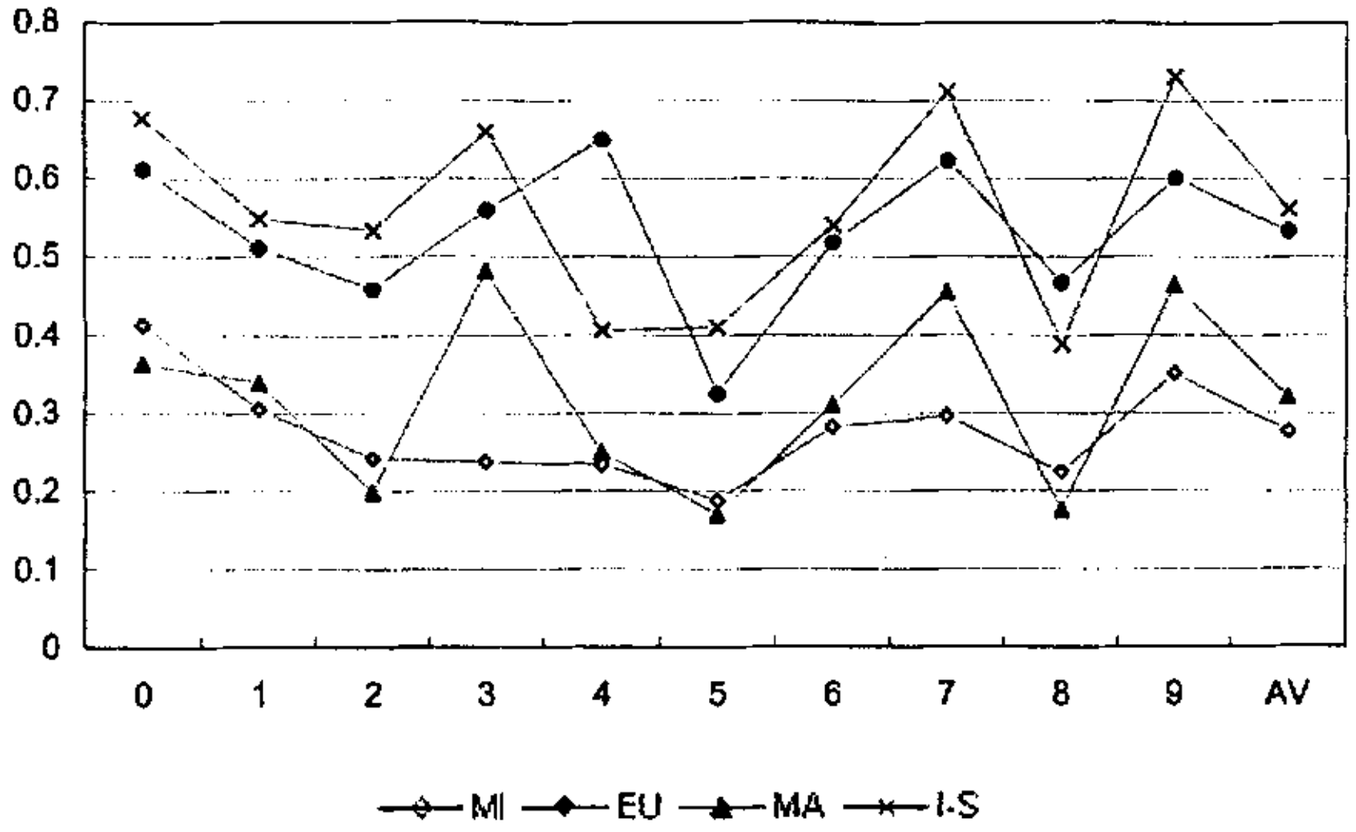


图4.4　模式类内凝聚度指标NOR_AV0

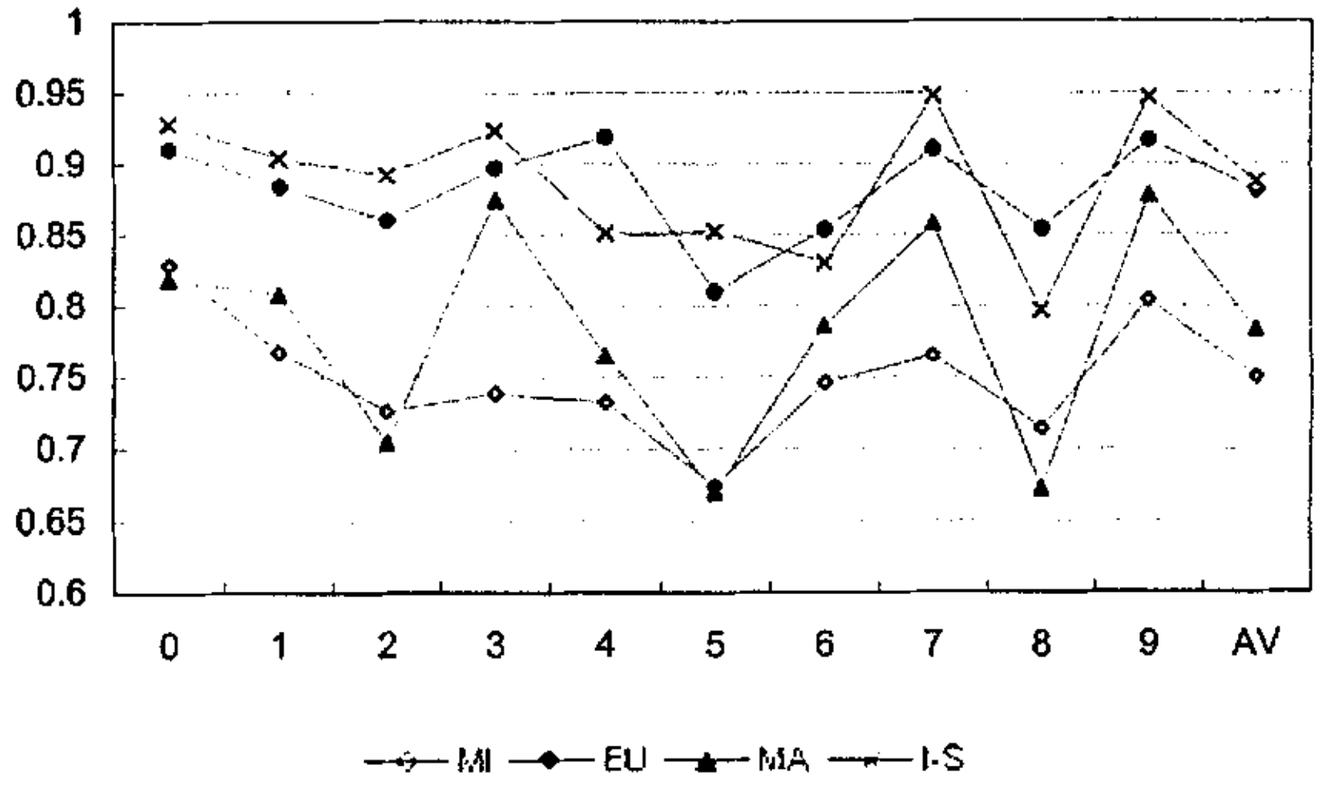


图 4.5　模式类内凝聚度指标 NOR_AV1

4.4.2 类间耦合度分析

类间耦合度反映不同模式类别间的分离程度，在本文分析中运用平均类间距离来衡量，平均类间距离越大，类间耦合度越小。一些基本的分析参数定义如下：

(1) $dt_{ij}^{k} = d(x_i^k; y_j),\ \ y_j \notin \omega_k$ ：计算两个不同类别模式样本之间的距离。

(2) $DT_i^k = \frac{1}{M}\sum_{j=1}^{M} dt_{ij}^{k}$ ：计算从模式样本 x_i^k 到其它类别各模式样本间的平均距离。

(3) $COVT_i^k = \sqrt{\frac{1}{M}\sum_{j=1}^{M}(dt_{ij}^{k} - DT_i^k)^2}$ ：计算样本 x_i^k 到其它类别各样本间距离之均方差。

(4) $MAXT_i^k = \underset{j}{MAX}\ \ dt_{ij}^{k}$ ：计算模式样本 x_i^k 到其它类别各样本间的最大距离。

(5) $MINT_i^k = \underset{j}{MIN}\ \ dt_{ij}^k$ ：计算模式样本 x_i^k 到其它类别各样本间的最小距离。

(6) $AVI=\frac{1}{N}\sum_{i=1}^{N} DT_i^k$ ：计算类别 k 到其它类别的平均距离。

(7) $COVI=\frac{1}{N}\sum_{i=1}^{N} COVT_i^k$ ：计算类别 k 到其它类别的距离之均方差。

(8) $MAXI=\frac{1}{N}\sum_{i=1}^{N} MAXT_i^k$ ：计算类别 k 到其它类别的平均最大距离。

(9) $MINI=\frac{1}{N}\sum_{i=1}^{N} MINT_i^k$ ：计算类别 k 到其它类别的平均最小距离。

以上各参数计算式中的 N、k 和 $d(\cdot)$的意义同类内凝聚度参数计算式中的意义一致；y_j表示非同类模式样本；M 表示非同类模式样本数，实际计算中为 108。在计算以上基本参数之后，定义如下两个归一化评价指标用于模式类间耦合度的评价。

(1) NOR_INTER0=(AVI-AV)/AVI

(2) NOR_INTER1=(AVI-AV)/(AVI+COVI)

各评价指标随模式类别变化的情况如图 4.6 和图 4.7 所示，总体而言，两个指标的分布情况基本是一致的，这与模式类内凝聚度指标的情况相同。从图中可以看到，MI 测度的总体平均类间距离最大，因此其类间耦合度最小。并且，MI 测度的类间平均距离表现出相当的平稳性，这表明 MI 测度的聚类特性相当稳定，语音信号声学语音特征变化的影响很小，这主要是因为在 MI 测度中考虑了语音信号的统计特征。I-S 测度的总体平均类间距离虽然与 MI 相近，但在类别 6 和 8 处表现出明显的下降，耦合度增加。EU 测度的总体类间平均距离最小，类间耦合度最高。

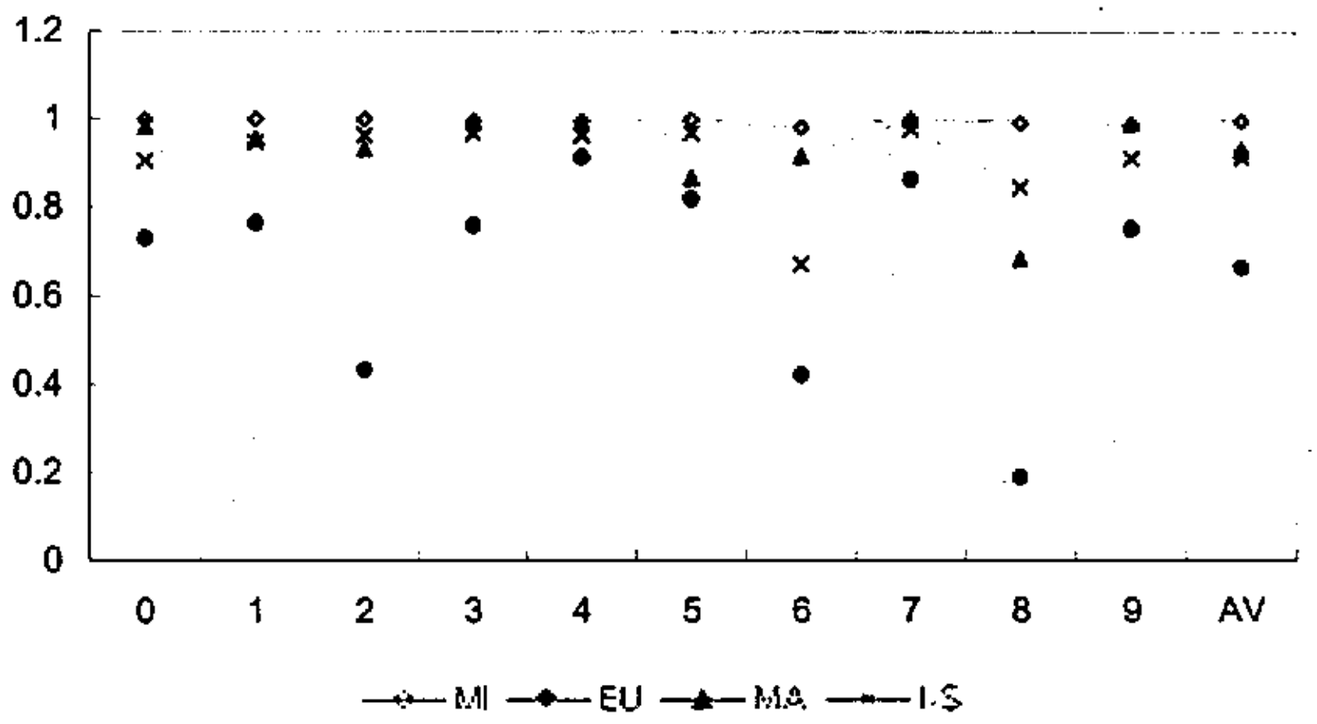


图 4.6　模式类间耦合度指标 NOR_INTER0

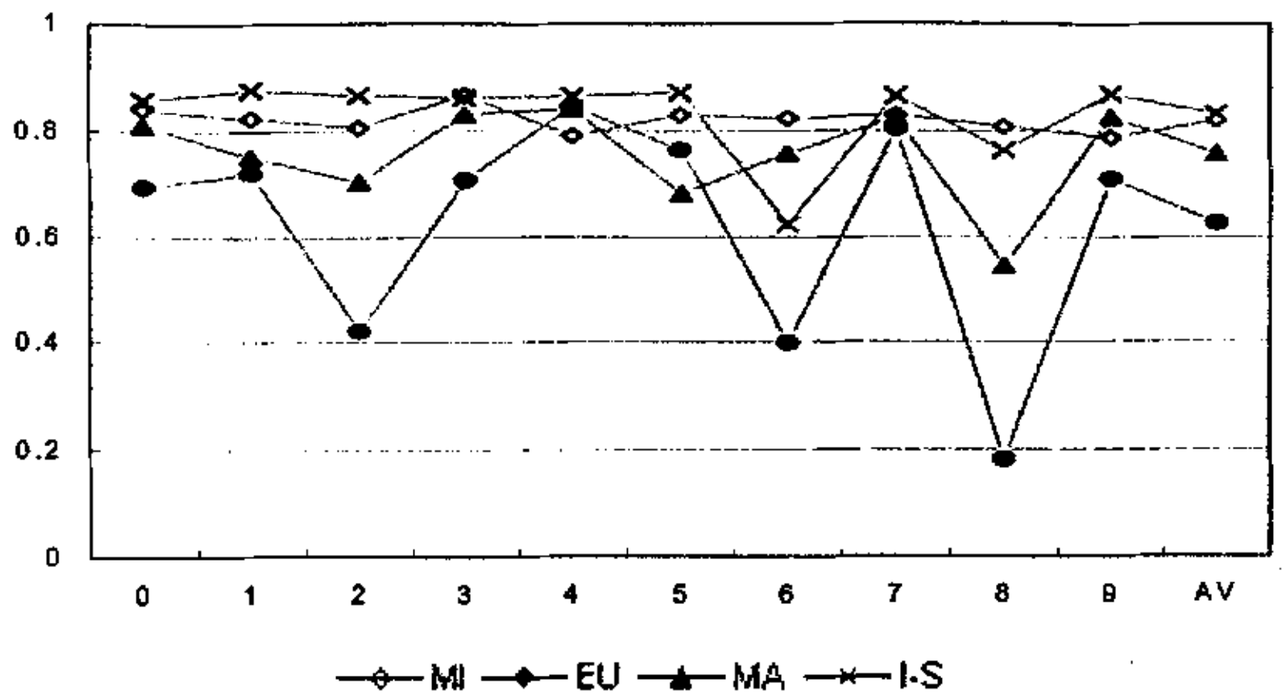


图 4.7　模式类间耦合度指标 NOR_INTER1

耦合度评价指标是一个相对指标，在其计算中同时运用了类内和类间平均距离参数。因此，根据耦合度来分析预测相应的识别性能相对可靠，但由于评价指标仅仅依据一维空间距离值得出，而实际语音识别中模式空间是多维的，所以仍然有非确定性。

4.4.3 类内类间平均距离比分析

类内凝聚度和类间耦合度指标从单方面反映了失真测度的分类性能，另一种常用的指标是类内类间平均距离比，即

$$r = \frac{\text{类内平均距离}}{\text{类间平均距离}}$$

类内类间平均距离比指标 r 是对凝聚度和耦合度的一个综合评价，指标值越小，说明模式分类性能越好，反之越差。实验分析中的具体计算公式如下：

$$NOR\_RATE = \frac{AV}{AVI}$$

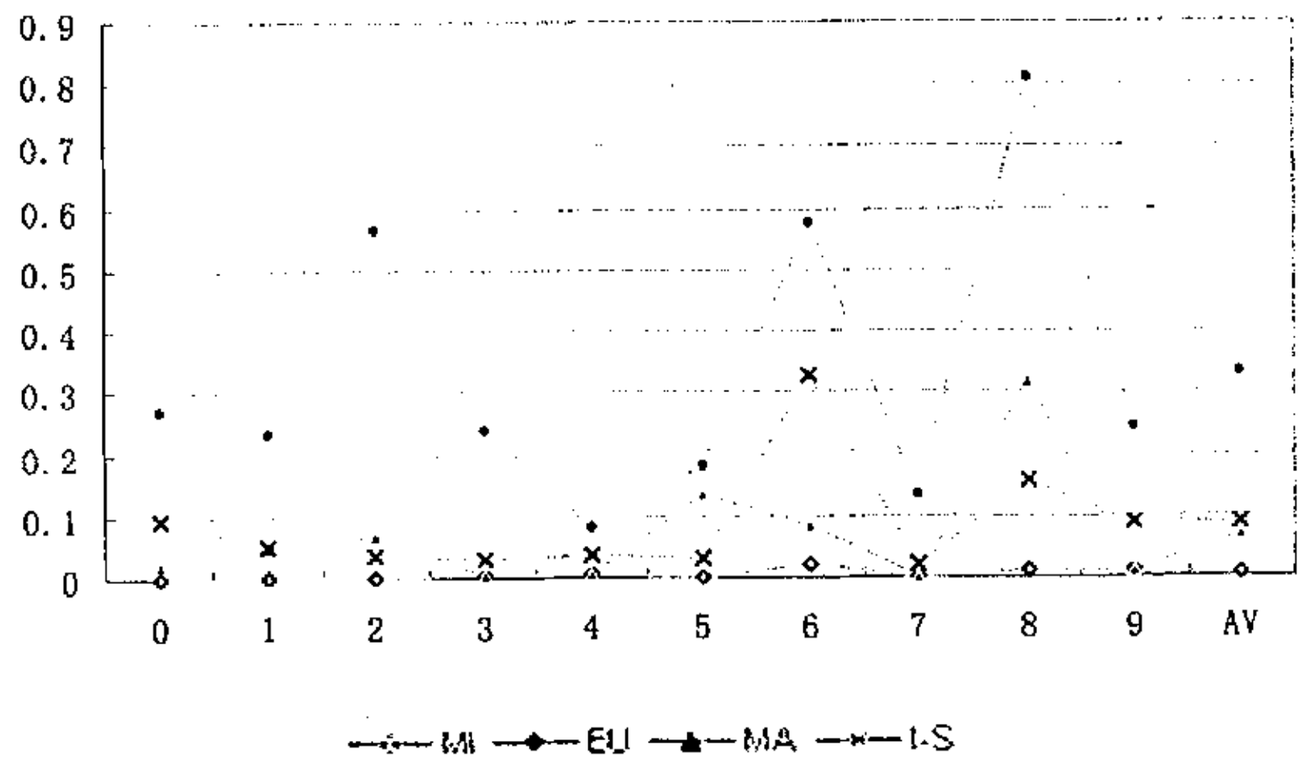


图 4.8　类内类间平均距离比 NOR_RATE

实验结果如图 4.8 所示。显然，MI 测度不仅具有最小的平均 NOR_RATE 指标值，而且几乎在所有类别上具有最小的指标值，说明 MI 测度的分类性能最好。EU 测度在所有类别上的指标值都是最大的，因此，分类性能最差。这些结果实际上与前面的实验分析结果是一致的。

4.4.4 类间重叠分析

类间重叠特性反映模式类在距离空间的交叉情况。交叉越多，重叠越多。它与耦合特性一样，反映模式类间的可分离程度。在本文分析中，根据以上类内与类间距离参数，定义四个归一化重叠特性评价指标来评价类间不同的重叠程度。虽然不能简单地根据一维空间的指标值去预测识别性能，但可以作为一种统计意义上的参考。四个重叠特性评价指标如下：

(1) COP1=(MINI-MAX)/AVI　　如果小于零，有极少的重叠(可能有少量误差)。

(2) COP2=(MINI-AV-COV)/AVI　　如果小于零，有少量的重叠(可能有误差)。

(3) COP3=(AVI-COVI-MAX)/AVI　　如果小于零，有一定的重叠(容易产生误差)。

(4) COP4=(MINI-AV)/AVI　　如果小于零，有重叠(有误差)。

以上指标是根据一维空间数据得出的，反映一维空间的重叠情况，如果用来预测实际识别情况的话则如括号中所示。各指标值随类别变化的曲线如图 4.9，4.10，4.11，4.12 所示。从图中可以看到，MI 和 I-S，EU 与 MA 具有相似的总体平均重叠指标值，前者的重叠情况较后者要轻。虽然 MI 和 I-S 在类 6 和类 8 处都有重叠，但 I-S 比 MI 要严重得多，并且 MI 同样表现出相当的稳定性。虽然 EU 和 MA 在总体上基本一致，但从图中可以看到，MA 在更多的类上具有负指标值，因此，可以认为 EU 的重叠情况较 MA 轻。

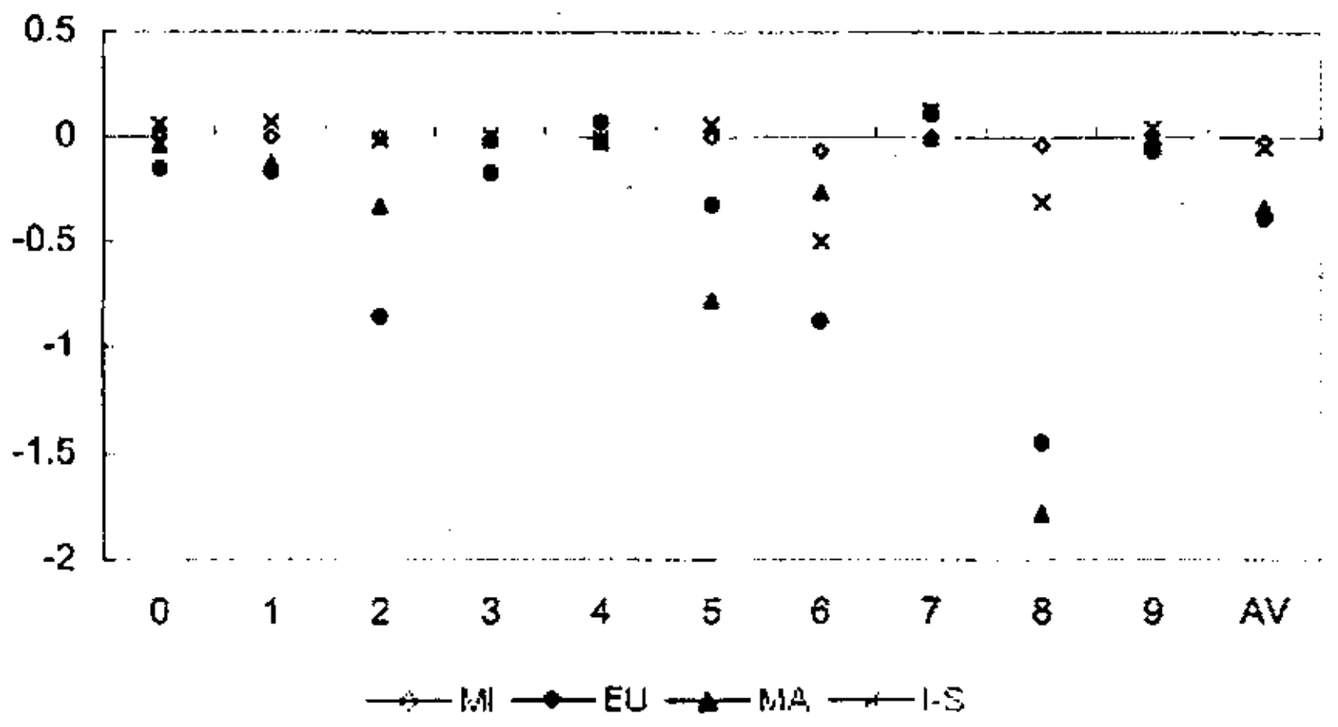


图 4.9　模式类间重叠指标 COP1

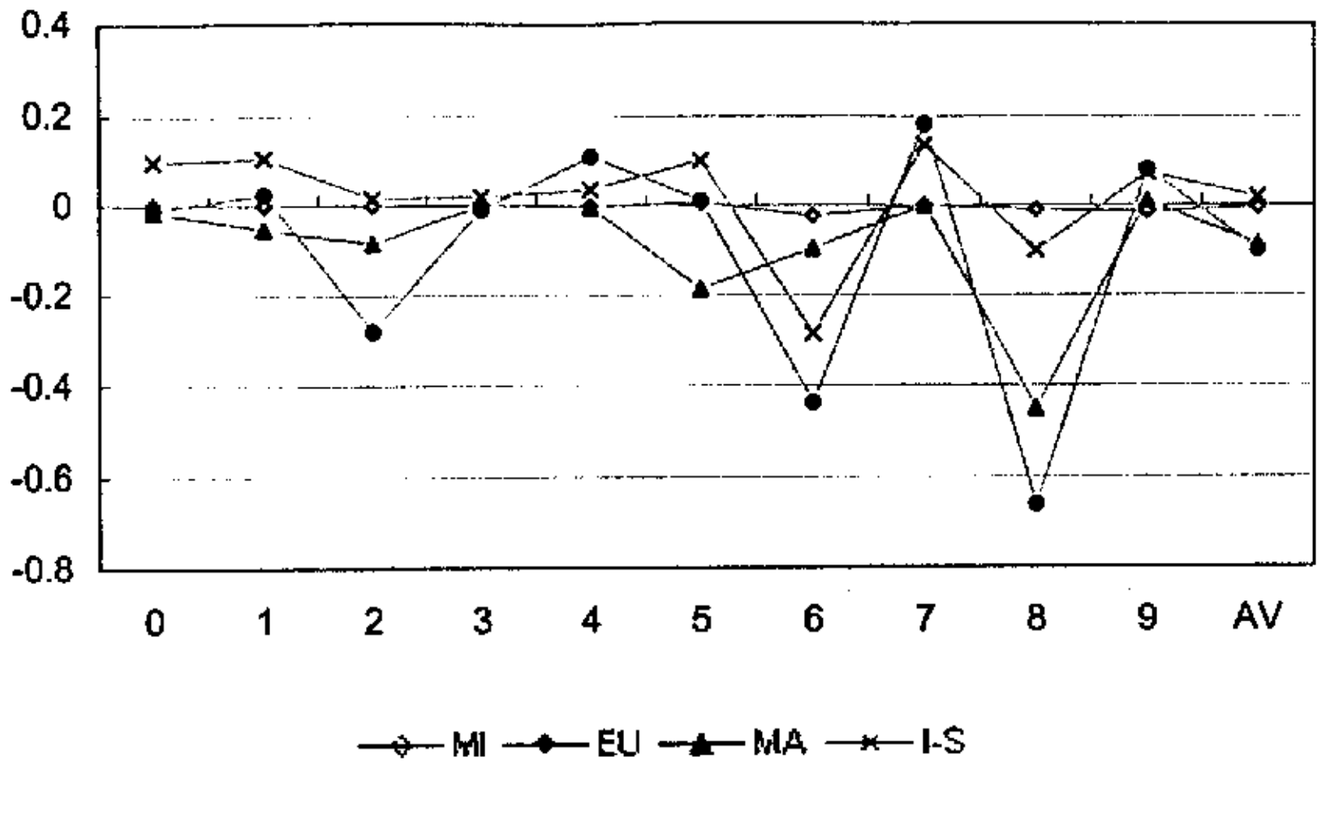


图 4.10　模式类间重叠指标 COP2

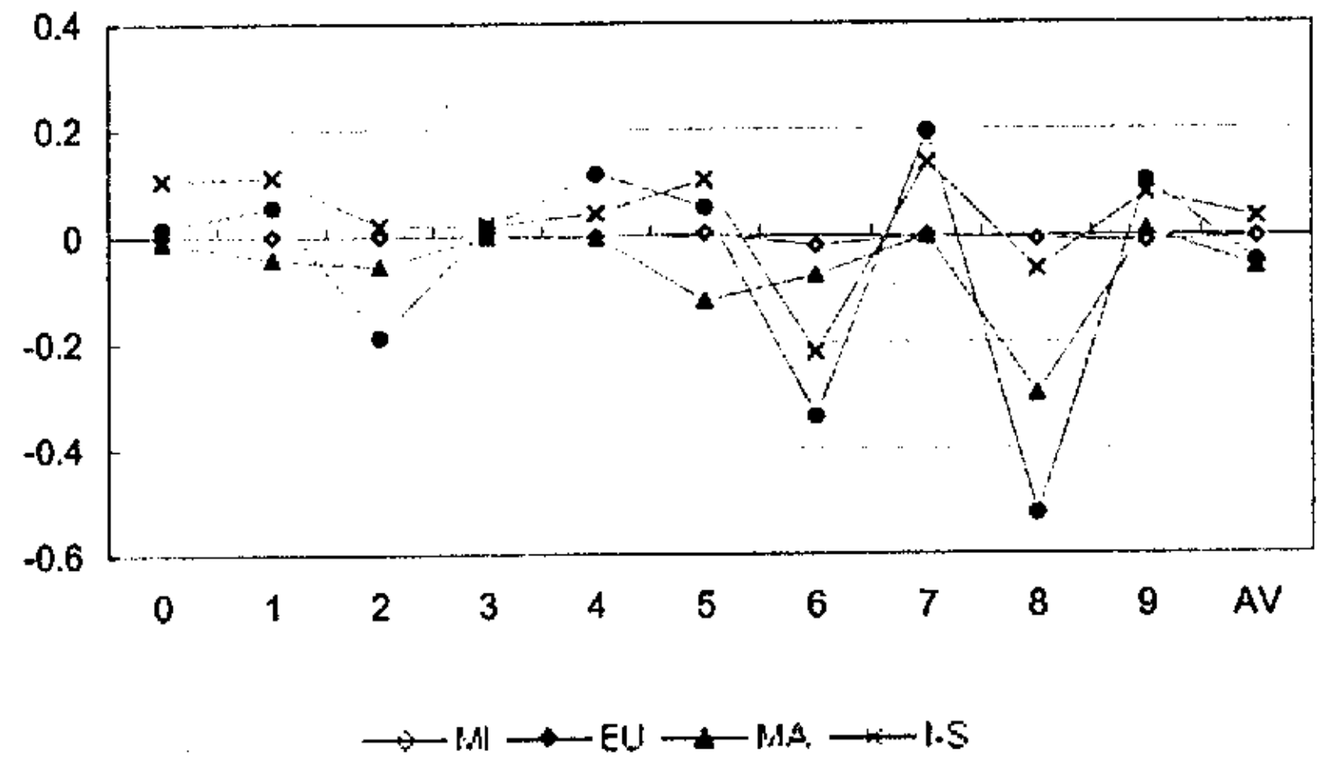


图 4.11　模式类间重叠指标 COP3

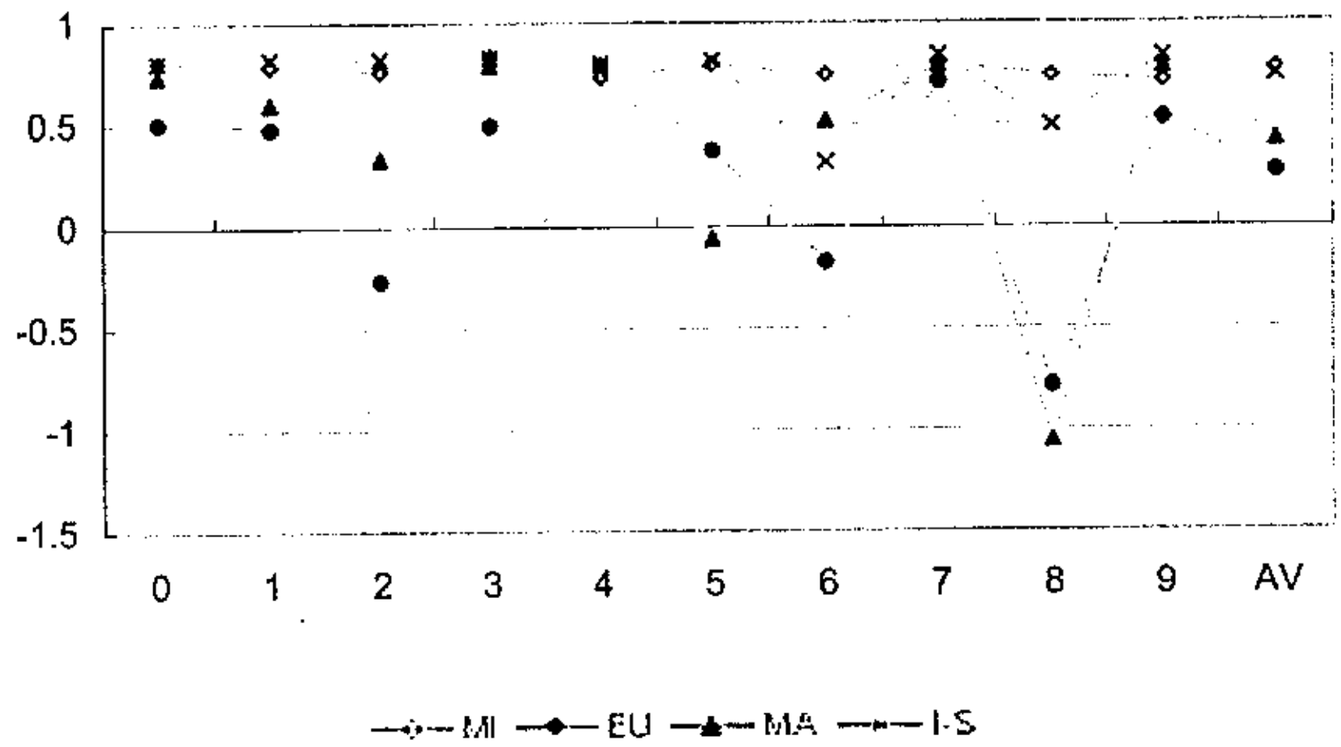


图 4.12　模式类间重叠指标 COP4

总之，耦合度指标和重叠指标都反映模式类间的分离程度，从指标分布曲线来看也基本相似。但从反映实际识别性能的作用来看，重叠指标要更加清晰些，因为正值表示无重叠，在实际的多维模式空间更不可能有重叠，负值表示有重叠，但在实际的多维模式空间可能不一定存在。以下的实际语音识别实验评价中也说明了这一点（见表4.1）。

4.5 基于互信息匹配的语音识别

语音信号模式的互信息测度充分考虑了语音信号的时变分布特征和统计分布特征，不同类型的特征信息得到了更多的提取和运用。互信息估计的非线性搜索算法进一步增强了对时变特征非线性波动的处理能力，对互信息测度性能的提高有较大的帮助。互信息作为一种模式匹配失真测度不仅可以应用于说话人识别，同样可以应用在语音识别中。当互信息测度应用于语音识别时，互信息指语音识别系统中每个待识别词汇对应语音信号之间的互信息，通过计算输入语音信号测试模式与各词汇对应语音信号参考模式之间的互信息值进行识别判决。

以下介绍互信息估计的非线性搜索算法在数字语音识别、连续语音识别以及因特网语音识别浏览器 VoiceIE 设计开发中的应用情况。

4.5.1 数字语音识别实验

数字语音识别实验对汉语中具有典型声学语音特征的数字集 0-9 进行了测试。实验环境及数据如下：

（1）识别对象：0-9 共 10 个数字语音，由单一参考模式表示。

（2）语音样本：利用声卡输入，220 个样本信号。其中，30 个用于训练，全部用于测试。

（3）短时帧与移位：每帧 18ms，矩形窗，相邻帧无重叠。

（4）特征参数：8 阶 LPC 参数，特征矢量采用前 4 个参数。

（5）测试环境：普通实验室。

（6）模式匹配：非线性搜索互信息匹配 NLM

为了重点观察互信息测度的性能，在信号预处理部分并没有实行预增强，短时窗的选择以及帧的移位不作特别考虑。实验结果如表 4.1 所示，其中列出了各特征参数下利用各种测度计算模式距离和似然度得到的各数字的错误识别数、数字的平均误识数（AV）和误识率。可以看到，当使用 LPC 特征参数时，互信息 MI 测度与 I-S 测度具有相近的识别性能，后者还要

略高一些，但都明显优于 EU 和 MA 测度的识别性能。当使用线性预测倒谱系数 LPCC 作为特征参数时，互信息测度 MI 的识别性能与 LPC 特征参数情况下相比有 65%的提高，并且优于 EU 测度的识别性能，误识率仅为后者的 50%。I-S 测度不能应用于 LPCC 特征参数，MA 测度的识别性能很差，没有列出。

表 4.1　互信息与各种测度下的数字语音识别实验结果

特征参数	测度	0	1	2	3	4	5	6	7	8	9	AV	误识率
LPC	MI	0	0	0	0	1	0	3	0	2	2	0.8	3.64%
	EU	0	1	2	0	0	0	6	0	4	0	1.3	5.91%
	MA	1	2	4	0	0	5	2	0	6	1	2.1	9.55%
	IS	0	0	0	0	0	0	4	0	3	0	0.7	3.18%
LPCC	MI	0	0	0	0	0	0	1	0	1	1	0.3	1.36%
	EU	0	0	0	0	0	0	3	0	3	0	0.6	2.73%

4.5.2 语音识别浏览器 VoiceIE

语音识别浏览器 VoiceIE[117]使得用户能够以语音口语方式进行因特网 Web 浏览，同时提供普通方式的因特网浏览。在普通工作方式下，浏览操作与 MS-IE 相同。在语音工作方式下，可以使用语音输入网址、浏览操作以及控制鼠标和窗口显示等。VoiceIE 的系统结构如图 4.13 所示。

VoiceIE 系统由 Web 浏览器 VoiceBrowser 和语音识别引擎 SDSE[123]这两大模块构成。语音指令库中包含常用的初始化指令 50 条，但这些语音指令可以根据需要扩充和更改、删除。例如，与网址搜索、下载和显示有关的指令可以根据新的需要随时改变，而浏览操作中光标和鼠标的移动控制、窗口管理等指令基本上是固定的。

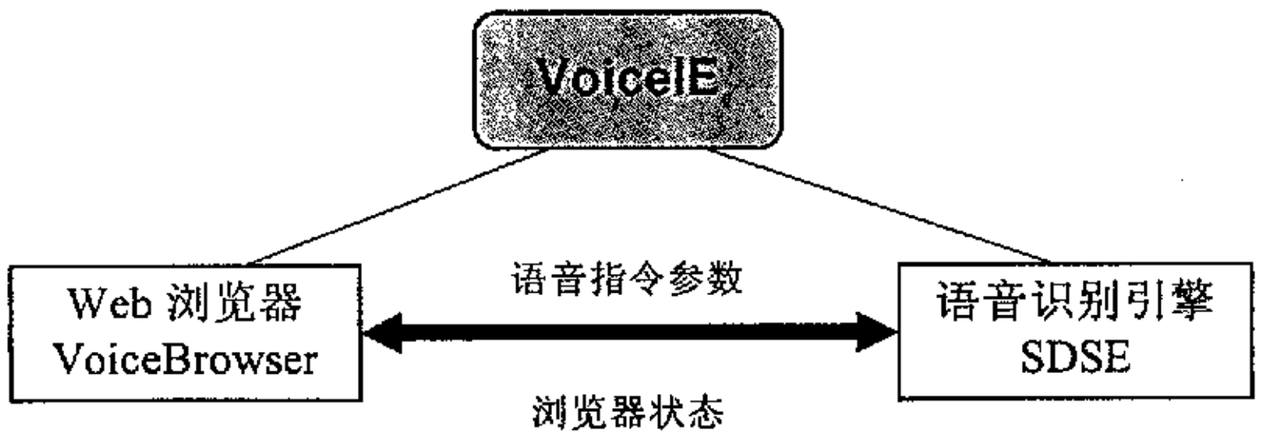


图 4.13　语音识别浏览器 VoiceIE 系统构成

在 VoiceIE 中，语音模式的表示除 LPCC 特征参数之外，还利用了语音指令的时长参数对语音指令模式进行预分类，模式匹配采用非线性搜索互信息匹配算法 NLM。作为一个实用系统应该尽量减少系统的训练量，因此，语音指令的一次训练识别率就很重要。在 29 个网址搜索指令和 21 个其它浏览操作指令的情况下，VoiceIE 一次训练识别率为 93.5%，三次训练后识别率达到 99%，并且完全实时处理。

4.5.3 连续语音识别实验

“旅游服务语音对话系统”是应用于苏州市虎丘旅游服务的非特定人语音识别系统[124]，能够识别理解 52 条有关旅游询问的语句，同时能够识别 139 个音节，速度基本实时。

训练时，通过对 10 名男性说话人和 10 名女性说话人的语音数据进行短时分析，提取 LPCC 系数构成特征矢量，在此基础上进行聚类分析，形成非特定人识别所需的音节参考模式。每一个音节经过训练后形成 8 个侯选参考模式，系统参考模式库共包含 1112 个标准模式。识别时，将连续输入语音所对应的音节序列中的每个音节测试模式与各音节参考模式进行互信息匹配，并根据互信息值的大小选择前五个最佳匹配音节作为侯选输入到语言处理部分进行关键词句法结构分析。

输入语音信号由哈明窗进行 30ms 的短时分段处理，相邻短时帧重叠 10ms。语音特征参数选择 LPCC 倒谱系数。对 139 个连续语音音节进行的非特定人识别测试表明：音节识别率平均为 82%，具体识别情况及结果如表 4.2 所示。

表 4.2 连续语音的音节识别实验结果

	非训练说话人			训练说话人		
测试音节数	5560			5560		
错误识别数	1223	1130	1101	932	824	806
音节识别率	78%	79.7%	80.2%	83.3%	85.2%	85.5%

语音对话系统在音节识别之后进一步进行关键词分析，运用句法结构知识进行语句识别，从而理解语句的含义，因此，实际的音节识别输出结果包含了前五个具有较大互信息的侯选识别音节。

4.6 结 论

语音信号之间的互信息揭示了相互之间的相似程度，因此，作为一种与似然度相似的测

度可以应用于模式识别中，包括语音识别和说话人识别。但是，由于语音信号复杂的非平稳时变特性以及无法直接得到其精确的统计分布，所以，互信息的计算只能通过估计来实现。

在互信息估计计算中，通过提出随机干扰信号描述语音信号之间的失真与差异，并运用随机信号统计分布理论分析得出其统计分布特性，在此基础上解决了互信息的计算问题。互信息的具体计算有线性映射匹配 LPM 和非线性搜索匹配 NLM 两种方法。前者在计算中考虑了语音信号的统计分布特性和时变特性，但无法处理信号的非线性变化，如语速的时快时慢等；后者采用非线性动态规划的方法搜索最佳匹配路径，得到最大化准则下的互信息估计值，同时很好地考虑了信号的统计和线性非线性时变特性。

互信息作为一种失真测度，与基于距离空间的 Euclidean 测度、Mahalanobis 测度和 Itakura-Saito 测度相比具有较高的类内凝聚度和较小的类间耦合度，从模式识别的角度分析，模式之间的分离性较高。语音识别实验和应用也证明了互信息估计算法的有效性以及基于互信息匹配的较高识别性能。

第五章　互信息应用在基于文本的说话人识别

本章提要：

- 提出说话人语音的多模板模型 MTM
- 基于 MTM 与 NLM 匹配算法建立新的说话人识别方法
- 识别实验分析与比较：MTM-NLM，DTW，GMM

基于文本的说话人识别系统中，每一个说话人有特定的识别文本，系统通过分析与识别文本对应的输入语音判别说话人的身份。显然，即便所有说话人采用相同的识别文本，同一个说话人在不同时间的发音相似度应该大于不同说话人发音之间的相似度，从语音信号所携带说话人特征信息的角度考虑，就是同一说话人的语音之间互信息较大，不同说话人的语音之间互信息较小。因此，可以利用语音信号之间的互信息进行说话人识别。基于文本的说话人识别系统原理如图5.1所示，输入语音与各侯选说话人参考语音进行互信息匹配，并将具有最大互信息的说话人作为识别结果。

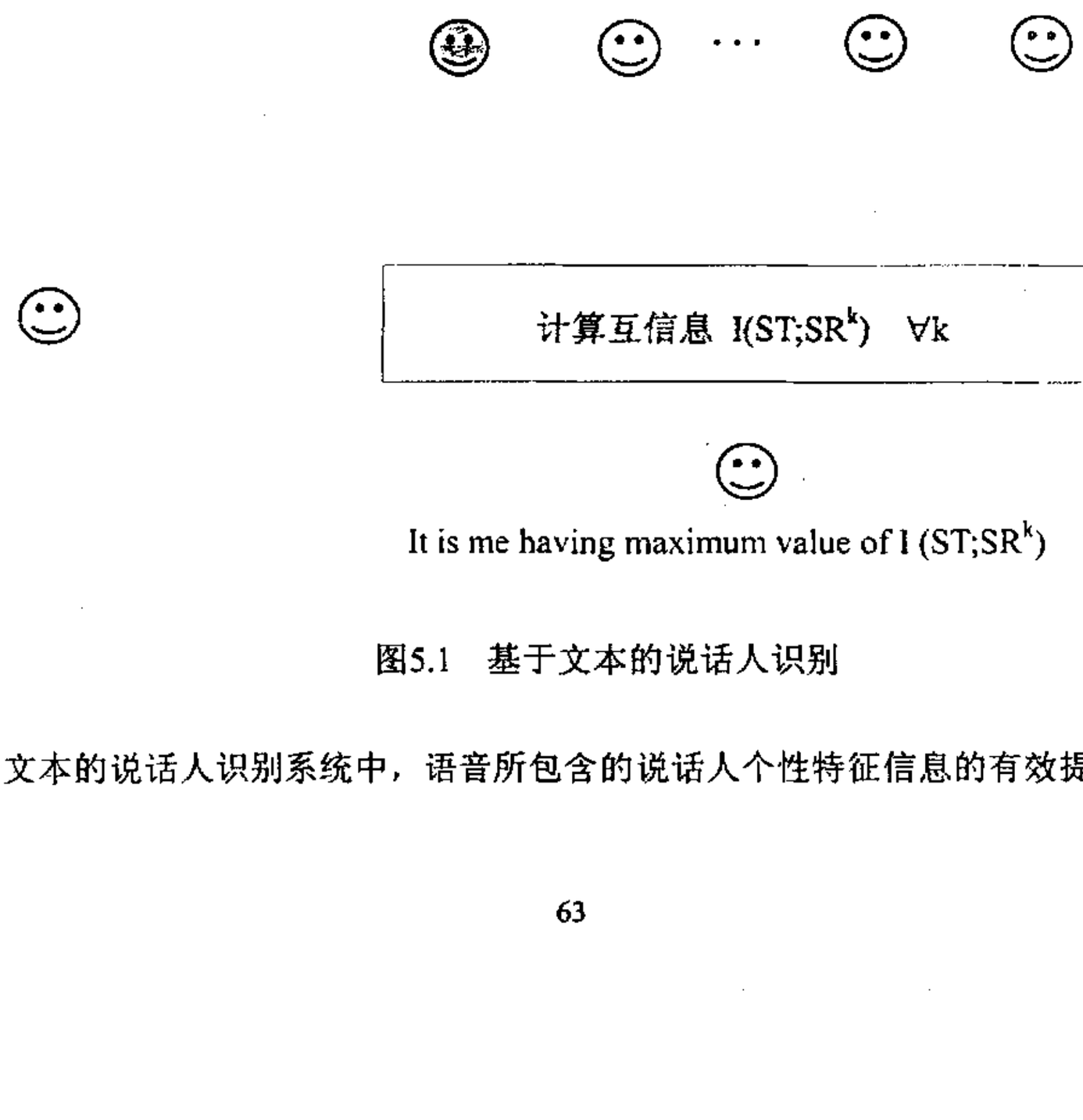


图5.1　基于文本的说话人识别

基于文本的说话人识别系统中，语音所包含的说话人个性特征信息的有效提取与识别是

关键所在。尽管在实际的说话人识别系统中可以考虑不同说话人采用不同的文本，以通过增加语义信息来提高识别性能，但研究过程中应该考虑所有说话人采用相同文本的较苛刻条件，即仅仅利用语音中包含的说话人个性特征信息进行识别。

在基于互信息匹配的说话人识别系统中，说话人语音模式之间的失真度采用互信息来计算，其值越大失真越小，根据输入语音信号对应的测试模式与系统各参考模式的互信息就可以进行识别判决。如前所述，互信息的计算可以选择线性映射和非线性搜索匹配算法计算语音之间的互信息。与其它匹配识别方法不同的是，互信息匹配识别方法能够综合运用和处理语音信号的统计分布特征与时变分布特征，具有较好的识别性能和鲁棒性 。这一章分析互信息匹配识别模型在基于文本的说话人识别中应用的可行性与性能特点，通过说话人辨认实验对互信息匹配识别模型的说话人识别性能进行分析，并与传统的说话人匹配识别方法 DTW 和 GMM 进行比较。

5.1　互信息匹配识别原理

基于文本的说话人识别系统结构如图 5.2 所示，包括输入语音信号预处理、特征提取、说话人模型建立以及模式匹配和判决等几个部分。互信息匹配识别与其它识别方法的主要区别是在说话人模型和模式匹配识别判决部分。

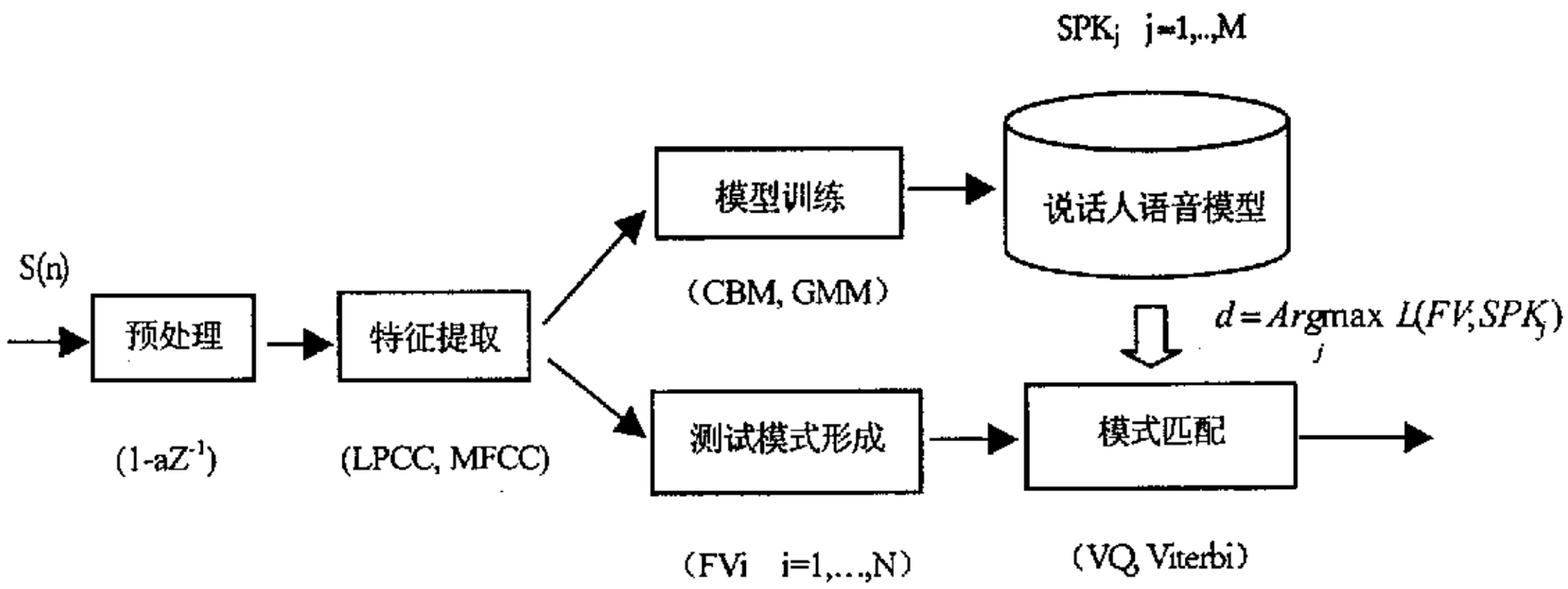


图 5.2 基于文本的说话人识别系统框图

5.1.1 多模板说话人模型 MTM

说话人模型反映说话人的语音发音特征。目前运用的说话人模型主要有两种，即模板模型和统计模型。前者属于非参数模型，如矢量量化模型CBM（Code-Book Model）[58]，后者

则属于参数模型，如高斯混合模型GMM（Gaussian Mixture Model）[59]。基于文本的说话人识别应该考虑语义信息的利用或保留，但CBM和GMM都失去了反映语义信息的时变特征，因此，这两种模型不会是基于文本的说话认识别应用中的理想模型。结合互信息匹配计算的条件和特点，这里提出一种多模板模型MTM（Multi-Template Model）应用于基于互信息匹配的文本有关说话人识别。

设某说话人SPK的训练语音集为 $S:\{S_1, S_2, \ldots, S_N\}$，所有的样本语音信号 $S_i, i=1\sim N$ 经过预处理和特征提取得到对应的特征序列 $SV_i:\{V_1^i, V_2^i, \ldots, V_L^i\}, i=1\sim N$，根据不同的特征参数决定特征矢量 V_l^i 的阶数和内容。另外，由于每次发音的时长有变化，所以每一个特征矢量序列的长度L并不都一样。

MTM的训练由两大部分构成，首先对训练用样本语音集S所对应的各个特征矢量序列进行聚类分析，按照指定的类别数将各特征矢量序列划分到各个子类中，然后求出各子类的中心特征矢量序列作为说话人模型，由于模型由多个子类的中心特征矢量序列构成，因此称为多模板模型MTM。设与训练语音集 $S:\{S_1, S_2, \ldots, S_N\}$ 对应的训练样本特征矢量序列集合为 $SV:\{SV_1, SV_2, \ldots, SV_N\}$，则具体的训练算法如下：

(1) 设定子类类别数M，并从 SV 中随机选择M个特征矢量序列作为初始子类中心或模板 $\{TM_1^o, TM_2^o, \ldots, TM_M^o\}$，例如，$TM_k^o = SV_k$，$k=1\sim M$。

(2) 计算各特征矢量序列与模板的距离，并将其归入具有最小距离的子类，即

$$SV_i \Rightarrow SVS_q;\quad q = \underset{k}{Argmin}\ d(SV_i, TM_k^o);\quad i=1\sim N$$

(3) 计算各子类 SVS_k；$k=1\sim M$ 的中心特征矢量序列，即

$$TM_k = Avr(SV_i \mid SV_i \in SVS_k);\quad k=1\sim M$$

(4) 如果存在新的模板与原模板不一致情况，即 $TM_k \neq TM_k^o$；$k\in\{1,2,\ldots,M\}$，则清除各子类，并更新模板，$TM_k^o = TM_k$；$k=1\sim M$，转（2）。

(5) 建立说话人多模板模型MTM，该模型由以上计算得到的各模板以线性同权值方式构成，如下：

$$MTM:\{TM_1, TM_2, \ldots, TM_M\}$$

以上模型训练是一个叠代计算处理过程，采用的聚类方式与K-means方法相似。特征矢量

序列之间的距离采用线性映射条件下的累积Euclidean 距离，公式如下：

$$d(\boldsymbol{SV}_i,\boldsymbol{TM}_k)=\frac{1}{L1}\sum_{j=1}^{L1}\|\boldsymbol{SV}_{f(j)}^i-\boldsymbol{TMV}_j^k\| \qquad (5.1)$$
$$f(j)=\frac{L2-1}{L1-1}(j-1)+1$$

其中，L1是模式$\boldsymbol{TM}_k$的长度（或特征矢量数），L2是特征矢量序列$\boldsymbol{SV}_i$的长度。

各子类中心特征矢量序列或模板$\boldsymbol{TM}_k$，$k=1\sim M$的计算通过对属于同一类的特征矢量序列求平均得到，其方法是首先求子类特征矢量序列的平均长度LAV，然后将子类中各特征矢量序列都线性映射为LAV长度，即

$$\boldsymbol{SV}_i=\{V_1^i,V_2^i,...,V_L^i\}\Rightarrow\boldsymbol{SVC}_i=\{\boldsymbol{VC}_j^i=V_{g(j)}^i\mid j=1\sim LAV\} \qquad (5.2)$$
$$g(j)=\frac{L-1}{LAV-1}(j-1)+1,\quad j=1\sim LAV$$

这样，子类中的所有特征矢量序列都有相同的长度LAV，因此，子类中心特征矢量序列就可以通过对所有序列求平均来得到，如下：

$$\boldsymbol{TM}_k=\frac{1}{S}\sum_{i=1}^{S}\boldsymbol{SVC}_i=\{\frac{1}{S}\sum_{i=1}^{S}\boldsymbol{VC}_1,\frac{1}{S}\sum_{i=1}^{S}\boldsymbol{VC}_2,...,\frac{1}{S}\sum_{i=1}^{S}\boldsymbol{VC}_{LAV}\},\quad \boldsymbol{SV}_i\in\boldsymbol{SVS}_k;\quad k=1\sim M \qquad (5.3)$$

这里，S代表子类的特征矢量数目。显然，为了使说话人模型MTM能够全面地反映说话人特定文本的语音发音特征，样本语音信号不仅要充分（N>>M），而且应该反映说话人发音的时变特性，即语音样本的采集有一定的时间分布。

5.1.2 基于模式的非线性搜索互信息匹配与识别判决

模式匹配部分进行输入语音信号模式与系统各说话人模型的比较，并根据最大互信息准则进行判决，识别输入语音说话人的身份。

设系统有N个侯选说话人$SPK_1,SPK_2,\ ...\ SPK_N$，相应的说话人模型为MTM_i，$i=1\sim N$，每个多模板模型MTM_i如前所述由反映说话人语音发音特征的多个模板$\boldsymbol{TM}_k$，$k=1\sim M$构成。输入语音信号S_T经过预处理和特征提取转换为测试语音特征序列$\boldsymbol{ST}:\{\boldsymbol{VT}_1,\boldsymbol{VT}_2,...,\boldsymbol{VT}_{LT}\}$。因此，互信息匹配需要计算$\boldsymbol{VT}$与$MTM_i:\{\boldsymbol{TM}_1^i,\boldsymbol{TM}_2^i,...,\boldsymbol{TM}_M^i\}$，$i=1\sim N$之间的互信息，得到$N\times M$个互信息值。

输入语音信号模式$\boldsymbol{VT}$与模型$\boldsymbol{TM}_k^i$，$k=1\sim M$　$i=1\sim N$的时长并不一致，因此，它们之间的互信息采用图4.3所示的非线性搜索匹配估计算法NLM进行计算。在其它基于模板的模式匹

配中，模式之间的距离或似然度的计算是通过短时帧之间距离的累积而形成，但基于互信息估计的模式匹配是直接通过模式的全体特征矢量并按照一定的时序进行估计的，并没有计算短时帧之间的互信息。因此，互信息匹配是基于模式的匹配计算方法。

识别判决的准则有最佳近邻准则N-N（Nearest Neighbor）和最大似然准则（Maximum Likelihood）。互信息匹配计算反映输入模式与说话人模型之间的相似程度，其判决准则采用最大互信息准则MMI（Maximum Mutual Information）。判决公式如下：

$$i^* = Arg\max_i \quad I(ST; MTM_i) \tag{5.4}$$

或

$$i^* = Arg\max_i \quad Max_k(I(ST; TM_k^i)) \tag{5.5}$$

即识别说话人所对应的模型与VT具有最大的互信息。

5.2 其它匹配识别方法

基于文本的说话人识别由于可以利用语义信息，因此也可以运用语音识别中相同的匹配识别方法，如基于模板模型的动态时间弯折 DTW（Dynamic Timing Warping）匹配识别方法、基于高斯混合模型 GMM（Gaussian Mixture Model）的最大似然匹配识别方法和基于神经网络模型 ANN（Artificial Neural Network）的匹配识别方法等。目前常用的有 DTW 和 GMM。

5.2.1 DTW 匹配识别方法

动态时间弯折 DTW 是一种依据模式距离进行识别的方法。在具体的距离计算中，使识别文本的参考语音模式与测试模式在一定的范围内沿时间轴进行非线性对准，以减少语音动态频谱的非线性波动对匹配的影响。DTW 以语音帧距离的非线性累加表示模式之间的距离，识别时以最小距离准则进行判决。

设对应第 i 个说话人的参考语音模式为 $RP_i = \{R_1^i, R_2^i, ..., R_M^i\}$，输入语音测试模式为 $TP = \{T_1, T_2, ..., T_N\}$，则由 DTW 匹配得到模式距离：

$$D(TP, RP_i) = \frac{1}{L}\sum_{l=0}^{L-1} dist(T_{f(l)}, R_{h(l)}^i) \tag{5.6}$$

其中，L 是最佳非线性搜索匹配路径的长度，$f(l)$，$h(l)$ 分别是测试模式时间轴与参考模式时间轴到非线性搜索匹配路径的映射函数。dist(.)是帧距离测度。DTW 识别模型将具有最小匹配

距离的参考模式所对应的说话人 S_d 作为识别结果，其判决准则如下：

$$d = \underset{i}{Argmin} \quad D(TP, RP_i) \tag{5.7}$$

5.2.2 GMM 匹配识别方法

高斯混合模型GMM的实质是用若干高斯概率分布的混合加权组合去拟合语音信号的实际概率分布，然后运用 Bayes 最大似然准则进行识别判决[59]，因此该模型纯碎是一种统计模型。GMM 模型对应的混合高斯分布如下：

$$p(x \mid \lambda) = \sum_{i=1}^{N} p_i b_i(x) \tag{5.8}$$

其中 p_i 是混合系数；$b_i(x)$ 是高斯概率分布，其相应的均值为 μ_i，协方差为 W_i；N 是混合成份数；λ 表示 GMM 模型的系数集合，由混合系数 p_i，高斯概率分布均值 μ_i 和协方差矩阵 W_i 构成，共有 N 组。

设有 M 个说话人 S_1, S_2，... S_M，每一说话人的 GMM 模型采用 EM 算法进行训练[59]，得到模型系数集 λ_i，$i = 1,2,...,M$，在识别时运用 Bayes 最大似然准则对输入语音信号 x_i 进行识别，判决准则如下：

$$d = \underset{k}{Arg\max} \quad p(x_i \mid \lambda_k) \tag{5.9}$$

即，如果说话人 S_d 的 GMM 模型具有最大概率值，则识别结果为 S_d。

5.3 实验分析

实验分析中采用了 SUDA2002-D1 语料库，其中包含了 30 名母语为日语的说话人的 3600 个文本语音。30 名说话人中，18 名为男性，12 名为女性，年龄分布在 21-44 区域。识别文本选择了具有代表性的 12 个地方名称，见表 5.1，其语音成份中包含了各种类型的日语语音音素。每个说话人输入每个文本语音 10 遍，30 名说话人共输入 3600 个文本语音，时间分布为两个月。语音信号在普通实验室环境下通过计算机声音系统输入，采样频率为 11025Hz，16 位量化。

为了观察不同文本语音下互信息匹配的识别性能，实验由 12 个阶段构成，每个阶段选择 12 个地名文本的一个进行模型训练并测试。每个说话人每一个文本语音的前 5 次发音用于模型训练，形成由 3 个模板构成的 MTM 模型，所有语音数据都用于测试。

表 5.1　识别文本

代码	文本与发音	代码	文本与发音
HD	北海道　Hokkaido	NR	奈良　Nara
HS	广岛　Hiroshima	NS	长崎　Nagasaki
KB	神户　Kobe	OS	大阪　Osaka
KS	九州　Kyushu	SD	仙台　Sendai
KT	京都　Kyoto	TK	东京　Tokyo
NG	名古屋　Nagoya	YH	横浜　Yokohama

5.3.1 识别性能分析

首先，在相同的数据和实验条件下对基于多模板模型 MTM 的互信息匹配识别方法的说话人识别性能进行分析，并与 DTW 识别方法和 GMM 识别方法进行比较。

在运用 DTW 的匹配识别实验中，作为说话人模型的参考模式同样采用每个文本的前 5 个输入语音训练形成，每个文本采用 K-Means 聚类形成三个参考模式。GMM 模型的训练所用数据与 MTM 训练所用数据完全一致，采用期望最大化叠代算法 EM（Expectation Maximum）得到每个说话人 GMM 的参数[59]。根据对局部数据所做的实验分析，GMM 模型的高斯分量数目取 10 时识别性能最佳，因此，实验中 GMM 模型都由 10 个高斯分量混合构成。

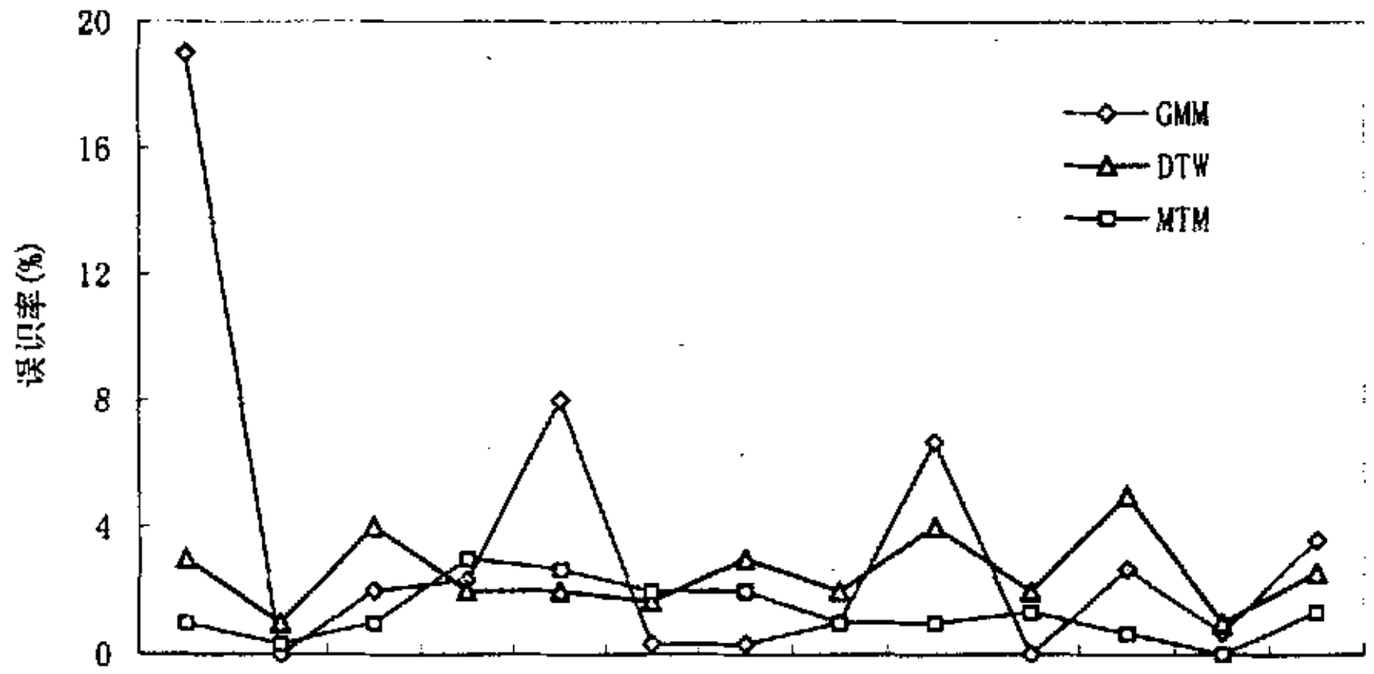


图 5.3　MTM、DTW 与 GMM 的说话人识别性能比较

当以 8 阶 LPCC 参数形成特征矢量时，各识别方法对应各个文本的说话人识别性能情况如图 5.3 所示，图中纵轴表示误识率，横轴为文本代码，其中 AV 表示所有文本的平均。MTM、DTW 和 GMM 的平均错误识别率分别为 1.33%、2.56%和 3.58%。其中，考虑到运行时间以及

各 LPCC 特征参数的权重，MTM 模型在识别时仅仅利用了 8 阶 LPCC 参数的前 4 个参数，并采用非线性搜索匹配算法 NLM 计算互信息值。DTW 采用 Euclidean 距离测度计算帧间距离，匹配采用动态规划算法进行，调整窗宽度为 6。

由图 5.3 可以看出，基于 MTM-NLM 的互信息匹配识别方法具有较好的平均识别性能，DTW 识别方法的平均识别性能其次，两者的识别性能总体上都好于 GMM 模型，但 GMM 模型在某些文本上的识别性能要好于 MTM 和 DTW，特别是文本语音中鼻音和摩擦音成份较多时。MTM-NLM 识别方法综合考虑了语音信号的统计与时变分布特征，因此说话人语音的特征能够在两个方面得到有效的提取与处理。DTW 仅仅考虑了语音的时变分布特征，而 GMM 仅仅考虑了语音的统计分布特征。从 DTW 识别性能好于 GMM 这一点来看，在基于文本的说话人识别中，语音信号的动态频谱时变特征的处理很重要。另一方面，从语音信号统计特征的提取与处理上分析，MTM 模型仅仅采用单个高斯概率分布表示语音信号的统计特征，而 GMM 采用了多个高斯分量的混合来逼近实际概率分布，但识别性能却是 MTM 更好，这同样说明了在基于文本的说话人识别中对说话人语音的时变特征的提取与处理很重要。对于“Nara”、“Hiroshima”等鼻音和擦音成分较多的语音，相应的语音动态频谱变化较平滑，时变特征上没有突变现象，因此，这种情况下 GMM 的识别性能较好，而对于“Hokkaido”、“Kyoto”等包括促音和拗音的语音，其动态频谱的变化存在较大的瞬间起伏，时变特征中有较多的突变现象，这些特征 GMM 不能有效地表达，因此识别性能较差。另外，从图 5.3 可以看出，MTM-NLM 识别方法对各类语音的识别性能比较稳定，这也是综合运用语音信号的统计与时变分布特征的结果。

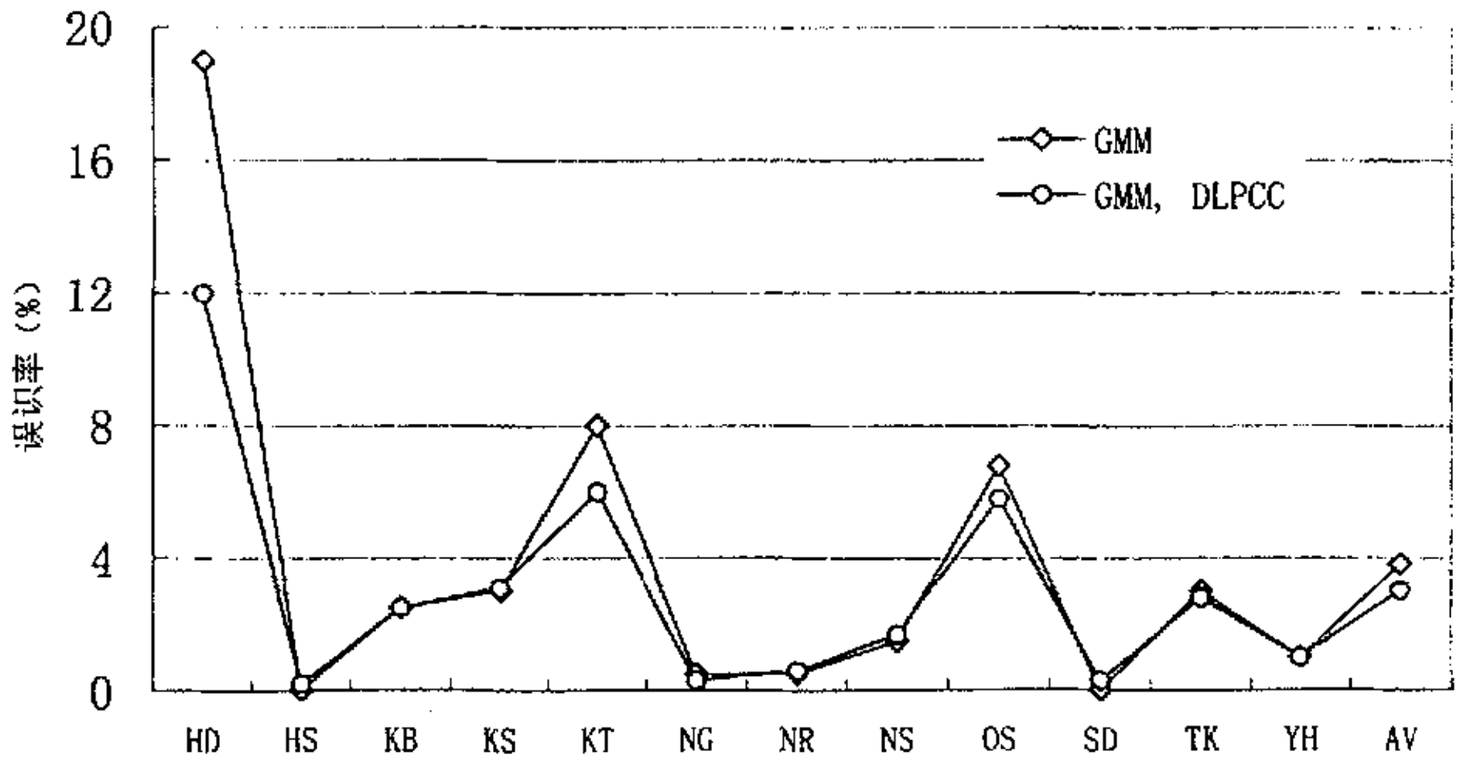


图 5.4　运用 DLPCC 时 GMM 的说话人识别性能改善

很难在 DTW 识别模型中融合语音信号的统计分布特征，但对于 GMM 模型，可以通过增加时变特征参数来克服模型本身仅仅利用语音信号统计分布特征的不足，提高识别性能。例如，当增加四阶帧间差分倒谱参数 DLPCC 时，在上述同样的实验条件下，其识别性能提高了 0.7%，对应各文本的具体识别情况如图 5.4 所示。

模型的训练对于识别性能的提高也很重要，在实际应用中可以进一步增加训练语音样本数据量来提高说话认模型的性能。以上实验中，由于语料库数据的限制，MTM 和 DTW 匹配识别方法中仅仅使用了每个文本的 5 个语音样本数据进行说话人模型的训练不是很充分，同样，GMM 训练时如果有更多的数据则模型会更完善。可以推断，当将以上实验系统转化为实际应用系统时，如果给予足够的训练数据，对应 MTM 模型的识别率还可以进一步提高。

5.3.2 识别性能与特征参数的关系

对于互信息匹配识别而言，当运用 LPC 或 LPCC 等特征参数时，在阶数一定的情况下一般仅需采用前几个权重较大的参数，采用更多的特征参数对识别性能的提高作用并不大[116] 。那么，当阶数发生变化时相应的选择范围是否也应该变化呢？图 5.5 表示当 LPCC 的阶数为 6、8 和 10 时，仅仅采用前 4 个特征参数时相应的误识率，其总体平均错误识别率分别为 1.67%、1.33%和 1.78%。

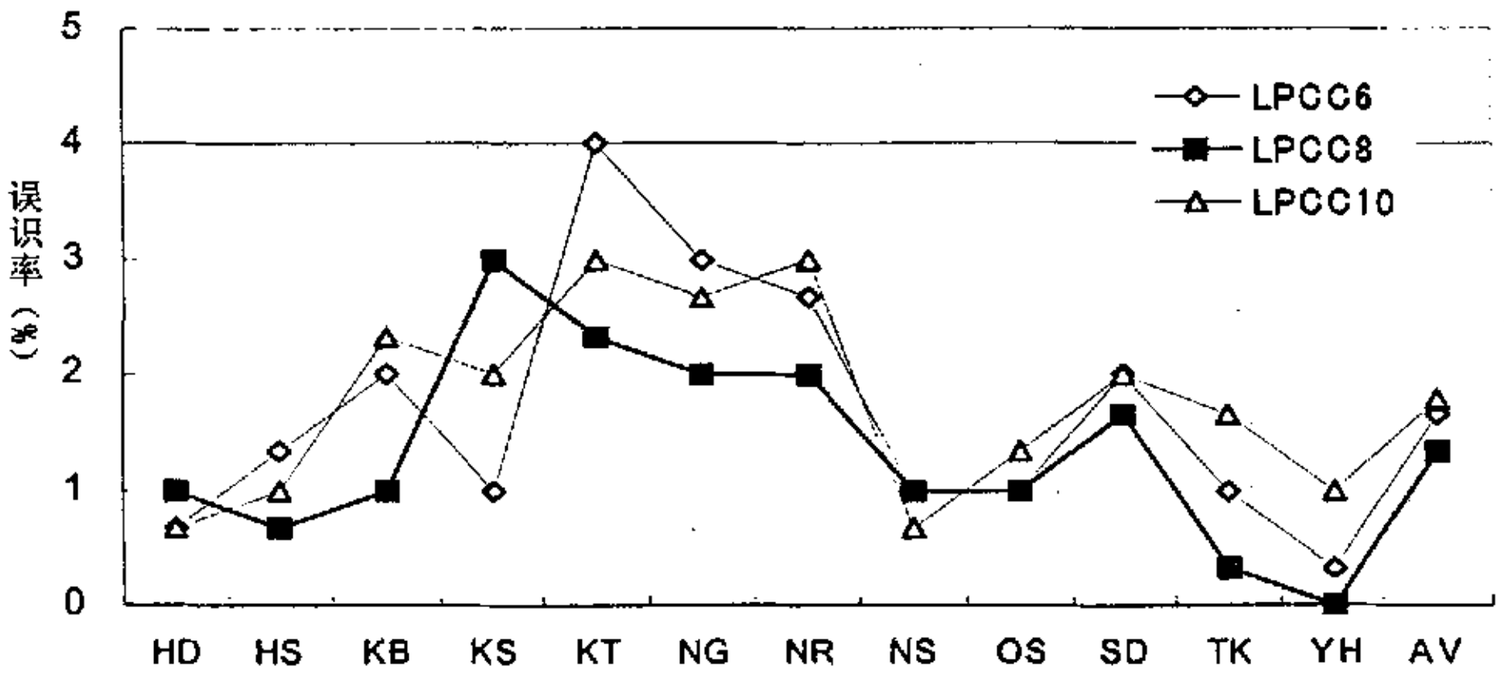


图 5.5　MTM 识别模型对应各阶 LPCC 特征参数的误识率

图 5.5 显示当采用 8 阶 LPCC 的前 4 个参数时 MTM-NLM 识别方法具有较好的识别性能。当 LPCC 阶数取 6 时，由于相应的预测模型拟合语音信号不如 8 阶时充分，因此采用其前 4 个参数时提取的说话人语音特征也不如 8 阶时充分，识别性能有所下降。当 LPCC 阶数取 10 时，虽然相应的线性预测模型可以更好地拟合语音信号，但前 4 个参数的权重减少了，说话

人的语音特征被更多地转移到其它参数，因此识别性能没有因为阶数的提高而提高，此时就应该采用更多的特征参数，例如前 6 个。本文对运用 LPC 参数时 MTM-NLM 识别方法的说话人识别情况也做了实验分析，实际结果表明运用 LPC 参数的识别性能不如 LPCC ，因此，MTM 与 DTW 和 GMM 的情况一样，LPCC 比 LPC 更适合作为说话人语音的特征参数。

5.3.3 运算效率分析

设特征矢量长度为 P，指数和对数函数的运算量相当于 5 次乘法，GMM 的高斯分量数为 N，则识别时 MTM-NLM、DTW 和 GMM 对应语音信号每一帧的乘法计算量分别为 P*(P+1)/2、P 和(3*P+12)*N。基于 MTM 模型的识别计算中，自协方差矩阵是一个正定对称矩阵，因此在每帧所对应的元素计算中仅仅需要计算约一半的元素，但在非线性路径搜索中的局部判决以及最终互信息值的计算中还需要进行矩阵运算，对应每帧的乘法计算量大约为 3*P!，这样，每帧所对应的乘法总计算量为 P*（P+1）/2+3*P!。DTW 在动态规划匹配时需要对局部路径进行计算选择，相应的乘法计算量增加为 3*P。若按以上识别实验中使用的参数个数计算，各模型每一帧的乘法计算量分别为 82、24 和 360，从运行效率上看，DTW 的效率较高，GMM 的效率较低。

5.4 结 论

提出了说话人语音的多模板模型 MTM，并基于这一模型与非线性搜索互信息匹配算法 NLM 在基于文本的说话人识别应用中进行了实验分析与评价，并与基于 DTW 和 GMM 的识别方法进行了比较。实验结果表明，MTM-NLM 识别方法在总体上具有较好的识别性能，并且，LPCC 系数比 LPC 系数更适合作为特征参数应用于说话人识别。在基于文本的说话人识别中，不仅统计分布特征，语音的时变分布特征的运用和处理也很重要，因此，在特征参数与匹配识别方法的选择中应该对此加以考虑。GMM 模型对鼻音以及摩擦音较多的文本语音识别性能很好，但由于没有考虑语音信号的时变特征，因此当文本语音的时变特征较丰富时识别性能相对较差，要提高其性能，可以考虑在特征矢量中增加时变特征参数，如帧间差分 LPCC 参数等。另外，MTM 和 DTW 识别方法的训练与识别算法都有较严格的理论基础和精确的计算推导，但在目前的 GMM 模型系数估计算法中一般假设各高斯分布的协方差矩阵为对角矩阵，亦即各特征参数是不相关的，这并不符合实际情况，而且，在迭代过程中必须对其值进行下限限制，不能保证在每次叠代计算时各高斯分量值小于 1，这些都会影响模型的识别性能，

尤其是在训练数据较少的情况下，因此，有关 GMM 模型的训练算法还有待改善。

在实际的基于文本的说话人识别系统中，可以考虑不同的说话人采用不同的识别文本，此时，由于利用了文本的语义信息，各说话人语音的统计与时变分布特征差进一步扩大，因此，各种识别方法的识别性能都会有所提高，而 MTM-NLM 识别方法由于综合运用了统计分布与时变分布特征，其识别性能的提高将更为显著。

第六章　互信息应用在文本无关的说话人识别

本章提要：

- 提出说话人语音的全特征矢量集模型 CFC
- 提出多级最小最大搜索互信息匹配算法 MMS
- 基于 CFC 和 MMS 建立文本无关说话人识别方法
- 识别实验分析与比较：CFC_MMS 与 GMM，LPCC 与 MFCC

目前，在文本无关说话人识别中常用的说话人模型有GMM和CBM [58,59,125] 。GMM高斯混合模型以多个正态分布逼近语音信号的实际统计分布，并运用 Bayes 分类器进行识别，这一模型通过EM算法较好地解决了语音信号统计分布特征的估计问题，分类器简单可行，但由于对时频动态变化的瞬时特征缺乏充分的建模能力，因此，当说话人语音包含较快的动态瞬时变化特征时，识别性能会明显下降。CBM 模型采用矢量量化的方法形成每个说话人的语音码书，并运用矢量距离累加计算进行匹配，由于没有充分利用语音信号的统计特征，因此识别性能一般，鲁棒性不够好。

这一章提出一种说话人的全特征矢量集模型 CFC（Complete Feature vector Corpus）和基于互信息评估的多级最小最大搜索匹配算法 MMS（Multi-step Mini-Max Search）。通过 MMS 算法计算输入语音与 CFC 之间的互信息，并运用最大互信息判决准则 MMI（Maximum Mutual Information）判别说话人的身份。实验对 CFC 模型和 MMS 算法在特征参数 LPCC 和 MFCC 两种情况下的识别性能进行了全面的分析，并与 GMM 模型在同样的测试环境与条件下进行了比较，结果显示，CFC-MMS 的识别性能较好，MFCC 特征参数比 LPCC 更能反映说话人特征。以下 6.1 节介绍全特征矢量集模型 CFC 的建立，6.2 节介绍基于互信息评估的多级最小最大搜索匹配算法 MMS，第 6.3 节给出实验分析结果。

6.1　说话人的全特征矢量集模型

在文本无关的说话人识别中，说话人的语音模型应该能够充分反映说话人语音发音的个

性特征，并实行语义及时序的归一化。

全特征矢量集模型 CFC 的基本思想是通过对一组包含说话人各种语音发音个性特征的数据进行分析处理，提取相应的代表性特征矢量表示说话人语音模型，其训练过程如下。

训练语音信号由 N 段语音 S_1，S_2，…，S_N 组成，其包含了说话人不同语音发音以及语音韵律的特征。预处理部分对这一组信号进行去噪处理，保留纯语音成份，并合并成一个完全由语音数据构成的训练语音信号 S。进一步对 S 进行短时分析处理，提取特征参数，如 LPCC 或 MFCC 等，形成一原始特征矢量序列$\{\mathbf{V}_i, \forall i\}$。聚类分析部分运用聚类分析算法对原始特征矢量进行聚类计算分析，提取代表性的特征矢量：$\{\mathbf{FV}_1, \mathbf{FV}_2, \ldots, \mathbf{FV}_M\}$作为说话人的全特征矢量集模型 CFC。聚类分析计算可以采用 K-Means 和 LBG[78]等多种算法来实现，本文采用的算法如下：

（1） 设置全特征矢量集 CFC 的大小 M，并以等间隔方式选取原始特征矢量序列$\{\mathbf{V}_i, \forall i\}$中的 M 个特征矢量作为 CFC 的初始值：$\{FV_1^o, FV_2^o, \ldots, FV_M^o\}$。

（2） 计算各原始特征矢量与 CFC 中各特征矢量之间的距离，并将原始特征矢量赋予与其距离最小的 CFC 特征矢量所在子集，即

$$V_i \Rightarrow FVS_q, \quad q = Arg\min_k d(V_i, FV_k^o); \forall i \text{。} \tag{6.1}$$

（3） 对每个 CFC 特征矢量子集中的原始特征矢量在特征空间计算其均值，并将其作为新的 CFC 特征矢量，即

$$FV_k'' = \frac{1}{L}\sum_{i=1}^{L} V_i, \quad V_i \in FVS_k; \forall k \tag{6.2}$$

（4） 如果计算得出的新的 CFC 特征矢量 $FV_k'', \forall k$ 与原 CFC 特征矢量 $FV_k^o, \forall k$ 完全一致，则结束，并将该 CFC 作为说话人的全特征矢量集模型，否则继续。

（5） 将 $FV_k'', \forall k$ 替代原 CFC 特征矢量 $FV_k^o, \forall k$，转（2）。

由于用于训练的语音数据 S_1，S_2，…，S_N 包含了说话人不同语音发音的声学与韵律特征，因而聚类训练所形成的特征矢量集 CFC 反映了说话人的全语音特征，并且，这样的特征矢量集并不包含语义与时序信息，实现了模型的语义及时序归一化。

6.2 多级最小最大搜索匹配算法与判决准则

目前，互信息理论在语音识别中的应用主要在 HMM 模型的训练和参数优化、距离尺度

的理论描述和多频带分配等[61,63,65]。第 4 章提出了语音信号之间互信息测度以及相应的估计算法，其特点是用信息量来表示语音信号之间的失真度。互信息匹配计算中同时考虑了信号的时变特性与统计特性，由前几章介绍可以看到基于互信息理论的模式匹配方法在语音识别和基于文本的说话人识别中的有效性。但是，在文本无关的说话人识别中，输入语音与说话人模型之间的互信息匹配计算和基于文本的说话人识别中的应用有很大的区别。第一，互信息匹配计算不能利用或保留语义信息，而只能利用说话人的个性特征信息；第二，互信息匹配计算中声学特性应该作归一化处理，包括时序归一化和统计特性归一化。

6.2.1 基于互信息评估的文本无关说话人识别原理

设有N个说话人SPK_1，SPK_2，．．．．，SPK_N，其对应的说话人语音模型采用全特征矢量集模型，分别为CFC_1，CFC_2，．．．．，CFC_N 。其中，每一个模型CFC_k ，∀k包含M个代表性特征矢量，即

$$CFC_k : \{FV_1^k, FV_2^k, \ \ldots \ , FV_M^k\} \tag{6.3}$$

当某一说话人的测试语音输入时，经过预处理和特征提取，得到一特征矢量序列*XFV*：{$\mathbf{XFV}_1$，$\mathbf{XFV}_2$，．．．．，$\mathbf{XFV}_L$}。*XFV*和各说话人模型之间的互信息I(XFV; CFC_k)反映了两者之间相互携带的信息量，根据互信息原理，信息量越大两者的相似程度越大。因此，可以根据最大互信息准则MMI来判决输入语音属于哪个说话人，即识别的说话人d应该满足下式：

$$d = Arg \max_k \ I(XFV; CFC_k) \tag{6.4}$$

如前所述，全特征矢量集模型CFC_k ，∀k已经实行了语义和时序的归一化，因此，在进行互信息计算中，只要通过适当的处理就可以实现语义与时序的归一化。

6.2.2 多级最小最大搜索匹配算法 MMS

依据互信息理论，输入语音信号 XFV 与说话人模型 CFC_k 之间的互信息可以由下式计算，

$$I(XFV; CFC_k) = H(XFV) - H(XFV \mid CFC_k) \tag{6.5}$$

其中，$H(XFV)$ 表示输入语音信号的熵，$H(XFV|CFC_k)$ 表示条件熵。

上式中第一项与模型无关，在识别判决中可以不加以考虑。第二项条件熵的计算需要运用条件概率，但这样的条件概率无法精确获取，需要通过一定的训练或估计获得。

对输入语音信号 XFV 的每一个特征矢量 $\boldsymbol{XFV}_j$ ∀j，求与说话人模型 CFC_k 的最佳匹配代

表性特征矢量 $\boldsymbol{BFV}_j^k$，得到一个 XFV 与说话人模型 CFC_k 的最佳匹配代表性特征矢量序列，如下：

$$\begin{aligned} &BFV_k:\{\boldsymbol{BFV}_1^k,\boldsymbol{BFV}_2^k,\ \cdots\ ,\boldsymbol{BFV}_L^k\} \\ &\boldsymbol{BFV}_j^k=\boldsymbol{FV}_o^k \\ &o=\arg\min_i\ \left\|\boldsymbol{XFV}_j-\boldsymbol{FV}_i^k\right\| \end{aligned} \tag{6.6}$$

对于特定的输入语音 XFV，由最佳匹配代表性特征矢量序列 BFV_k 代表说话人模型 CFC_k 求相互之间的互信息，计算公式如下：

$$I(XFV;CFC_k)\Rightarrow I(XFV;BFV_k)=H(XFV)-H(XFV\mid BFV_k) \tag{6.7}$$

求 XFV 与最佳匹配代表性特征矢量序列 BFV_k 的差，形成一个特征差矢量序列 DFV_k，如下：

$$\begin{aligned} DFV_k:&XFV-BFV_k \\ &=\{\boldsymbol{DFV}_1^k,\boldsymbol{DFV}_2^k,\ \cdots\ ,\boldsymbol{DFV}_L^k\} \\ &=\{\boldsymbol{XFV}_1-\boldsymbol{BFV}_1^k,\boldsymbol{XFV}_2-\boldsymbol{BFV}_2^k,\ \cdots\ ,\boldsymbol{XFV}_L-\boldsymbol{BFV}_L^k\} \end{aligned} \tag{6.8}$$

特征差矢量序列 DFV_k 中的每个特征差矢量与输入信号特征矢量和最佳匹配代表性特征矢量之间存在关系：$\boldsymbol{DFV}^k=\boldsymbol{XFV}-\boldsymbol{BFV}^k$。由于特征差矢量 $\boldsymbol{DFV}^k$ 是两个随机特征矢量的差，统计分布上可以认为与 $\boldsymbol{BFV}^k$ 是近似独立的。互信息计算中的条件熵可以由下列公式推导得到。

$$\begin{aligned} &\boldsymbol{XFV}=\boldsymbol{DFV}^k+\boldsymbol{BFV}^k \\ &p(\boldsymbol{XFV}\mid\boldsymbol{BFV}^k)=p(\boldsymbol{DFV}^k+\boldsymbol{BFV}^k\mid\boldsymbol{BFV}^k)=p(\boldsymbol{DFV}^k) \\ &H(XFV\mid BFV_k)=-\iint p(\boldsymbol{BFV}^k)p(\boldsymbol{XFV}\mid\boldsymbol{BFV}^k)\log p(\boldsymbol{XFV}\mid\boldsymbol{BFV}^k)d\boldsymbol{XFV}d\boldsymbol{BFV}^k \\ &\qquad=-\int p(\boldsymbol{DFV}^k)\log p(\boldsymbol{DFV}^k)d\boldsymbol{DFV}^k \\ &\qquad=H(DFV_k) \end{aligned} \tag{6.9}$$

根据概率统计理论，当 $\boldsymbol{XFV}$ 和 $\boldsymbol{BFV}^k$ 采用相同的特征参数，并且具有高斯正态分布统计特征时，特征差矢量 $\boldsymbol{DFV}^k$ 也具有高斯正态分布统计特征，其均值和方差可以根据 $\boldsymbol{XFV}$ 和 $\boldsymbol{BFV}^k$ 的均值与方差计算得到，也可以直接根据其数据样本估计得到。线性预测系数 LPC 以及 LPCC、MFCC 等参数具有近似的高斯正态分布统计特征，因此，当采用这些特征参数时，特征差矢量 $\boldsymbol{DFV}^k$ 的统计分布特征可以由概率密度函数 $N(m_{dfv}^k,W_{dfv}^k)$ 表示，以上条件熵

$H(DFV_k)$及输入语音信号 XFV 与说话人模型 CFC_k之间的互信息可以计算如下：

$$\begin{aligned} H(DFV_k) &= \frac{p}{2}log(2\pi e)+\frac{1}{2}log|W_{dfv}^k| \\ H(XFV) &= \frac{p}{2}\log(2\pi e)+\frac{1}{2}\log|W_{xfv}| \\ I(XFV;BFV_k) &= H(XFV)-H(DFV_k) = \frac{1}{2}\log\frac{|W_{xfv}|}{|W_{dfv}^k|} \end{aligned} \tag{6.10}$$

上式中，p 是特征矢量的维数，W_{xfv}和W_{dfv}^k分别是输入语音信号特征矢量和特征差矢量的协方差矩阵，前者由于是模型无关的，因此在实际的识别判决中并不需要计算。W_{dfv}^k可以依据最大似然估值理论由实际的特征差矢量分布数据估计获得，其估值计算公式如下：

$$\begin{aligned} W_{dfv}^k &= \frac{1}{L}\sum_{j=1}^{L}(DFV_j^k - m_{dfv}^k)(DFV_j^k - m_{dfv}^k)^T \\ &= \frac{1}{L}\sum_{j=1}^{L}(XFV_j - BFV_j^k - m_{xfv} + m_{bfv}^k)(XFV_j - BFV_j^k - m_{xfv} + m_{bfv}^k)^T \end{aligned} \tag{6.11}$$

各特征矢量的均值$m_{dfv}^k, m_{xfv}, m_{bfv}^k$同样根据实际数据估计，但由于涉及统计估计的有效性，在实际应用中需要根据不同的情况作一定的变化。当输入信号的长度 L 较短时，采用相同的均值效果好些，而当 L 较大时，采用不同的均值效果更好，并且，最佳匹配代表性特征矢量的均值m_{bfv}^k由相应的模型 CFC_k直接估计效果更好。

6.2.3 最大互信息判决 MMI

输入语音与说话人模型之间的互信息 $I(XFV; CFC_k)$反映了两者的似然度，其值越大表示两者越相似。根据互信息评估算法 MMS，互信息 $I(XFV; CFC_k)$可以由输入信号与最佳匹配代表性特征矢量序列之间的互信息 $I(XFV; BFV_k)$来替代，并且，最终由特征差矢量序列的协方差矩阵W_{dfv}^k与输入语音信号特征矢量序列的协方差矩阵W_{xfv}来决定。最大互信息判决准则 MMI 如下：

$$d = \arg\max_k \; I(XFV;BFV_k) = \arg\max_k \; \frac{1}{2}\log\frac{|W_{xfv}|}{|W_{dfv}^k|} \tag{6.12}$$

在对所有的说话人模型进行匹配判决过程中，输入语音信号特征矢量序列的协方差矩阵是不变的，所以对判决并没影响，可以在具体计算中不加考虑。另外，考虑到对数函数的单调递增特点，可以免除对数计算，最终 MMI 准则简化如下：

$$d = Arg\min_{k} \ |W_{dfv}^{k}| \tag{6.13}$$

即识别说话人所对应的模型具有最小的特征差矢量序列协方差矩阵值。

6.3 实验分析与比较

基于全特征矢量集模型 CFC 与多级最小最大搜索匹配互信息评估算法 MMS 的文本无关说话人识别，需要研究分析的问题包括：（1）汉语说话人全特征矢量集模型 CFC 的大小，或 CFC 中需要多少代表性特征矢量才能充分表示说话人的语音特征？（2）训练一个说话人的全特征矢量集模型 CFC 一般需要多少语音数据？（3）互信息评估算法 MMS 的有效性或识别性能？（4）不同特征参数，例如 LPCC、MFCC 情况下，CFC-MMS 的说话人识别性能？（5）不同长度测试语音输入时，说话人识别性能的变化趋势？（6）在相同训练语音数据、实验环境和条件下， CFC-MMS 与 GMM 的识别性能比较分析。

6.3.1 实验数据、环境与条件

语音数据选择 SD2002-D2 数据库，该数据库中包含了在普通实验室环境下通过计算机声音系统采集得到的 40 个说话人的 280 条语音片段，其中，男声 26 人，女声 14 人，每人分别输入 7 段语音，如表 6.1 所示，每段语音包括停顿间隙的长度为 12 秒。语音采样率为 11025Hz，16 位量化，单声道输入。实验中，每说话人的前 4 段语音用于模型训练，后 3 段用于测试。

在模型训练和识别测试中，预处理部分首先消除输入语音信号的背景噪声，保留纯语音数据。短时分析采用 20ms 矩形窗。特征参数采用 12 阶 LPCC 和 MFCC 系数，其中，LPCC 由 12 阶 LPC 线性预测系数推导得到，MFCC 是基于 Mel 频率尺度的倒谱系数，通过计算 Mel 频率域均匀分布的 19 个三角滤波器组的 DFT 输出，并经 DCT 变换得到[22]，实验中选取第 1~12 个系数作为特征参数。

根据以往实验分析的结果，当训练语音数据达到 30 秒时模型性能最佳[33]。因此，在本文的实验中，说话人模型 CFC 均采用每说话人的前 4 段语音信号进行训练，其纯语音成分的长度平均为 32 秒。测试实验采用每说话人的后 3 段语音。实验中首先对互信息计算中特征差矢量序列的协方差矩阵计算采用统一均值和非统一均值的差别进行了分析，并在后续的实验中均采用统一均值方法处理，以便消除当测试语音长度较短时（小于 3 秒）均值估计误差带来的影响。

表 6.1　CFC 模型训练与测试语音样本

序号	文　　本
1	苏州，典型的江南水乡，园林众多，河流交错，素有东方威尼斯之称。北京，中国的首都，政治文化中心，高校林立，人才荟萃，清华大学、万里长城、中关村软件园、天安门就坐落在那里。
2	据信息产业部负责人透露，今年年底手机有望实现单向收费。从国家教育厅获悉，全国重点大学本科生部分专业课程将使用英文原版教材进行教学。没有什么比“吃”更令人心情愉快的了。
3	惟独让漂亮美眉兴趣大增的，当然是些模样精致的糕点，还有散发浓香的咖啡——制做这些东西的原料，多是清爽的鸡蛋、牛奶、苹果、草莓，再加上巧克力酱、干果等配料。
4	由刘德华、梁朝伟主演的《无间道》获香港金像奖最佳影片。张艺谋的《英雄》落选奥斯卡。通信原理、计算机网络、操作系统、程序设计、多媒体技术、程控交换、数字信号处理是电子工程学院的核心课程。
5	广州和深圳位于珠江三角洲，经济发达，人口密集。杭州，重要的旅游城市，历史悠久，四大爱情传说之一的许仙和白娘子的故事就发渊于此，著名景点有西湖、雷锋塔。与伦敦、巴黎、米兰、东京齐名。
6	武汉、上海，大连、南京、昆明，重庆、成都、兰州、无锡、天津、济南、合肥、青岛、石家庄、江阴、沈阳、西安、秦皇岛、辽宁、山东、浙江、温州。
7	广州、南宁、三亚、深圳、郑州、台湾、长春、江苏、四川、太原、山西、澳门、海口、南昌、厦门、福州、银川、哈尔滨、湖北、河南、海南、福建、安徽。

6.3.2 全特征矢量集大小分析

说话人的全特征矢量集模型 CFC 表达了说话人的语音特征，这种语音特征是语义和时序归一化的，因此体现了文本无关的特点。但是，对于汉语说话人来说，CFC 的大小应该确定为多少比较合适？如果 CFC 太小，其代表性特征矢量就不能完备地反映说话人的所有语音特征，反之，如果很大，不仅计算量增大，而且，各说话人模型出现相似代表性特征矢量的概率就增加，模型之间的耦合度将变大。因此，选择合适的 CFC 大小对模型的准确性和识别性能有较大影响。

实验中分别对代表性特征矢量数为 100、200、300、400、500 的五种不同大小的 CFC 的识别性能进行了分析，输入测试语音的长度分别为 1 秒和 2 秒，其具体测试中的误识率如图 6.1 所示。

从图 6.1 可以看到，如果采用统一均值，当测试语音长度为 1 秒时，随着模型中特征矢量数从 100 增加到 200，误识率从 8.8%下降到 6.63%，之后不断缓慢下降，但总体上较大；当测试语音长度为 2 秒时，误识率随模型大小的变化幅度不超过 1%，基本上是稳定的。如果采用非统一均值，两种测试语音长度下都在特征矢量数为 200 时得到最小误识率。因此，说话人全特征矢量集模型 CFC 选择 200 个代表性特征矢量较合适。

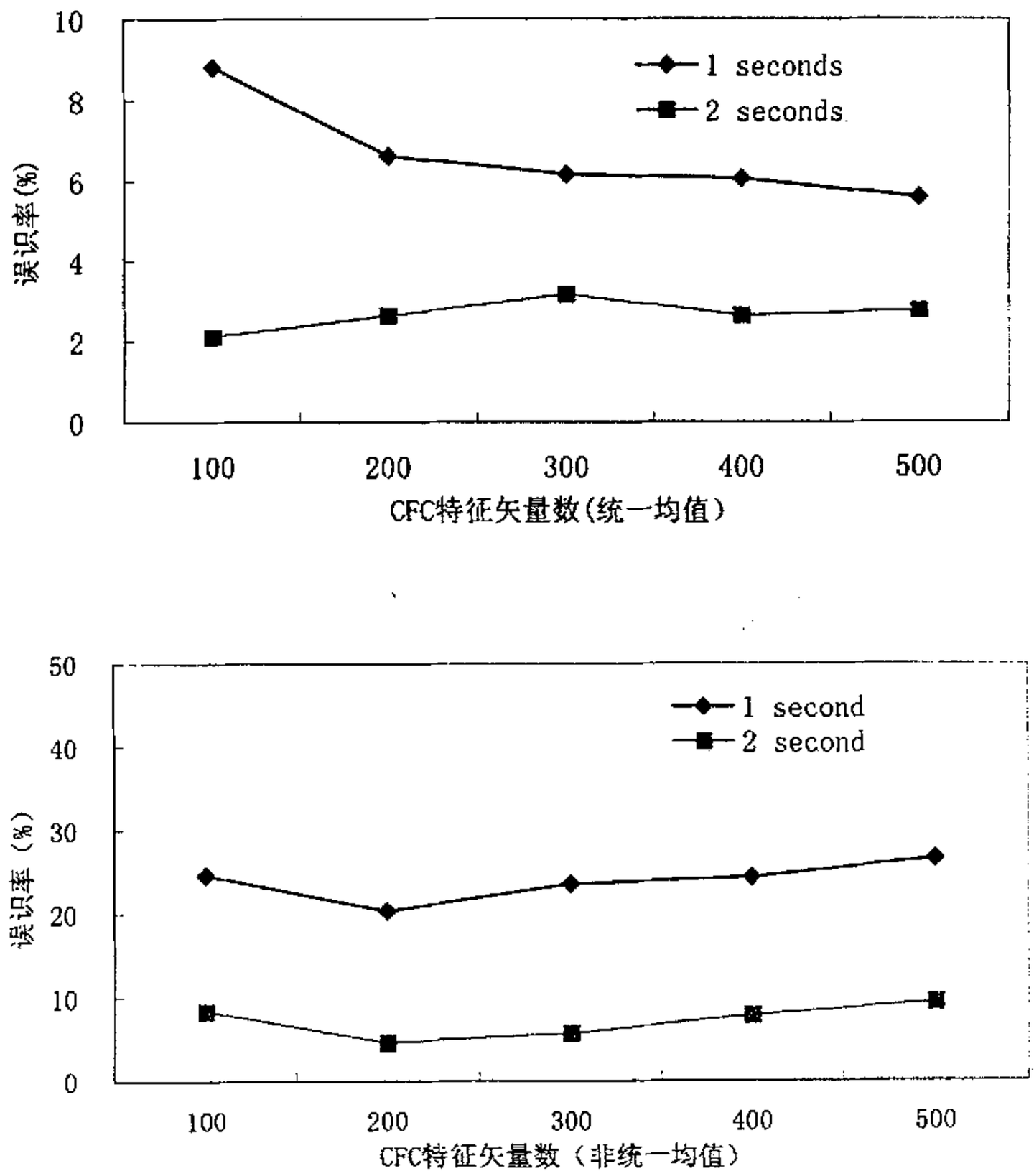


图 6.1　CFC 大小与识别性能的关系

6.3.3 CFC-MMS 的识别性能分析

根据以往的说话人识别研究表明，线性预测倒谱系数 LPCC 以及 Mel 频率倒谱系数 MFCC 是说话人识别中比较有效的两种特征参数，实验分析了当采用以上两种参数作为特征参数时 CFC-MMS 的说话人识别性能，具体结果如图 6.2 和图 6.3 所示，其中 CFC 的大小为 200，特征矢量维数为 12。

图 6.2 是在 12 阶 LPCC 特征参数情况下，采用统一均值（CFC_MMS_LPCC）和非统一均值(CFC_MMS_LPCC_M)得到的识别结果。可以看到，当测试语音时长分布为 1~3 秒时，由于无法从较短的语音信号中估计精确可靠的均值，采用统一均值的互信息计算方法识别性能更加优越。但当输入语音信号时长达到 4 秒以上时，此时采用非统一均值的互信息计算方法其识别性能比采用统一均值的方法要好，并在 5 秒时达到了 100%，这是因为当时长达到一定长度时，均值估计就比较精确可靠。从这一实验可以得到结论，在语音信号小于 4 秒时，互信息计算应该采用统一均值方法计算，而当超过 4 秒时，则应该采用非统一均值计算。同时，这一实验结果也说明了语音信号特征矢量的均值只有通过大于 4 秒的语音信号数据获得才是可靠、有意义的。

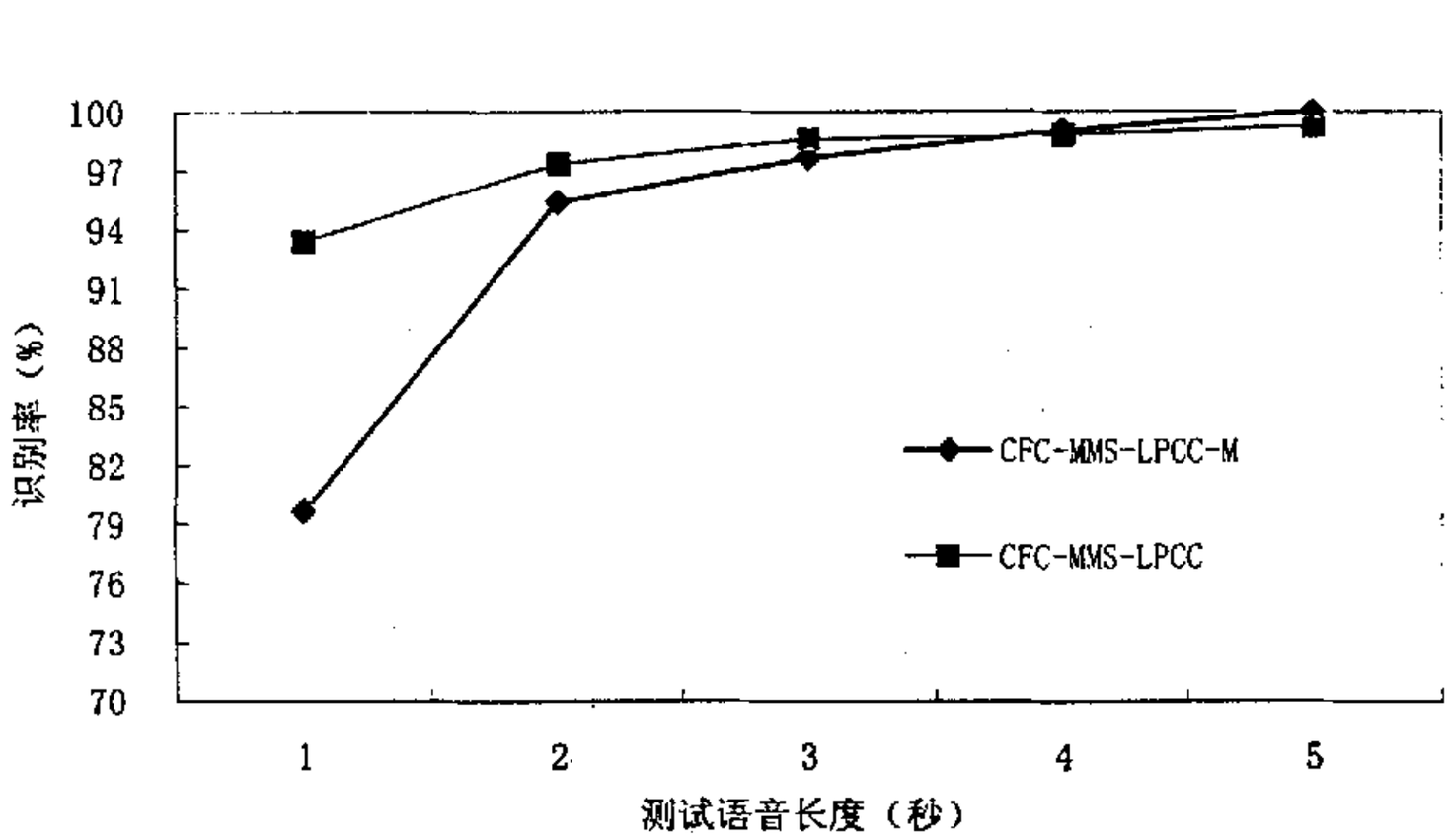


图 6.2　采用 LPCC 参数的识别性能（统一均值与非统一均值）

图 6.3 是采用统一均值情况下，分别采用 12 阶 MFCC 和 LPCC 特征参数时的识别性能比较。可以看出，总体上 CFC-MMS 具有很好的识别性能，并且，MFCC 作为特征参数比 LPCC 的性能更加优越。当输入测试语音为 1 秒时，尽管长度较短，但两种特征参数下识别率分别达到了 94.29%和 93.37%。当输入测试语音长度增加到 2 秒时，识别率分别提高到 99.07%和

97.35%，并且，当输入测试语音长度增加到 3 秒时，MFCC 特征参数情况下的识别率达到了 100%，LPCC 特征参数情况下的识别率也达到了 98.58%。

MFCC 参数无论在说话人识别应用中，还是在语音识别中都表现出比 LPCC 优越的识别性能，说明它在描述语音信号的语义和说话人个性特征两个方面都比 LPCC 更加有效。

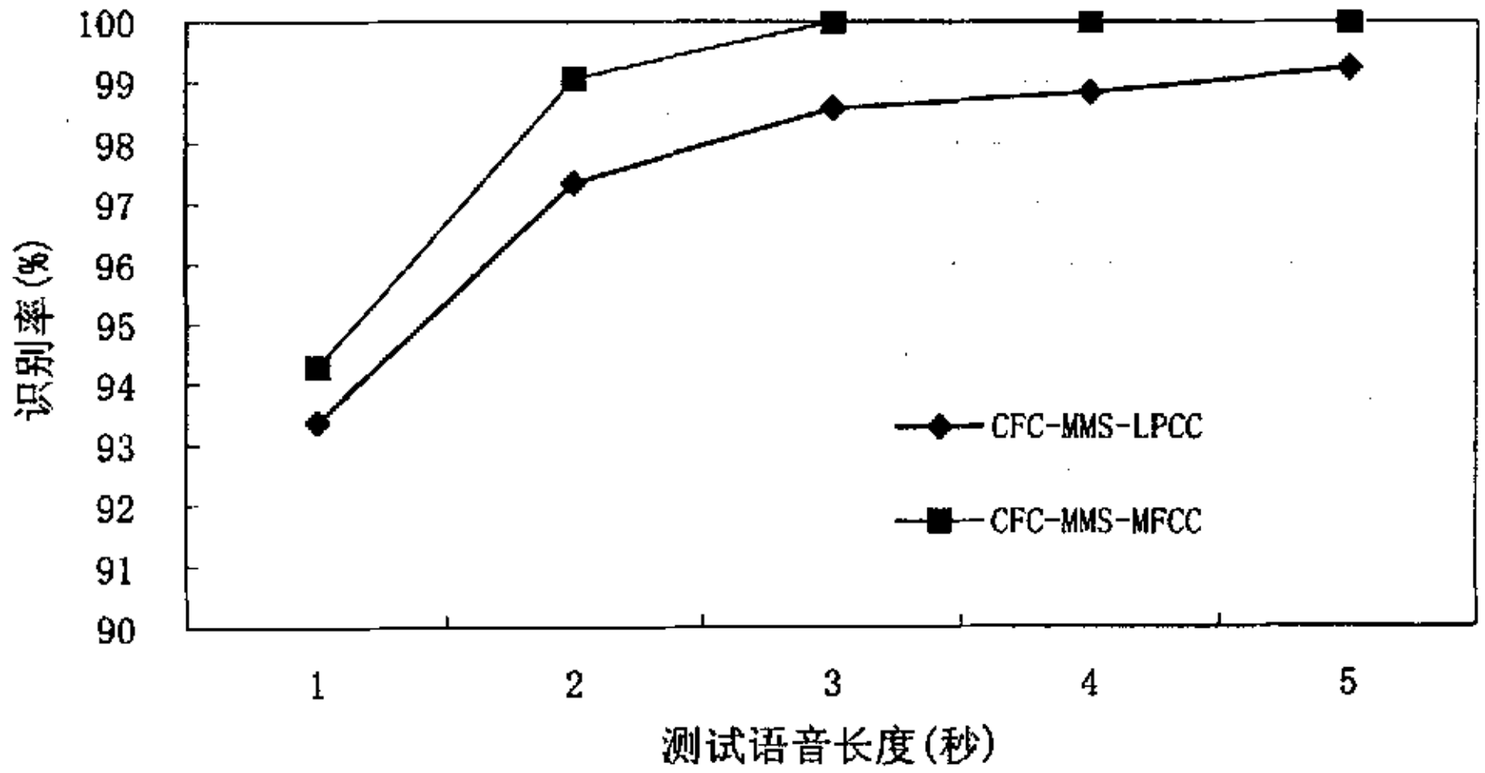


图 6.3　不同特征参数下识别性能比较

6.3.4　CFC-MMS 与 GMM 的识别性能比较

高斯混合模型 GMM 是目前文本无关说话人识别中应用较广、性能较好的一种模型，该模型用若干高斯正态分布概率密度函数的混合加权来拟合实际说话人语音信号频谱的统计分布，并利用 Bayes 最大似然准则进行判决。

比较分析中，GMM 所采用的训练和测试语音数据、特征矢量的计算方法与维数、实验条件与环境都与 CFC-MMS 一致，其混合分量数为 16。表 6.2 和图 6.4 表示了不同测试语音长度情况下 GMM 与 CFC-MMS 采用统一均值进行互信息匹配计算时的识别性能比较，其中，特征参数为 MFCC，CFC 的大小为 200。

表 6.2 各种测试语音长度下的识别率

Test Times	1	2	3	4	5
CFC-MMS-MFCC	94.29	99.07	100	100	100
GMM-MFCC	92.79	98.15	99.04	99.42	100

图 6.4 显示，在各种测试语音长度下，基于全特征矢量集模型 CFC 与互信息评估算法

MMS 的识别性能要优于 GMM 模型的识别性能，并在测试语音长度到达 5 秒时趋向一致。

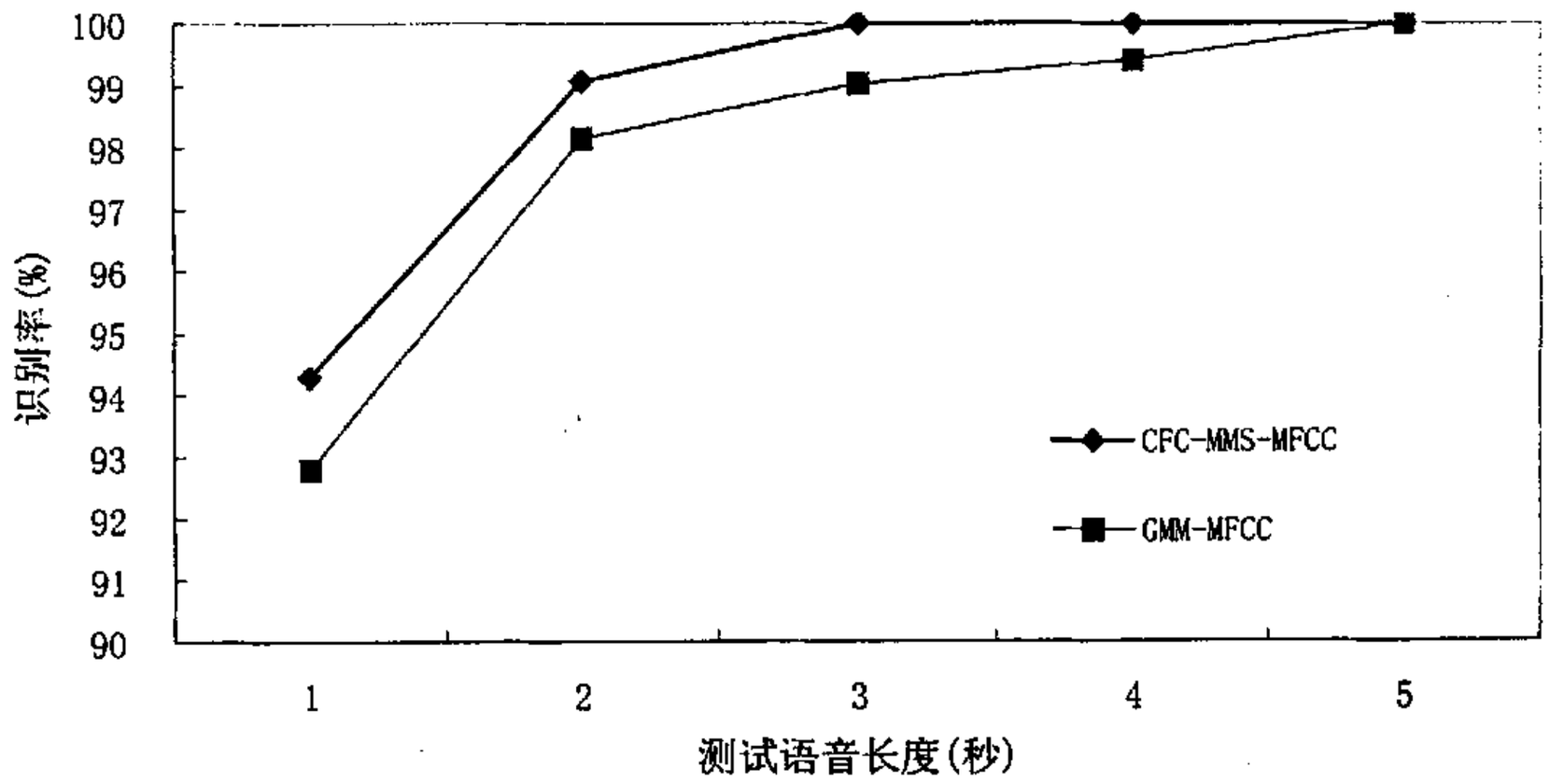


图 6.4　CFC-MMS 与 GMM 的识别性能比较

6.4 结　论

这一章提出了说话人的全特征矢量集模型 CFC 以及输入语音与说话人模型之间互信息匹配计算的多级最小最大搜索算法 MMS，并通过实验分析了基于 CFC-MMS 的文本无关说话人识别性能。实验表明：（1）CFC-MMS 对文本无关的说话人识别是有效的，并具有很高的说话人识别性能，当特征参数为 MFCC，输入测试语音长度为 3 秒时，识别率可以达到 100%；（2）相对 GMM 模型，CFC-MMS 的识别性能更加优越，在测试语音长度为 1~5 秒范围时识别率平均高出 1%左右，并在输入语音长度达到 5 秒时趋于一致；（3）对于 CFC-MMS，当特征参数采用 MFCC 时，其识别性能优于 LPCC，这与应用 GMM 等其它模型情况下得出的结论一致；（4）对于汉语说话人，全特征矢量集模型 CFC 的大小取 200 较合适，太小不能充分反映说话人的发音特征，太大则会出现特征信息的冗余。（5）互信息匹配计算时可以根据输入测试语音的长度自适应地选择采用统一均值或非统一均值进行，并以 4 秒为界。

CFC 的训练速度很快，同样训练数据下，比 GMM 的 EM 训练算法快许多，并且，对于语音信号来说，代表性特征矢量初始值的选取既可以按等间隔方式从原始特征矢量中选取，也可以以随机的方式选取，不同的选取方式对识别性能的影响很小。互信息匹配计算采用统一均值的好处是可以避免输入测试语音较短时均值估计误差引起识别性能的下降，例如在输入测试语音长度为 1 秒，特征参数采用 LPCC 时，采用统一均值和分别估计均值两种情况下

的识别率分别为 93.37%和 79.66%，说明这一处理是有效的。但是，当输入测试语音较长时，采用分别估计均值的方式或许更好些，例如，在输入语音长度为 5 秒的情况下，统一均值和分别估计均值两种情况下的识别率分别为 99.24%和 100%。

第七章　总结、讨论与展望

本章提要：

- 总结基于互信息理论的说话人识别研究结果
- 讨论自动说话人识别研究领域存在的核心问题及解决方法
- 今后的可开展的进一步研究：特征子空间分离，模型自适应

运用互信息理论对语音信号进行分析，揭示语音信号之间的信息相关程度，并将互信息理论进行推广、应用于说话人识别。本文解决了语音信号之间互信息的计算问题，对互信息作为一种失真测度的有效性进行了分析，针对基于文本和文本无关两种情况下的说话人识别提出了相应的说话人模型和互信息匹配算法。当然，本文研究的内容仅仅涉及到了说话人识别问题的某些方面，还有一系列的问题有待进一步作分析与研究。以下7.1节对本文的主要工作做一个总结，并对某些问题进行讨论。7.2节将讨论特征参数有效性评价问题，并简单介绍评价特征参数语义和说话人个性特征表达性能的4S方法。7.3节讨论语音特征子空间分离的问题，分析运用主成分分析PCA（Principal Component Analysis）方法和最小子空间熵MSE（Minimum Subspace Entropy）方法从语音特征空间分离语义特征子空间和说话人个性特征子空间的可行性。最后，7.4节讨论了说话人模型的自适应问题。

7.1　互信息理论的说话人识别应用

互信息理论使得自动说话人识别问题的研究有了一种区别于传统距离空间方法和统计方法的新思路。从信息量的角度考察分析语音模式之间的相似程度或特征相关性具有更加广泛的意义，可以证明，传统的方法是能够从信息论的理论角度得到解释和推导引出的[61]。但是，理论上的分析不完全代表实际应用的可行性，将互信息理论应用到说话人识别的一个首要问题是怎样计算语音信号或语音模式之间的互信息。

由于无法得到语音信号之间的条件概率分布特性，本文提出了随机干扰信号的概念来解释和描述语音信号之间的失真，并从随机信号的特征以及随机信号分析理论推导出这一信号的统计分布特性，最终，互信息的计算归结到该随机干扰信号的熵的计算并得到了解决。聚

类分析和识别实验结果都表明了互信息理论应用在说话人识别中是有效的，并且具有很好的模式分类特性和较高的识别性能。在基于文本和文本无关的说话人识别中都显示出比流行的GMM方法更加优越的识别性能。

本文提出了语音信号之间互信息计算的线性映射匹配算法LPM和非线性搜索匹配算法NLM。LPM算法通过线性映射将语音信号特征矢量序列规整到相同的时域进行互信息计算，而非线性搜索匹配则通过动态规划方法将两个语音信号特征序列以非线性的方式进行互信息匹配计算。NLM算法相对LPM的优点是考虑了对信号时域非线性波动的处理，但缺点是增加了运算复杂度。

互信息作为一种模式失真测度与传统的Euclidean、Mahalanobis测度和 Itakura-Saito测度等相比具有一定的优越性。并且，互信息的计算是基于模式而非短时帧，很好地融合了信号统计特征和时变特征的处理。但是，本文提出的方法也有一定的前提，即信号本身或相应的特征参数必须是高斯（Gaussian）分布。幸运的是语音信号的LPC、LPCC和MFCC等大部分特征参数都具有近似的高斯分布统计特征。对于其它的统计分布情况下如何计算互信息有待今后进一步研究，一个可以考虑的方法是用混合高斯分布GM(Gaussian Mixture)去逼近实际统计分布，并依据GM分布计算互信息。

基于文本的说话人识别在许多安全认证系统中有重要的应用价值。本文针对这一领域的说话人识别应用提出了说话人多模板模型MTM和非线性搜索匹配相结合的系统解决方案，并通过实验证明了这一方案的有效性以及与GMM和DTW方法相比在识别性能上的优越性。在这个实验分析的最初阶段曾经考虑单一模板说话人模型的方案，后来发现这样的误识率较高，当将模型用MTM替代后，识别性能明显提高了。但是，由于训练数据的原因，没有对模板个数的优化进行分析。

文本无关的说话人识别在电子监听、司法鉴定方面有重要的应用价值，相对基于文本的说话人识别来说更具有挑战性。本文提出了说话人的全特征矢量集模型CFC和多级最小最大搜索（MMS）互信息匹配相结合的系统解决方案。CFC模型的基本概念是，在无法从语音信号特征中分离语义和说话人个性特征的前提下，说话人模型应该全面地充分反映说话人所发各种语音的特点。虽然可以设计特定的训练样本数据来训练CFC模型，但对于实际应用系统来说，随机选取训练样本数据也许更加现实。为了研究训练一个说话人模型需要多少语音样本数据，曾通过基于GMM的文本无关说话人识别对此进行了分析[33]，发现随机选取的30秒语音是一个很充分的量。另一个与CFC模型有关的问题是特征矢量集大小的问题，本文的实验表明，200

个特征矢量对于CFC是一个合适的数字。多级最小最大搜索匹配算法MMS首先从CFC中提取与输入语音特征矢量距离最小的CFC特征矢量，并进一步采用线性映射匹配的方式计算互信息，依据最大互信息准则进行识别判决。MMS算法同时运用了距离空间和互信息意义下的最佳匹配思想，语音信号的统计特征和时变特征都得到了处理。

有关噪声环境下的识别实验有待进一步研究。例如，可以考虑应用小波变换提取具有鲁棒性的特征参数[72,115]，或采用噪声补偿[126]的处理方法消除噪声。从理论上分析，本文提出的系统方案中，由于互信息计算运用了语音信号的统计分布特征，因此，与基于距离空间的方法相比会表现出一定的抗噪声特点，特别是高斯白噪声的干扰，但需要实验进一步证明。

互信息理论在模式识别和语音处理中的应用已经有一些研究人员做了工作，但提出直接计算语音信号之间互信息计算的线性和非线性搜索匹配算法是本文的主要贡献，另外，将互信息理论应用到说话人识别中也是本文的开创性工作之一[127]。实际上，本文的思路以及所提出的互信息计算方法可以推广到目标识别、图象识别等许多应用领域。

7.2 特征参数的有效性分析

目前，在说话人识别中采用的特征参数几乎与语音识别中所采用的特征参数一样，还没有一种有效的仅仅表达说话人个性特征的参数，常用的特征参数有LPCC、MFCC等。而实际上，特征参数对于识别性能起着重要的作用，因此，对特征参数在不同应用领域的有效性进行分析评估很有必要。但到目前为止，还没有一套评价方法能够对特征参数的语义和说话人个性特征的表达能力进行分析，大部分情况下通过实际的识别实验进行分析说明。但识别实验只能反映特征参数在某一方面的有效性，并不能说明特征参数在描述语义信息和说话人个性特征信息方面的优劣以及它们的比例等。

在这里介绍一种作者提出的4S（2-Speech and 2-Speaker）参数评价方法，该方法可以定量地对特征参数表达语义信息和说话人个性特征信息的比例进行分析，其示意图如图7.1所示。

在相同模型结构下，对说话人A的语音A建立模型A，并对说话人B的语音B以相同的方法建立模型B。然后，分别对说话人A的语音B和说话人B的语音A进行测试，计算与两个模型之间的距离D1，D2，D3，D4。这里所谓语音A和语音B是指对应的单词文本分别为文本A和文本B。各距离反映的信息如图所示，四个距离中两个表示语义差，两个表示个性特征差，距离值越大说明特征参数能够更好地反映相应信息。判决规则为：

$$\frac{D1+D4}{D2+D3}=\lambda=\begin{cases}<1 & \text{特征参数较多地包含了个性特征信息}\\ >1 & \text{特征参数较多地包含了语义信息}\\ =1 & \text{特征参数不能很好地区分语义和个性特征信息}\end{cases}$$

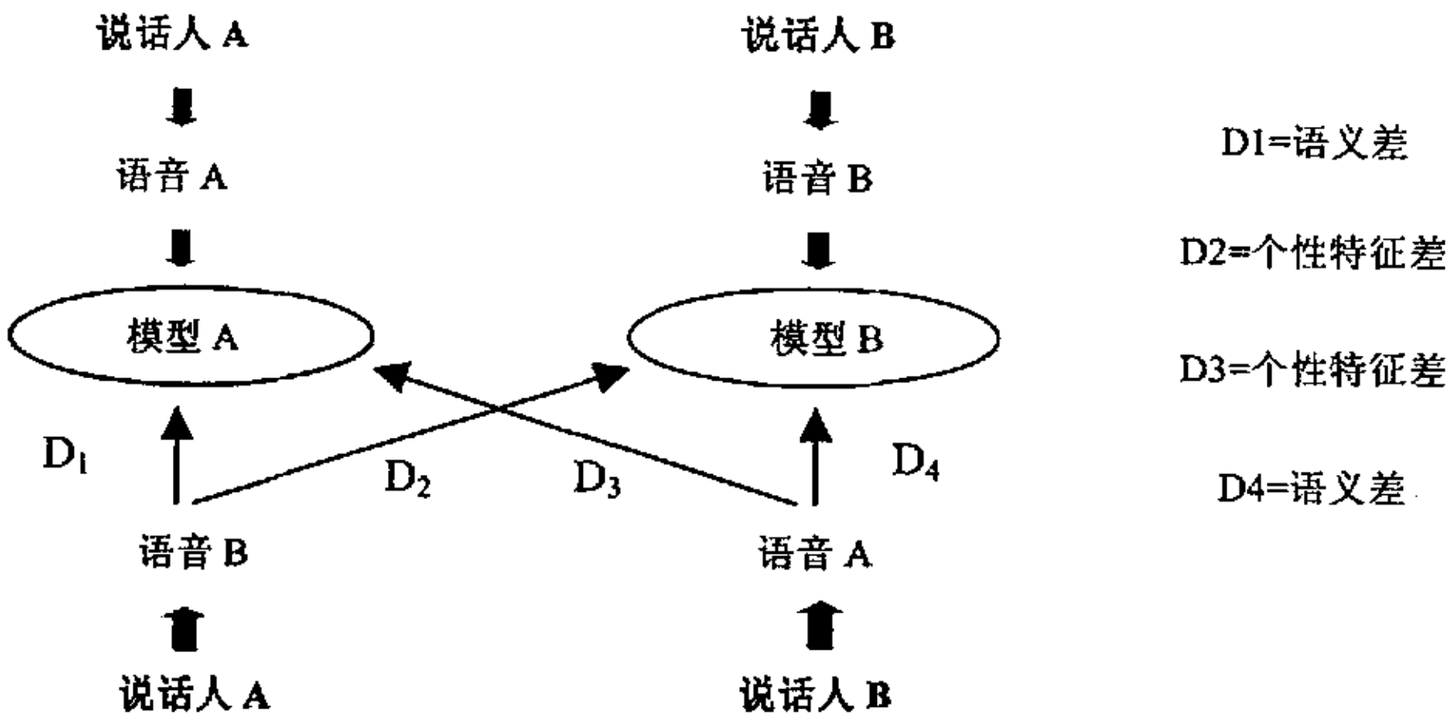


图7.1　特征参数评价的4S方法

比值λ就是特征参数在一定模型和匹配距离尺度下所能表达的语义信息和说话人信息的比例。可以选择一组说话人和语音样本数据对特征参数应用4S方法进行测试，并对距离进行统计分析后运用以上规则进行判决。

对常用的特征参数LPC、LPCC和MFCC进行的评价表明：这些参数表达语义信息比说话人个性特征信息更加充分。并且，三种特征参数中，MFCC的λ值最大。另外，对说话人A为男性，说话人B为女性的男女成对情况进行的评价表明：各种参数下λ值仍然大于1，但比一般情况下小，说明了此时特征参数中的说话人个性特征成分增强了。

7.3　说话人特征子空间分离

语音是一种复杂的非平稳随机信号。语音信号特征中既包含有语义特征，又包含说话人的个性特征，两种特征以复杂的形式相互交织在一起，并通过信号的声学特性反映出来。目前在说话人识别中主要采用的特征参数有线性预测倒谱系数 LPCC、反映人类听觉特性的 Mel 尺度谱系数 MFCC、线谱对系数 LSP、动态特征系数Δ、Δ^2 以及韵律特征参数基音等。这些参数同时表达了语义和说话人的特征信息，无法实现两者的分离描述。因此，在进行说话人识别时，语义特征对说话人个性特征的干扰甚至淹没将对识别性能产生很大的影响。例如，LPCC 参数被认为较好地反映了说话人的声道结构特性，当然，一个说话人本身具有区别于其他人的固有声道结构，但是，当发不同的语音时这个声道结构将会有不同的改变，这种改变导致不同说话人声道结构之间的差异减少，甚至混淆，而 LPCC 参数是通过测量语音信号获得的，它反映的是改变后的声道结构。

影响说话人识别性能的主要因素是由语义的变化和发音时间的变化引起的信号声学特性变化。由于当前采用的特征参数和模型主要反映说话人语音信号的声学特性，并无法对语音特征中所包含的语义和说话人特征进行分离，因此，这些变化使得输入语音的特征和相应说话人模型得不到很好的匹配，系统的识别性能将大幅下降。虽然可以采用特征参数归一化和似然度归一化的方法来处理这些变化，但效果是有限的。如果能够找到一种方法将语音特征中的语义特征和说话人个性特征进行分离，或强化说话人个性特征，则语义变化的影响就可以消除或抑制，发音时间变化的影响可以减少。另外，依据说话人个性特征建立的说话人模型将变得更加可靠和具有鲁棒性，而且仅仅需要少量的训练数据，系统的自适应学习也变得容易实现。当然，要实现语义特征与说话人个性特征的完全分离并非容易，但可以尝试利用主元分析（PCA: Principal Component Analysis）方法和最小子空间熵（MSE: Minimum Subspace Entropy）原理在语音特征空间中进行子空间分离来研究这一问题。

从语音特征空间来看，每一个说话人的语音特征分布在一个特定的区域范围内。从语音特征包含语义特征和说话人个性特征的实际情况分析，由于说话人个性特征相对稳定、变化小，而语义特征非稳定、变化大。因此，说话人语音特征的分布区域应该是以说话人个性特征为中心，语义特征的变化引起的分散而形成。可以利用主元分析方法（PCA）对说话人的语音特征进行分析，并在其特征分布区域构成说话人语音特征子空间，进一步对这一子空间进行分析，依据最小子空间熵（MSE）原理分离出说话人个性特征子空间，应用于说话人识别。

7.4 说话人模型的自适应

说话人模型描述和反映说话人的语音特征，目前主要采用的模型为（1）非参数模型（template model),如 VQ 码本模型 CBM(Code Book Model)；（2)参数模型(parameter model)，如高斯混合模型 GMM（Gaussion Mixture Model、神经网络模型（Artificial Neural Network）和支持向量机模型 SVM（Support Vector Machine)。非参数模型以特征矢量集的形式直接描述说话人语音特征，模型训练较简单，对数据量要求不高，但没有考虑语音信号统计特征的利用；参数模型以一定结构的分布描述说话人语音特征，模型参数的训练利用了语音的统计特征，但对训练数据量有较高要求，训练时间较长，并且会产生欠学习和过学习的问题。由于这些模型的建立总体上都是采用数据挖掘技术和统计学习技术实现，模型的可靠性与鲁棒性与训练的数据量和分布有很大关系，数据的不足和分布不合理会造成建立的模型缺乏可靠性和鲁棒性。事实上，一个模型在训练时是不可能将系统应用时的所有情况都考虑到的。特别

是，模型训练的基本数据元素是语音的特征参数，在不能实行说话人个性特征与语义特征分离的前提下，建立的说话人模型实际上只是说话人的语音特征模型，虽然通过语义归一化手段减少了语义特征的影响，但无法有效消除，客观上语义特征的存在仍然会增加模型之间的耦合度，影响模型的可靠性和鲁棒性，导致识别性能的下降。

说话人本身的个性特征也是时变的，即便是同一个语音，在其它条件不变的前提下，不同时刻发音的声学特性表现并不完全一致。引起这些现象的主要原因是说话人的生理、心理以及情绪状态等发生了变化。因此，一个说话人识别系统需要能够根据说话人个性特征的变化而自动学习新的特征、更新说话人模型，使模型自动适应新的特征，保持良好的识别性能。虽然已经提出了监督学习自适应和无监督学习自适应方法，但由于目前的特征参数无法实现说话人个性特征的分离描述，因此这些方法只能捕捉到语音特征的变化，自适应学习也仅仅针对语音特征进行，无法很好地实现对说话人个性特征的自适应学习。

参考文献

[1] S.Prabhakar, S.Pankanti, and A.Jain, "Biometric recognition: security and privacy concerns", *IEEE Security & Privacy Magazine,* 1, 2003, pp33–42.

[2] http://www.biometrics.org/. the Biometric Consortium. WWWpage, December 2003.

[3] A.K. Jain, R. Bolle and S. Pankanti, "BIOMETRICS: Personal identification in networked society", Kluwer Academic Publishers, 1999.

[4] J.Campbell, "Speaker recognition: a tutorial", *Proceedings of the IEEE,* 85(9):1437–1462,1997.

[5] S.Furui, "Recent advances in speaker recognition", *Pattern Recognition Letters,* 18, 1997, pp.859–872.

[6] G.Doddington, "Speaker recognition - identifying people by their voices", *Proceedings of the IEEE,73(* 11):1651–1164,1985.

[7] S.Furui, "Digital Speech Processing, Synthesis, and Recognition", second ed. *Marcel Dekker, Inc.*, New York, 2001.

[8] Q.Li, B.-H.Juang, and C.-H.Lee, "Automatic verbal information verification for user authentication", *IEEE Trans. on Speech and Audio Processing,* 8(5):585–596,2000.

[9] A.Schmidt-Nielsen, and T. Crystal, "Speaker verification by human listeners: experiments comparing human and machine performance using the nist 1998 speaker evaluation data", *Digital Signal Processing,* 10 , 2000, pp.249–266.

[10] J.Kerstholt, E.Jansen, A.Amelsvoort, and A.Broeders, "Earwitness line-ups: e®ects of speech duration, retention interval and acoustic environment on identification accuracy", *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, Geneva, Switzerland, 2003, pp.709–712.

[11] L.Liu, J.He, and G. Palm, "A comparison of human and machine in speaker recognition", *Proc. 5th European Conference on Speech Communication and Technology(Eurospeech 1997)* , Rhodos, Greece, 1997, pp. 2327–2330.

[12] K.Sullivan, and J. Pelecanos, "Revisiting Carl Bildt's impostor: would a speaker verification system foil him?", *Proc. Audio- and Video-Based Biometric Authentication (AVBPA 2001),* Halmstad, Sweden, 2001, pp. 144–149.

[13] http://www.nist.org/speech. Speaker system evaluation results of National Institute of Speech Technology.

[14] G.Doddington, "Speaker recognition based on idiolectal di®erences between speakers", *Proc. 7th European Conference on Speech Communication and Technology(Eurospeech 2001)* , Aalborg, Denmark, 2001, pp. 2521–2524.

[15] B.Xiang, "Text-independent speaker verification with dynamic trajectory model", *IEEE Signal Processing Letters,* 10 , 2003, pp.141–143.

[16] D.Reynolds, W.Andrews, J.Campbell, etc., "The SuperSID project: exploiting high-level information for high-accuracy speaker recognition", *Proc. Int. Conf. on Acoustics,Speech, and Signal Processing (ICASSP 2003)*, Hong Kong, 2003, pp.784–787.

[17] J.Campbell, D.Reynolds, and R.Dunn, "Fusing high- and low-level features for speaker recognition. In *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003)* , Geneva, Switzerland, 2003, pp. 2665–2668.

[18] P.Rose, "Forensic Speaker Identification", *Taylor & Francis*, London, 2002.

[19] H.Sako, S.Chiba, "Dynamic programming optimization for spoken word recognition", *IEEE Trans. On ASSP*, 26(10):43~49,1978.

[20] L.R.Rabiner, S.E.levinson, "On the application of vector quantization and hidden Markov models to speaker independent isolated word recognition", *Bell Syst. Tech.* 12(4):1035~1074,1983.

[21] L.Rabinar, B.H.Juang, S.E.levison, "Recognition of isolated digits using hidden Markov models with continuous mixture density", *AT&T Tech.J.*, 64(6):1211-1234, 1985.

[22] L.Rabiner, and B.-H.Juang, "Fundamentals of Speech Recognition", *Prentice-Hall*, New Jersey, U.S.A, 1993.

[23] 俞一彪，袁保宗, "连续语音识别中句法结构知识的利用", *电子学报*，18(6):68~74, 1990

[24] R.Kuhn, J.-C.Junqua, P.Nguyen, etc. , "Rapid speaker adaptation in eigenvoice space", *IEEE Trans. on Speech and Audio Processing*, 8(6):695–707,2000.

[25] A.Martin, and M.Przybocki, "Speaker recognition in a multi-speaker environment", *Proc. 7th European Conference on Speech Communication and Technology(Eurospeech 2001)* , Aalborg, Denmark, 2001, pp. 787–790.

[26] I.Lapidot, H.Guterman, and A.Cohen, "Unsupervised speaker recognition based on competition between self-organizing maps", *IEEE Transactions on Neural Networks*, 13(8): 877–887,2002.

[27] S.Kwon, and S.Narayanan, "Speaker change detection using a new weighted distance measure", *Proc. Int. Conf. on Spoken Language Processing (ICSLP 2002)*, Denver, Colorado, USA, 2002, pp. 2537–2540.

[28] D.Liu, and F.Kubala, "Fast speaker change detection for broadcast news transcription and indexing", *Proc. 6th European Conference on Speech Communication and Technology (Eurospeech 1999)* , Budapest, Hungary, 1999, pp. 1031–1034.

[29] R.Brunelli, and D. Falavigna, "Person identification using multiple cues", *IEEE Trans. on Pattern Analysis and Machine Intelligence*,17(10):955–966,1995.

[30] J.Kittler, and M.Nixon, Proc. 4th Internation Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA 2003), Lecture Notes in Computer Science, Springer-Verlag, Berlin, 2003.

[31] E.Zetterholm, "The significance of phonetics in voice imitation", *Proc. 8th Australian Int. Conf. on Speech Science and Technology*, 2000, pp. 342–347.

[32] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, "Clustering speakers by their voices", *Proc. International Conference on Acoustic Speech and Signal Processing*, 2, 1998, pp.557-560.

[33] 许允喜,俞一彪, "基于 GMM 的汉语说话人识别特性分析", *通信技术*，2，2004，pp.120-123

[34] G.Doddington, W Liggett, A.Martin, etc., "Sheeps, goats, lambs and wolves: a statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation", *Proc. Int. Conf. on Spoken Language Processing (ICSLP 1998)* , Sydney, Australia, 1998.

[35] W.Majewski , and G.Mazur-Majewska, "Speech signal parametrization for speaker recognition under voice disguise conditions", *Proc. 6th European Conference on Speech Communication and Technology (Eurospeech 1999)* , Budapest,Hungary, 1999, pp. 1227–1230.

[36] D.Genoud, and G.Chollet, "Speech pre-processing against intentional imposture in speaker recognition", *Proc. Int. Conf. on Spoken Language Processing(ICSLP 1998)* ,Sydney, Australia, 1998.

[37] D.van Leeuwen, "Speaker verification systems and security considerations", *Proc.8th European Conference on Speech Communication and Technology (Eurospeech2003)*, Geneva, Switzerland, 2003, pp. 1661–1664.

[38] X.Huang, A.Acero, and H.-W.Hon, "Spoken Language Processing: a Guide to Theory, Algorithm, and System Developmen", Prentice-Hall, New Jersey, U.S.A, 2001.

[39] T.Quatieri, D.Reynolds, and G. O'Leary, "Estimation of handset nonlinearity with application to speaker recognition", *IEEE Trans. on Speech and Audio Processing* , 8(5):567–584, 2000.

[40] M.Phythian, J.Ingram, and S.Sridharan, "Effects of speech coding on text-dependent speaker recognition", Proceedings of IEEE Region 10 Annual Conference on Speech and Image Technologies for Computing and Telecommunications(TENCON'97) , 1997, pp.137–140.

[41] L.Besacier, S.Grassi, A.Dufaux, etc., "GSM speech coding and speaker recognition", *Proc. Int. Conf. on Acoustics,Speech, and Signal Processing (ICASSP 2000)*, Istanbul, Turkey, 2000, pp.1085–1088.

[42] X.Li, M.Mak, and S.Kung, "Robust speaker verification over the telephone by feature recuperation", *Proc. 2001 Int. Symposium on Intelligent Multimedia,Video, and Speech Processing*, Hong Kong, 2001, pp. 433–436.

[43] R.Mammone, X.Zhang, and R.Ramachandran, "Robust speaker recognition: a feature based approach", *IEEE Signal Processing Magazine*, 13(5):58–71,1996.

[44] M.Zilovic, R.Ramachandran, and R.Mammone, "Speaker identification based on the use of robust cepstral features obtained from pole-zero transfer functions", *IEEE Trans. on Speech and Audio Processing*, 6(3):260–267, 1998.

[45] J.Ortega-Garc_ia, , and J. Gonz_alez-Rodr_iguez, "Overview of speech enhancement techniques for automatic speaker recognition", *Proc. Int. Conf. on Spoken Language Processing (ICSLP 1996)*, Philadelphia, Pennsylvania, USA, 1996, pp. 929–932.

[46] D. Reynolds, "An overview of automatic speaker recognition technology", *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2002)* , Orlando, Florida, USA, 2002, pp. 4072–4075.

[47] F.Soong, and A.Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition", *IEEE Trans. on Acoustics, Speech and Signal Processing*, 36(6): 871–879, 1988.

[48] J. Hansen, and J.Proakis, "*Discrete-Time Processing of Speech Signals*", second ed, *IEEE Press*, New York, 2000.

[49] J. M.Naik, "Speaker Verification: A Tutorial", *IEEE Communications Magazine*, January 1990, pp.42-48.

[50] C.Miyajima,Y.hattori, "Text-independent seaker identification using Gaussion mixture model based on multi-space probability distribution", *IEICE Tans. INF.&SYST.*, 84(7):847-855,2001

[51] R. Auckenthaler, M.Carey , and H.Lloyd-Thomas, "Score normalization for text-independent speaker verification systems", *Digital Signal Processing*, 10(1):42–54, 2000.

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

[53] 岳喜才,伍晓宇,郑崇勋,"用神经阵列网络进行文本无关的说话人识别"，*声学学报*，25(3):230-234, 2000.

[54] 侯风雷,王炳锡, "基于支持向量机的说话人辨认研究" , *通信学报*, 23(6):61-67, 2002.

[55] L. Rabinar, B.H. Juang, "Fundamentals of Speech Recognition", *Prentice Hall*, pp.141-195, 1993.

[56] M.Nishida,Y.Ariki, "Speaker recognition by projecting to speaker space with less phonetic information", *IEICE Tans. INF.&SYST*, 85(4):554-562, 2002.

[57] H.Ney, "The use of one-stage dynamic programming algorithm for connected word recognition", *IEEE Trans. On ASSP*, 32(2):263-271, 1984.

[58] J.K.Flanagan,D.R.Morell, "Vector quantization codebook generation using simulated annealing", *Proc. of ICASSP*, 1989, pp.1759-1762

[59] D.Reynolds, and R.Rose, "Robust text-independent speaker identification using gaussian mixture speaker models", *IEEE Trans. on Speech and Audio Processing*, 3(1): 72–83, 1995.

[60] K.Farrell, R.Mammone, and K.Assaleh, "Speaker recognition using neural networks and conventional classifiers", *IEEE Trans. on Speech and Audio Processing*, 2(1):194–205, 1994.

[61] Y.-T. Lee, "Information-theoretic distortion measures for speech recognition", *IEEE Trans. On ASSP*, 39(3):330-335, 1991.

[62] S. Okawa, T.Kobayashi, K. Shirai, "Automatic training of phoneme dictionary based on mutual information criterion" , *Proc. Of ICASSP*, 1994, pp.241-244

[63] S. Okawa, "A recombination strategy for multi-band automatic speech recognition based on mutual information criterion", Research Report of Chiba University, 47, 2000, pp.141-147

[64] L.R.Bahl, P.F.Brown,etc., "Maximum mutual information estimation of hidden Markov model parameters for speech recognition", *Proc. Of ICASSP*, Tokyo, 1986, pp.49-52.

[65] E.Herakao, "Deriving multiplication-type FVQ/HMM from a new information-source model and a viewpoint of the information theory, and showing the relation of the model to discrete and continuous HMMs", *IEICE Tans. INF.&SYST*, 84(12):2501-2506, 2001.

[66] S.Nakagawa, "A survey on automatic speech recognition", *IEICE Tans. INF.&SYST*, 85(3):465-486, 2002.

[67] R.Ramachandran, K.Farrell, R.Ramachandran, and R.Mammone, "Speaker recognition - general classifier approaches and data fusion methods", *PatternRecognition, 35*, 2002, pp.2801–2821.

[68] H.Gish, and M.Schmidt, "Text-independent speaker identification", *IEEE Signal Processing Magazine*, 11, 1994, pp.18–32.

[69] M.Huggins, and J.Grieco, "Confidence metrics for speaker identification", *Proc. Int. Conf. on Spoken Language Processing (ICSLP 2002)*, Denver, Colorado, USA, 2002, pp. 1381–1384.

[70] A.Jain, R.P.W.Duin, and J.Mao, "Statistical pattern recognition: a review", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(1): 4–37, 2000.

[71] A.Jain, and D.Zongker, "Feature selection: evaluation, application, and small sample performance", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19, 1997, pp.153–158.

[72] 芮贤义，俞一彪, "基于小波变换的鲁棒型特征提取及说话人识别"，*电路与系统学报*(待发表), 2005.

[73] K.Fukunaga, "Introduction to Statistical Pattern Recognition", second ed. *Academic Press*, London, 1990.

[74] R.Duda, P.Hart, and D.Stork, "Pattern Classification", second ed. *Wiley Interscience*, New York, 2000.

[75] A.Hyvarinen, J.Karhunen, , and E.Oja, "Independent Component Analysis", *John Wiley & Sons, Inc.*, New York, 2001.

[76] 边肇祺，张学工，“模式识别”，清华大学出版社，2000.

[77] F.Soong, B.-H.Juang, and L.Rabiner, “A vector quantization approach to speaker recognition”. *AT & T Technical Journal*, 66(1): 14–26, 1987.

[78] Y. Linde, A. Buzo, R. Gray, “An algorithm for Vector Quantizer Design”, *IEEE Trans. on Communications*, 28 (1): 84-95, January 1980.

[79] A.Gersho, and R. Gray, “Vector Quantization and Signal Compression”, *Kluwer Academic Publishers*, Boston, 1991.

[80] J.He, L.Liu, and G. Palm, “A discriminative training algorithm for VQ-based speaker identification”, *IEEE Trans. on Speech and Audio Processing*, 7(3):353–356, 1999.

[81] T.Kinnunen, T.Kilpelainen, and P.Franti, “Comparison of clustering algorithms in speaker identification”, *Proc. IASTED Int. Conf. Signal Processing and Communications (SPC 2000)*, Marbella, Spain, 2000, pp. 222–227.

[82] T.Kinnunen, I.Karkkainen, and P.Franti, “Is speech data clustered?- statistical analysis of cepstral features”, *Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001)* , Aalborg, Denmark, 2001, pp. 2627–2630.

[83] 俞一彪，王朔中，“语音识别互信息测度的聚类特性与实验评价”，*信号处理*, 18(5), 2002, pp.24-27

[84] R.-H.Wang, L.-S.He, and H. Fujisaki, “A weighted distance measure based on the fine structure of feature space: application to speaker recognition”, *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 1990)*, Albuquerque, New Mexico, USA, 1990, pp. 273–276.

[85] G.Kolano, and P. Regel-Brietzmann, “Combination of vector quantization and gaussian mixture models for speaker verification”, *Proc. 6th European Conference on Speech Communication and Technology (Eurospeech 1999)*, Budapest, Hungary, 1999, pp. 1203–1206.

[86] J.Pelecanos, S.Myers, S.Sridharan, and V.Chandran, “Vector quantization based gaussian mixture modeling for speaker verification”, *Proc. Int. Conf. on Pattern Recognition (ICPR 2000)*, Barcelona, Spain, 2000, pp. 3298–3301.

[87] G.Singh, A.Panda, S.Bhattacharyya, and T.Srikanthan, “Vector quantization techniques for GMM based speaker verification”, *Proc. Int. Conf. On Acoustics, Speech, and Signal Processing (ICASSP 2003)* , Hong Kong, 2003.

[88] F.Bimbot, M.Blomberg, L.Boves, D.Genoud, etc., “An overview of the CAVE project research activities in speaker verification”, *Speech Communications*, 31(2): 155–180, 2000.

[89] S.Mallat, “A Wavelet Tour of Signal Processing”, *Academic Press*, New York, 1999.

[90] J.Proakis, and D.Manolakis, “Digital Signal Prosessing. Principles, Algorithms and Applications”, second ed. Macmillan Publishing Company, New York, 1992.

[91] J.Harrington, and S. Cassidy, “Techniques in Speech Acoustics”, *Kluwer Academic Publishers*, Dordrecht, 1999.

[92] J.Laver, “Principles of Phonetics”, *Cambridge University Press*, Cambridge, 1994.

[93] J.Clark, and C.Yallop, “Introduction to Phonetics and Phonology”. *Basil Blackwell*, Wiltshire, 1990.

[94] G.Ashour, and I.Gath, “Characterization of speech during imitation”, *Proc. 6th European Conference on Speech Communication and Technology (Eurospeech 1999)* , Budapest, Hungary, 1999, pp. 1187–1190.

[95] E.Zwicker, and E.Terhardt, “Analytical expressions for critical band rate and critical bandwidth as a function of frequency”, *Journal of the Acoustic Society of America*, 68,1980, 1523–1525.

[96] F.Nolan, "The Phonetic Bases of Speaker Recognition", *Cambridge University Press*, Cambridge, 1983.

[97] J.Markel, B.Oshika, and A.H. Gray, "Long-term feature averaging for speaker recognition", *IEEE Trans. Acoustics, Speech, and Signal Processing*, 25(4): 330–337, 1997.

[98] M.Carey, E.Parris, H.Lloyd-Thomas, and S. Bennett, "Robust prosodic features for speaker identification", *Proc. Int. Conf. on Spoken Language Processing(ICSLP 1996)*, Philadelphia, ennsylvania, USA, 1996, pp. 1800–1803.

[99] P.Mokhtari, "An Acoustic-Phonetic and Articulatory Study of Speech-Speaker Dichotomy", *PhD thesis*, School of Computer Science, University of New South Wales, Canberra, Australia, 1998.

[100] J.Eatock, and J.Mason, "A quantitative assesment of the relative speaker discriminating properties of phonemes", *Proc. Int. Conf. on Acoustics, Speech,and Signal Processing (ICASSP 1994)*, Adelaide, Australia, 1994, pp. 133–136.

[101] M.Sambur, "Selection of acoustic features for speaker identification", *IEEE Trans. Acoustics, Speech, and Signal Processing*, 23(2):176–182, 1975.

[102] L.Besacier, J.Bonastre, and C.Fredouille, "Localization and selection of speaker-specific information with statistical modeling", *Speech Communications, 31*, 2000, pp.89–106.

[103] P.Sivakumaran, A.Ariyaeeinia, and M.Loomes, "Sub-band based textdependent speaker verification", *Speech Communications*, 41, 2003, pp.485–509.

[104] T.Kinnunen, "Designing a speaker-discriminative filter bank for speaker recognition", *Proc. Int. Conf. on Spoken Language Processing (ICSLP 2002)*, Denver, Colorado, USA, 2002, pp. 2325–2328.

[105] K.Yoshida, T K.akagi, and K.Ozeki, "Speaker identification using subband HMMs", *Proc. 6th European Conference on Speech Communication and Technology(Eurospeech 1999)* ,Budapest, Hungary, 1999, pp. 1019–1022.

[106] B.Atal, "Automatic speaker recognition based on pitch contours", *Journal of the Acoustic Society of America*, 52(6):1687–1697, 1972.

[107] R A.osenberg, "Automatic speaker verification: a review", *Proceedings of the IEEE*, 64(4):475–487, 1976.

[108] Y.Kyung, and H.-S.Lee, "Text independent speaker recognition using microprosody", *Proc. Int. Conf. on Spoken Language Processing (ICSLP 1998)*, Sydney, Australia, 1998.

[109] F.Weber, L.Manganaro, B.Peskin, and E. Shriberg, "Using prosodic and lexical information for speaker identification", *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2002)* , Orlando, Florida, USA, 2002, pp. 141–144.

[110] http://www.ldc.upenn.edu/. Linguistic data consortium. WWW page, December 2003.

[111] K.Paliwal, and L.Alsteris, "Usefulness of phase spectrum in human speech perception", *Proc. 8th European Conference on Speech Communication and Technology(Eurospeech 2003)* , Geneva, Switzerland, 2003, pp.2117–2120.

[112] J. R. Deller, J. H. L. Hansen, J. G. Proakis, "*Discrete-Time Processing of Speech Signals*", Piscataway, IEEE Press, 2000.

[113] 杨福生,"小波变换的工程分析与应用", *科学出版社*, 1999

[114] D.Maltoni, A.Jain, D.Maio, and S.Prabhakar, "Handbook of Fingerprint Recognition", *Springer Verlag*, New York, 2003.

[115] Ching-Tang HSIEH, and You-Chuang WANG, "A Robust Speaker Identification System Based on Wavelet Transform", *IEICE Trans. On Inf.&Syst.*, 84(7):839-846, 2001.

[116] Y.B. Yu, H.M. Zhao, "Speech Recognition Based on Estimation of Mutual Information", *Proc. Of Int. Conf. on Spoken language Processing(ICSLP2000)*, Beijing, 2000,pp.3194-3197.

[117] 俞一彪，赵鹤鸣, "语音识别浏览器 VoiceIE 设计与实现"，*数据采集与处理*，17(1), 2002, pp.35-38

[118] L. R.Rabiner, R.W.Schafer, "Digital Processing of Speech Signal", *Science Publication Inc.*, Beijing,1983.

[119] A. H. Gray, J. D. Markel, "Distance measures for speech processing", *IEEE Trans. On ASSP*, **24**(5): 380-391, 1976.

[120] R. M. Gray, A. Buzo, A. H. Gray, "Distortion measures for speech processing", *IEEE Trans. On ASSP*, **28**(4): 367-376, 1980.

[121] D. O'Shaughnessy, "Speech Communications-Human and Machine", *IEEE Press*, 2000, pp378-383,

[122] 俞一彪，赵鹤鸣，"语音信号互信息估计的非线性搜索算法及识别应用"，*信号处理*, **18**(2):102-106, 2002.

[123] 俞一彪，赵鹤鸣，"基于 MIM 模型的语音识别引擎 SDSE",*计算机工程与应用*, 38(11):pp91-93, 2002.

[124] 俞一彪，赵鹤鸣，"运用互信息匹配及关键词分析的语音对话系统",*小型微型计算机系统*, 24(1):147-150, 2003.

[125] F.K.Soong, A.E.Rosenberg, L.R.Rabiner, "A vector quantization approach to speaker recognition", *Proc. Of ICASSP*, 1985, pp.387-390.

[126] L.Wong, M.Russell, "Text-dependent speaker verification under noisy conditions parallel model combination", *Proc. Of ICASSP*, 2001.

[127] Y.B.Yu, S.Z.Wang, "Speaker Identification based complete feature corpus and multi-step mini-max search matching", *Proc. Of International Conference on Signal Processing*, Beijing, 2004.

攻读博士学位期间公开发表和已录用的论文

题　　目	刊物与日期	作者	类别
1. Speaker identification based on complete feature corpus and multi-step mini-max search matching	ICSP2004 2004, Beijing	Yi-biao Yu, Shuo-zhong Wang	ISTP
2. 文本无关说话人识别的全特征矢量集模型及互信息评估算法	声学学报（待发表）	俞一彪，王朔中	EI
3. 采用遗传算法的 VQ 码本设计及说话人识别	信号处理 2005 年第 3 期	芮贤义，俞一彪	
4. 基于互信息匹配模型的说话人识别	声学学报 2004 年第 5 期	俞一彪，王朔中	EI
5. 特征空间映射提高说话人识别性能	通信技术 2004 年第 3 期	许允喜，俞一彪	
6. 基于 GMM 的汉语说话人识别特性分析	通信技术 2004 年第 2 期	许允喜，俞一彪	
7. 基于 MI-OneStage 的连续数字语音识别	通信技术 2003 年第 3 期	徐　华，俞一彪	
8. 基于互信息匹配及关键词分析的语音对话系统	小型微型计算机系统 2003 年第 1 期	俞一彪，赵鹤鸣	EI
9. 一种适于计算声场景分析的混叠语音基音检测方法	电子学报 2003 年第 1 期	赵鹤鸣，俞一彪	EI
10. 吴语文语转换系统中语音韵律的控制	通信技术 2002 年第 10 期	俞一彪，段凯宇	
11. 语音识别互信息测度的聚类特性分析与实验评价	信号处理 2002 年第 5 期	俞一彪，王朔中	
12. Comparative study of mutual information measure for speech recognition	Forum on Information Technology 2002 Tokyo, Sept. 2002	Yi-biao Yu, Shuo-zhong Wang	
13. 基于 MIM 模型的语音识别引擎 SDSE	计算机工程与应用 2002 年第 11 期	俞一彪，赵鹤鸣	
14. 基于基音同步帧的吴语语音合成	通信技术 2002 年第 3 期	段凯宇，俞一彪	
15. 语音信号互信息估计的非线性搜索算法及识别应用	信号处理 2002 年第 2 期	俞一彪，赵鹤鸣	
16. 语音识别浏览器 VoiceIE 设计与实现	数据采集与处理 2002 年第 1 期	俞一彪，赵鹤鸣	EI

致　谢

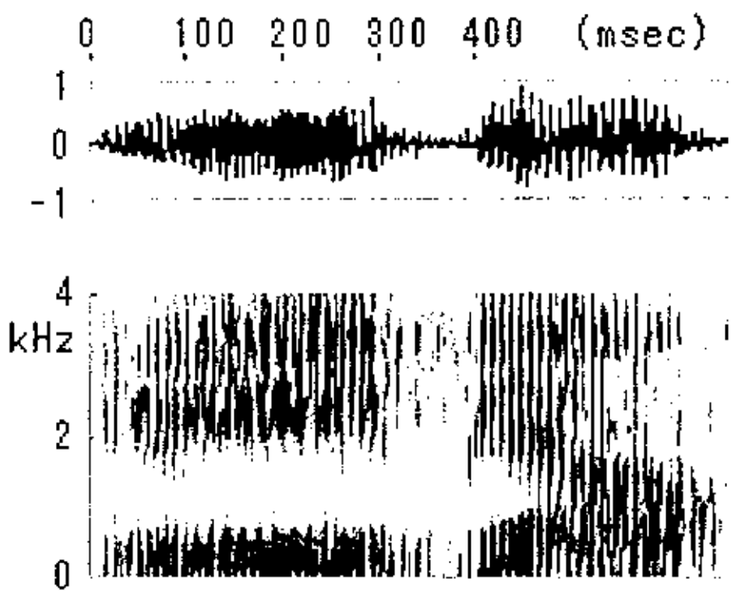
经过三年多的努力，终于完成了博士课程的学习与研究工作。回顾这三年来攻读博士学位紧张又充实的历程，伴随着不断的奔波和辛苦，得到的是许多学习的体会和研究的收获。今天，我把学习和研究的成果通过这篇论文呈现出来，以此对这几年来给予我教育和帮助的老师、同行和家人表示深深的感谢。

首先，我要感谢指导老师王朔中教授。王老师治学严谨、知识渊博，对科学研究有着深刻的认识和精益求精的执着。与王老师的讨论有益又透彻，正是在他的指导和鼓励下完成了这篇博士论文。

感谢日本爱知工业大学情报通信工学科的井研治（Inomoto Kenji）教授，他的帮助使我在该校做访问学者期间的研究工作得以顺利进行，并有机会参加了 2002 年东京信息技术论坛会议。

在博士课程学习中，能够有机会倾听汪敏教授、方勇教授的精彩讲解，学到了很多新的知识，在此表示感谢。

我也要感谢通信与信息工程学院 2002、2003 级博士班的同学，他们的一些帮助节约了我的时间。



俞一彪

2004 年 6 月