2004年上海大学博士学位论文 59

基于互信息理论的说话人识别研究

作　者：俞一彪
专　业：通信与信息系统
导　师：王朔中

上海大学出版社
·上海·

2004 年上海大学博士学位论文

基于互信息理论的说话人识别研究

作　者：俞一彪
专　业：通信与信息系统
导　师：王朔中

上海大学出版社
· 上海 ·

Shanghai University Doctoral Dissertation (2004)

The Research of Speaker Recognition Based on Mutual Information Theory

Candidate: Yu Yi-biao
Major: Communication and Information System
Supervisor: Prof. Wang Shuo-zhong

Shanghai University Press
• Shanghai •

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.

答辩委员会名单:

主任：	陈泳冠	教授，同济大学电子信息工程系	200433
委员：	张兆扬	教授，上海大学电子信息工程系	200072
	王治钢	研究员，上海航天局 809 研究所	200031
	吴亚明	研究员，中科院微系统所	200050
	万旺根	教授，上海大学通信工程系	200072
导师：	王朔中	教授，上海大学	200072

评阅人名单：

袁保宗	教授，北京交通大学	100044
吴镇扬	教授，东南大学无线电工程系	210096
梁庆林	教授，北京大学电子学系	100871

评议人名单：

张立明	教授，复旦大学电子工程系	200433
翁默颖	教授，华东师大电子科学系	200062
王炳锡	教授，解放军信息工程大学	450002
张兆扬	教授，上海大学电子信息工程系	200072
顾亚平	研究员，中科院东海站	200032
方　勇	教授，上海大学通信与信息工程学院	200072

答辩委员会对论文的评语

俞一彪的博士论文研究说话人识别，是信息系统中身份确认以及司法鉴定等领域的重要内容和信息技术的前沿课题，具有重要的学术意义和实际应用价值. 论文运用互信息理论进行说话人识别研究，针对独立可加性模型提出了有效的方法，并通过实验进行了验证.

论文的创新成果包括：

(1) 从信息量的角度出发研究语音信号之间的特征相关性，将语音信号之间互信息的计算归结为随机干扰信号的熵的计算.

(2) 通过类内凝聚度、类间耦合度、类间重叠三大指标对互信息测度的聚类特性进行分析，与其他几种常用测度进行比较，表明了互信息测度的有效性和优越性.

(3) 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 来计算互信息.

(4) 利用互信息测度，针对不同识别要求提出基于文本的说话人识别的多模板模型 MTM 以及文本无关说话人识别的全特征矢量集模型 CFC. 实验表明这些模型能充分表达说话人的语音特征.

(5) 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特性，提出多级最小最大搜索匹配算法计算全特征矢量集模型 CFC 和语音信号的互信息.该算法优于 GMM 的模型识别算法.

论文立论正确，论据充分；结构合理，表达通顺；实验方法合理，结果可信；在答辩中叙述清楚、回答问题正确. 该论文表明作者具有坚实的专业理论基础，分析问题解决问题的能力强. 论文已达到博士学位水平.

答辩委员会表决结果

答辩委员会经无记名投票一致通过论文答辩. 建议校学位委员会授予工学博士学位.

答辩委员会主席: 陈泳恩

2004 年 9 月 13 日

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：

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.

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.

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.

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，并与其他常用测度 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.

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.

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 on 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 (LMP) 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.

Key words Speaker recognition, Mutual information, Matching, Linguistic property, Individual property

目　录

第一章 绪 论

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；②PIN 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. 第三种方法是一种基于生物特征的身份鉴定识别方法[1~3]，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.