声纹识别系统原理及其关键技术

朱浩冰，郭东辉
（厦门大学，福建 厦门 361005）

摘 要：以声纹为特征的身份识别技术具有十分广阔的应用前景。该文介绍了声纹识别系统的应用分类及其基本技术原理，重点分析了声纹识别系统中的特征参数提取、模式匹配判断等关键技术问题，并总结声纹识别技术的研究进展。

关键词：声纹识别；特征参数提取；模式匹配判断

Principles and Key Technologies of Voiceprint Recognition System

ZHU Hao-bing, GUO Dong-hui
(Xiamen University, Xiamen 361005, P.R.China)

Abstract: The identity recognition technology which uses voiceprint as feature has very broad application foreground. In this paper, the applied classification and basic technology principles of voiceprint recognition system are presented. Based on this, two key technologies of feature parameter extraction and pattern matching judgment in voiceprint recognition system are analyzed in detail, then the research development of voiceprint recognition technology are summarized.

Key words: voiceprint recognition; feature parameter extraction; pattern matching judgment

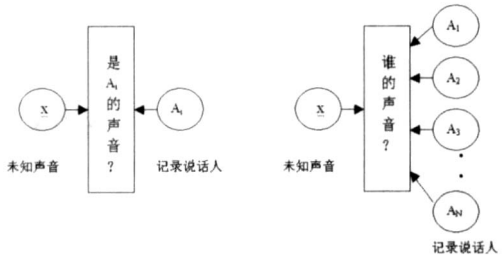
1 引言

伴随着信息技术和网络技术的迅猛发展，人们对身份识别技术的需求越来越多，对其安全可靠性的要求也越来越严格。基于传统密码认证的身份识别技术在实际信息网络应用中已经暴露出许多不足之处，而基于生物特征辨别的身份识别技术近年来也日益成熟并在实际应用中展现出极大的优越性[1]。其中，声纹识别技术便是近年来发展起来的一种新的更有效的身份识别技术之一。

声纹是指说话人语音频谱的信息图。由于每个人的发音器官不同，所发出来的声音及其音调各不相同，因此，声纹作为基本特征来实现人的身份识别具有实际的不可替代性和稳定性，使声纹识别技术广泛地应用于信息网络的各个领域。尽管至今已有许多介绍声纹识别技术及应用的相关论文发表[2-5]，但是，多数论文仅局限于介绍声纹识别技术的某一具体方法改进或某一缺点问题克服。为此，本文希望能够通过综述性地介绍声纹识别系统的基本原理及其关键技术，并总结分析声纹识别的技术研究进展及其应用方向，为人们进一步研究声纹识别技术及其应用提供技术参考。

2 声纹识别系统及其技术实现原理

声纹识别系统是基于对说话人的语音识别或鉴别的应用系统，它是根据人所说语音信息而表征出来的说话人的生理和行为特征来自动识别或鉴别说话人身份的技术系统[3,6]。基于声纹识别系统的不同应用，声纹识别系统的技术实现基本上可以分归两类，如图1所示，即说话人确认技术和说话人辨认技术[3]。前者是用于判断未知说话人是否为某个指定人；后者则是用于辨认未知说话人是已记录说话人中的哪一位。因此，声纹识别系统最终要解决的技术问题就是体现在“一对一”的匹配判断问题或“多选一”的比较判断问题。



(a) 说话人确认　(b) 说话人辨认

图1 说话人确认与辨认

从声纹识别系统的使用场合来看，需要判别的声音其来源基本可分为3种情况，即文本提示型、文本相关型和文本无关型[4]。其中，文本提示型的声纹识别系统要求被鉴别的人需要根据给定的文字进行发音判别，即要求用户配合发音，才能实现识别功能；文本有关型的声纹识别系统要求系统录制有被判别人一定数量的规定文本内容的声音，只要判别人发出相关内容的声音就可以实现判别功

基金项目：本文得到福建省自然科学基金计划资助项目（A0410007）、国家教育部新世纪人才计划项目和国家人事部留学人员创业基金项目的联合资助。

能；而文本无关型的声纹识别系统则不规定说话人的发音内容，只要系统中录有说话人的声音，就能够识别是否为该说话人。可见，文本无关型的声纹识别系统的技术含量要求比较高，它不仅仅需要解决匹配判断问题，还需要预先提取说话人的语音特征，才能进行判断识别。

此外，从声纹识别的目标对象来看，声纹识别系统的适用范围可以分为两类，即闭集识别和开集识别[7,8]。前者是指对特定人群中的说话人识别，即被判定的说话人是在已记录说话人集合内，而后者是指被判定的说话人可能不在已被记录的这个集合内。相比于闭集识别系统，开集识别系统需要增加一个阈值来判断未知说话人是否在已记录说话人集合内。如果不在集合内，系统需要重新进行语音记录和训练。因此，适用于开集识别的声纹识别系统还需要解决训练学习的技术问题。

总的看来，一个典型的声纹识别系统的技术实现原理可以用如图 2 所示的框图来概括。即声纹识别系统的工作过程一般可以分为两个过程：训练过程和识别过程。无论训练还是识别，都需要首先对输入的原始语音信号进行预处理，如采样、量化、预加重和加窗等处理过程[9]，以实现语音特征的提取功能。在训练过程中，声纹识别系统要对所提取出来的说话人语音特征进行学习训练，建立声纹模板或语音模型库，或者对系统中已有的声纹模板或语音模型库进行适应性修改。在识别过程中，声纹识别系统要根据系统已有的声纹模板或语音模型库对输入语音的特征参数进行模式匹配计算，从而实现识别判断，得出识别结果。

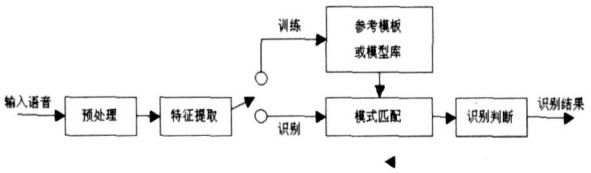


图 2　声纹识别系统的技术实现原理框图

3 声纹识别的关键技术

从声纹识别系统技术实现的基本原理来看，其关键技术在于语音预处理后的特征参数提取技术、系统训练过程中的建模学习技术及系统识别过程中的模式匹配识别判断技术。其中，声纹识别系统中应用的建模学习技术类同其他样本学习技术[10-12]，因此，这里主要介绍语音特征参数提取技术和模式匹配识别技术。

3.1 语音特征参数提取技术

特征参数提取的目的就是从说话人语音中提取出能够表征说话人特定器官结构或习惯行为的特征参数。该特征参数对同一说话人具有相对稳定性，不能随时间或环境变化而不一致，对同一说话人的不同话语也应该是一致的；而对于不同的说话人即使说同样的话语也应该易于区分，具有不易模仿性和较强的抗噪性。目前常用语音特征参数的提取技术主要体现在以下的几种特征参数提取：

3.1.1 语音频谱参数

这种参数的提取主要是基于说话人发声器官，如声门、声道和鼻腔等的特殊结构而提取出说话人语音的短时谱特征（即基音频率谱及其轮廓）[6]。它是表征说话人声音的激励源和声道的固有特征，可以反映说话人语音器官的差异，而短时谱随时间或幅度变化的特征，在一定程度上反映了说话人的发音习惯。因此，语音频谱参数在声纹识别中的应用主要体现在基音频谱及其轮廓[13]、基音帧的能量[6,14]、基音共振峰的出现频率及其轨迹[15]等的参数表征与模式识别。

3.1.2 线性预测参数

这种参数的提取则是以若干“过去”的语音抽样或已有的数学模型来逼近当前的语音抽样，用相应的逼近参数来估计的语音特征[16]。 它能够实现用少量的参数有效地表现语音的波形和频谱特性，具有计算效率高、应用灵活的特点。目前声纹识别中广泛应用的线性预测参数提取方法主要包括有：线性预测倒谱（LPCC）[17]、线谱对（LSP）[18]、自相关和对数面积比[19]、Mel 频率倒谱（MFCC）[9,17,20]、感知线性预测（PLP）[21]等不同方法的特征系数提取。

3.1.3 小波特征参数

这种参数的提取是利用小波变换技术[22]对语音信号进行分析处理以获得表示语音特征的小波系数。小波变换具有分辨率可变、无平稳性要求和时频域兼容表征等优点，能够有效地表征说话人的个性信息。因此，它在声纹识别系统中实际应用体现出计算量小、复杂度低、识别效果好等特点[22]，是近年来语音特征参数提取技术的研究热点。

此外，不同方法提取出来的特征参数如果其之间相关性不大时，说明它们分别反映了语音信号的不同特征，因此，也可以通过不同特征参数的组合技术[23]来获得更适用于模式匹配识别判断的语音特征参数模型。

3.2 模式匹配识别判断技术

模式匹配识别判断的目的在于获取表现说话人个性的特征参数的基础上，将待识别的特征参数模板或模型与训

练学习时得到的模板或模型库作相似性匹配，得到特征模式之间的相似性距离度量，并选取适当的距离度量作为门限值，从而识别判断出可能结果中最好的结果。由识别系统输出。目前常用模式匹配识别判断技术主要体现在以下几种模型：

3.2.1 矢量化模型

这种模型通过某种矢量化方法，将提取的说话人特征参数编辑为某种具有代表性的特定矢量，识别时将待识别参数按此特定矢量进行模型编辑，依照一定的判决标准如：量化时产生的失真度来得出识别结果。矢量化模型在声纹识别系统中的应用主要包括：动态时间规整（DTW）[3]、矢量量化（VQ）[24,25] 及支持向量机（SVM）[11,26] 等。

3.2.2 随机模型

这种模型是一种基于转移概率和传输概率的模型。在使用随机模型进行识别时，为每个说话人建立发声模型，通过训练得到状态转移概率矩阵和符号输出概率矩阵，识别时计算待识别语音在状态转移过程中的最大概率，根据最大概率对应的模型进行识别判断。其优点是计算有效，性能较好，因此成为主流的模式匹配识别判断技术。随机模型在声纹识别系统中的应用主要包括：隐马尔可夫模型（HMM）[4,24,27]、高斯混合模型（GMM）[28]。

3.2.3 神经网络模型

神经网络模型 [29] 在某种程度上模拟了生物的感知特性，它是一种分布式并行处理结构的网络模型，具有自组织和自学习能力、很强的复杂分类边界区分能力以及对不完全信息的鲁棒性，在训练过程中能不断调整自身的参数权值和结构拓扑，以适应环境和系统性能优化的需求。其优点是速度快、识别率高，近几年来不断地被完善 [30]。

此外，为了提高声纹识别系统的准确率，将不同的模式匹配方法融合起来进行识别，也是声纹识别系统研究的一个方向。

4 声纹识别技术的研究进展

声纹识别技术的研究始于 20 世纪 30 年代，从技术特点上看可以分为以下几个发展阶段 [3]：(1) 技术启蒙阶段即 20 世纪 30 年代，研究工作主要集中在人耳听辨实验和探讨听音识别的可能性方面。(2) 技术突破阶段即 20 世纪 60 至 70 年代早期，研究重点主要在各种识别参数的提取、选择和试验上，并将倒谱比较和线性预测分析等线性处理和简单模式匹配方法实际应用于声纹识别 [31]。(3) 技术发展阶段，即从 20 世纪 70 年代末开始随着计算机技术的飞速发展，声纹识别的研究转向对各种声学特征参数的非线性处理及新的模式匹配方法上 [32]。

其中在特征参数提取技术方面，在 20 世纪 70 年代末，如，小波特征参数及不同特征参数的线性预测组合等，非线性处理方法相继提出并得到广泛地应用。特别是近年来，DSP 芯片计算技术的采用，使目前的语音特征参数提取技术达到比较成熟的阶段。而在模式匹配判断技术方面：20 世纪 70 年代末，动态时间规整 [3] 与矢量量化 [25] 技术首先被应用到声纹识别上，使声纹识别系统的性能得到有效的提高；20 世纪 80 年代开始，隐马尔可夫模型 [27]、人工神经网络 [29] 等技术在声纹识别方面得到有效的利用，逐渐成为声纹识别系统主流的模式匹配方法；进入 90 年代后，高斯混合模型 [28] 技术由于简单、有效及较好的鲁棒性也迅速成为重要的声纹识别技术；步入 21 世纪以来，支持向量机 [26] 技术及多种模式匹配方法融合也得到不断深入研究与发展，并进入了商业化实用阶段。

5 结束语

与其他生物识别技术相比，由于声纹识别技术具有简便、准确、经济及可扩展性良好等众多优势，目前在世界范围内广泛应用于各个领域。例如：在公共安全系统中，声纹识别技术可以在一段录音中查找出犯罪嫌疑人或缩小侦察范围，并且还可在法庭上提供犯罪嫌疑人身份确认的旁证；在互联网应用及通信领域，声纹识别技术可以运用于诸如电话银行、电子商务、安全控制、计算机远程登录等领域；在军事领域，可以实现战场环境监听，辨认出敌方指挥员。

当然，声纹识别技术还存在一些技术难点，如：用很短的语音进行模型训练和识别、有效地区分模仿声音和真正声音、消除或减弱声音变化带来的影响；消除信道差异和背景噪音带来的影响等。不过，总的来说，声纹识别技术作为生物识别技术的代表之一，具有十分广阔的应用前景。

参考文献：

[1] Simon Liu, Mark Silverman. A Practical Guide to Biometric Security Technology [J]. IEEE Computer Society, IT Pro-Security, 2001, 3(1): 27-32.

[2] 李财莲，赵小阳，王丽娟，岳振军．说话人识别中关键技术的现状与展望 [J]. 军事通信技术，2005,26(2):62-65.

[3] Joseph P. Campbell, Jr. Speaker recognition: a tutorial [J]. Proceedings of the IEEE, 1997, 85: 1437-1462.
[4] Chi-Wei Che, Qi-guang Lin, Dong-Suk Yuk. An HMM Approach to Text-Prompted Speaker Verification [A]. The 1996 IEEE International Conference on Acoustics, Speech and Signal Processing Conference Proceedings[C], 1996, 2: 673-676.
[5] 宁飞，陈频．说话人识别的几种方法[J]．电声技术，2001,12(1):9 — 15.
[6] 蔡莲红，黄德智，蔡锐．现代语音技术基础与应用[M]．北京：清华大学出版社，2003.
[7] 何致远，胡起秀，徐光祐．两级决策的开集说话人辨认方法[J]．清华大学学报（自然科学版），2003,43(4):516-520.
[8] 王金明，张雄伟．一种模糊高斯混合说话人识别模型[J]．解放军理工大学学报（自然科学版），2006, 7(3):214-219.
[9] 张万里，刘桥．Mel频率倒谱系数提取及其在声纹识别中的作用[J]．贵州大学学报，2005,22(2):207-210.
[10] 李虎生，刘加，刘润生．语音识别说话人自适应研究现状及发展趋势[J]．电子学报，2003,31(1):103-108.
[11] 刘江华，程君实，陈佳品．支持向量机训练算法综述[J]．信息与控制，2002,31(1):45-50.
[12] 闫友彪，陈元琰．机器学习的主要策略综述[J]．计算机应用研究，2004,(7):4-13.
[13] 王文剑，王长富，戴蓓倩，陆伟．基于藤崎模型的汉语语音基频轮廓的参数提取[J]．小型微型计算机系统，1999，20(10):756-759.
[14] 李桦，安钢，樊新海．短时能频值在语音端点检测中的应用[J]．测试技术学报，1999,13(1):21-27.
[15] 章文义，朱杰，陈斐利．一种新的共振峰参数提取算法及在语音识别中的应用[J]．计算机工程，2003,29(13):67-68.
[16] 余铁城．语波的线性预测原理及其应用[J]．声学学报，1980,(4):291-300.
[17] 王让定，柴佩琪．语音倒谱特征的研究[J]．计算机工程，2003,29(13):31-33.
[18] C.S. Liu, M.T. Lin, W.J. Wang, H.C. Wang. Study of line spectrum pair frequencies for speaker recognition [A]. The 1990 IEEE International Conference on Acoustics, Speech and Signal Processing Conference Proceedings[C], 1990, 1: 277-280.
[19] N.Mohankrishnan, M.Sridhar, Sid-Ahmed. A composite scheme for text-independent speaker recognition [A]. The 1990 IEEE International Conference on Acoustics, Speech and Signal Processing Conference Proceedings[C], 1982, 7: 1653-1656.
[20] Davis S B, Mermelstein P. Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences [J]. IEEE Transactions on Speech Acoustic Processing, 1980, 28: 357-366.
[21] Hermansky H. Perceptual Linear Predictive (PLP) Analysis of Speech [J]. Journal of the Acoustical Society of America, 1990, 87(4): 1738-1752.
[22] C.T. Hsieh, Y.C Wang. A Robust Speaker Identification System Based on Wavelet Transform [J]. Trans. IEICE on Information and Systems, 2001, E84-D (7): 839-846.
[23] 汪峥，连翰，王建军．说话人识别中特征参数提取的一种新方法[J]．复旦学报（自然科学版），2005,44(1):197-200.
[24] Matsui T., Furui S. Comparison of Text-Independent Speaker Recognition Methods Using VQ-Distortion and Discrete/Continuous HMM's [J]. IEEE Transactions on Speech and Audio Processing, 1994, 2(3): 456-459.
[25] 张炜，胡起秀，吴文虎．距离加权矢量量化文本无关的说话人识别[J]．清华大学学报（自然科学版），1997,37(3):20-23.
[26] 侯风雷，王炳锡．基于支持向量机的说话人辨认研究[J]．通信学报，2002,23(6):61-67.
[27] D Charlet, D Jouvet, O Collin. An alternative normalization scheme in HMM-based text-dependent speaker verification [J]. Speech Communication, 2000, 31: 113-120.
[28] Douglas A Reynolds, Richard C Rose. Robust text-independent speaker identification using Gaussian mixture speaker models [J]. IEEE Trans. on Speech and Audio Processing, 1995, 3(1): 77- 83.
[29] Farrell K.R, Mammone R.J, Assaleh K.T. Speaker recognition using neural networks and conventional classifiers [J]. IEEE Trans on Speech and Audio Processing, 1994, 2 (1): 194-205.
[30] 白莹，赵振东，戚银城，王斌，郭建勇．基于小波神经网络的与文本无关说活人识别方法研究[J]．电子与信息学报，2006,28(6):1036-1039.
[31] S.Pruzansky. Pattern-Matching Procedure for Automatic Talker Recognition [J]. J.Acoust.Soc.Am, 1963, 35: 354-358.
[32] B.S.Atal. Automatic Speaker Recognition based on pitch contours [J]. J.Acoust.Soc.Am, 1972, 52: 1687-1697.

作者简介：朱浩冰（1982年—），男，福建龙岩人，厦门大学硕士研究生，研究方向：人工智能，语音信号处理。
收稿日期：2007-07-25