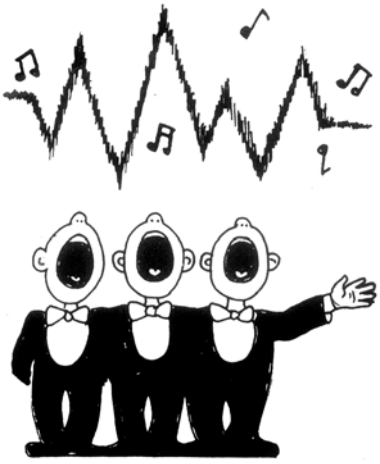
音频信息识别与检索技术



颜永红

随着互联网（Internet）和电信网等信息网络的蓬勃发展，人们的信息交互变得越来越方便，除了文字信息，语音、音乐、图像等多媒体信息也越来越多地进入了人们的日常生活中。例如，在广播或电视媒体中每天都在增加的语音文档或视频文档，在日常生活中的音视频电子邮件等。目前针对文字信息的检索已经有许多成功的应用，如：Google、Yahoo、百度等搜索引擎，但如何利用计算机对非结构化的海量多媒体数据做信息检索是目前急需解决的一个难题。音频信息中主要包含语音、音乐、说话人、语种等内容信息。相应的音频识别技术主要包含以下几类：语音识别技术、说话人识别技术、语种识别技术、音乐识别检索技术。语音识别技术可以将音频中语音转换为文字，说话人识别技术可以确定音频信息中的说话人身份，语种识别技术可以确定音频信息中所用语言的种类，音乐识别检索技术可以识别检索出音频中的音乐旋律片断。通俗一些说，给定一个音频文件，运用上述技术可以自动从音频文件中获得这段音频的内容信息：“由谁说的、用的什么语言、说的内容是什么”。因此，利用音频信息识别与检索技术可以对多媒体文档中的音频信息自动建立索引，以解决对非结构化的海量多媒体数据的信息检索。以下将主要介绍音频信息识别与检索的几项关键技术。

一、语音识别技术

语音是人与人之间交互最自然的手段，语音识别技术的目的是要使机器听懂人说话，因此语音识别技术正成为信息技术中人机接口的关键技术。

语音识别的研究始于 20 世纪 50 年代，贝尔实验室的 Rabiner 等研究人员在 20 世纪 80 年代率先将隐含马尔可夫模型（Hidden Markov Model, HMM）方法应用于语音识别，并成功地应用在了非特定人连续语音识别系统中，从那时开始隐含马尔可夫模型成为了语音识别的主流方法，20 世纪 90 年代以后，世界许多著名的国际大公司如：IBM、Apple、AT&T、NTT、MicroSoft 等都逐渐开始了语音识别技术的学术及实用化研究，并推出了一系列的语音识别应用产品，其中的典型代表是 IBM 公司 1998 年推出的 ViaVoice 系统以及 Dragon 公司的 Naturally Speaking 系统，这些系统基于大词表、非特定人连续语音识别技术，在 PC 机上可以实现语音输入。进入到 21 世纪，随着微电子技术的发展，语音识别技术从实验室开始逐渐走向了实际应用，在语音识别应用市场上，美国 Nuance 公司是目前世界最大的语音识别技术提供商，其语音识别产品涵盖电信平台、嵌入式平台应用。

我国语音识别技术研究工作始于 20 世纪 50 年代，中科院声学所率先开始了语音识别技术的研究，经过多年的努力，我国语音识别技术的研究水平基本与国外同步。长期以来，国家对语音识别技术的研究给予了较大的支持，国家科委（科技部）863 智能计算机专家组从 20 世纪 90 年代初开始举办语音识别技术的评测，较好地促进了国内语音识别技术的繁荣发展。目前在国内许多研究单位如中科院声学所、中科院自动化所、北京大学、清华大学、北京邮电大学、北京理工大学、哈尔滨工业大学等都有相应的实验室从事语音识别技术的研究工作。

一个语音识别系统通常由以下模块组成：前端预处理、特征提取、解码识别，如图 1 所示。

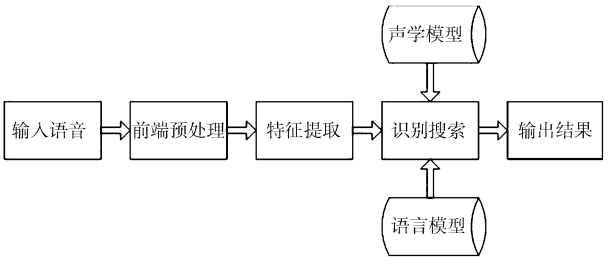


图 1　语音识别系统框图

下面简要介绍图 1 中的主要技术模块：

1. 前端预处理

前端预处理模块主要是采用数字信号处理的方法，消除输入语音中受环境噪声和信道影响，同时

该模块对语音信号进行端点检测，检测到语音的起始点和终止点，以消除噪声对识别结果的影响，提高识别性能。

2. 特征提取

检测后的语音数据流被送入特征提取模块，提取得到语音信号的特征矢量流。目前语音识别用的特征是短时频谱特征，因此需要对语音信号进行预加重、分帧、加窗、频域变换、倒谱变换、差分等处理，最终得到多维的语音特征矢量，识别特征通常用的帧长为 25ms，帧移为 10ms，因此通常一秒钟的语音有 100 帧的语音特征矢量序列。目前常用的语音特征参数有：梅尔刻度式倒频谱参数（MFCC）及基于感知线性预测分析提取的感知线性预测倒谱（MFPLP）等。

3. 识别搜索

经过特征提取模块后，语音特征矢量序列被送入识别搜索模块。在这个模块中，未知语音信号的特征矢量序列与系统中的声学模型、语言模型相配合，进行识别搜索，常用的搜索算法为 Viterbi 算法，运用该算法可以找出一条与未知语音信号相匹配的最佳路径，该最佳路径对应的汉字序列即为识别的结果。这个模块是语音识别系统的核心，其中声学模型和语言模型与语音识别系统性能密切相关。

声学模型是语音识别系统最核心的资源文件，包含了对于语音信号频谱和时间序列特征的精确描述。声学模型通常是对大量的训练语音数据通过统计训练获得的，目前语音识别中常用的声学模型是基于连续高斯概率密度分布的隐含马尔可夫模型（Continuous Density Hidden Markov Model, CDHMM），传统的 HMM 训练方法是基于最大似然估计准则（Maximum Likelihood, ML）的 Baum-Welch 算法，近年来，基于最大互信息准则（Maximum Mutual Information Estimation, MMIE）和基于最小词/音素错误准则（Minimum Word Error, MWE/Minimum Phone Error, MPE）声学模型鉴别性训练方法被实验证明比基于 ML 的方法更有效，因此被广泛采用。

语言模型是大词汇量连续语音识别系统的一个重要组成部分，在声学层识别出错时，可以根据语言学模型、语法结构、语义学知识进行判断纠正，特别是对于汉语来说，有多音字问题，如果单纯靠声学层的知识是不可能将声音转换成正确文字的，必须通过语言层的知识才能确定其对应的正确文字。目前在语音识别系统中得到成功的语言模型是基于统计规则的 N-Gram 语言模型，如：二元文法及三元文法语言模型，此外，在解码过程中引入语言学的知识可以有助于减少识别的搜索路径，从而提高识别的速度和识别的准确性。

语音识别技术经过多年的发展，在正常办公室环境下朗读式语音的正确识别率可以达到 90%以上，因此目前的语音识别技术针对正常的办公室环境下的朗读语音，已经能够达到实用水平；然而在人们自由交谈方式以及在实际应用环境下，语音识别的正确率会急剧下降。因此目前语音识别技术难点在于如何解决语音识别的鲁棒性问题，也就是说如何解决环境噪声、方言口音等实际应用中常见的现象对识别性能的影响。

环境噪声是影响语音识别性能的主要因素之一，目前常用的解决方法有以下几种：在信号层面，用信号处理的手段过滤掉噪声；在语音特征层面，对语音特征做一些归整处理，如倒谱均值减及方差规整、特征的高斯化处理等；在模型层面，可以在声学模型的训练数据中加入一些实际应用环境中的带噪数据；这些方法均可以有效提高噪声环境下的语音识别性能。另一方面，方言口音对识别性能也有较大的影响，常用的解决方法有以下几种：在发音层面，可以总结方言口音的发音变异规则，将这些发音变异规则应用到识别系统的发音字典中，从而允许说话人带有一定的方言口音；在模型层面，在声学模型训练数据中加入方言口音数据，能够有效地提高方言口音的识别性能。上述的方法均能够提高语音识别系统在实际应用环境下的性能，但还不能根本解决，因此语音识别的鲁棒性问题是目前的研究热点之一。目前有些研究人员在研究人的听感知机理，也就是希望搞清楚人是如何听懂语音的，希望将来这方面的研究成果能够对语音识别带来帮助。

二、语种识别技术

语种识别的目标是自动提取音频中的语种信息，也就是识别出音频中所用语言的种类。语种识别技术按照识别任务的不同可以分为：语种辨识（Language Identification）和语种确认（Language Verification）。语种辨识是指辨识待测输入语音属于目标语种中的哪一种语言，是个多择一的问题。而

语种确认则是判断待测输入语音是否属于某一特定语种，是个二择一的问题。它们之间的最根本的差别是决策数目的不同，所以语种辨识的性能会随着系统中目标语种的增多而下降；而语种确认却与系统中目标语种的增多无关。此外对于语种识别技术来说，还存在着开集测试和闭集测试的不同。所谓的闭集测试，就是指待测语音一定是待识别语种中的一个，系统不需要给出拒识；而开集测试，则是指待测语音可能属于待识别语种中的一个，也可能是待识别语种之外的语言，系统需要对此给出拒识。

一个典型的语种识别系统框图如图 2 所示。在近年的研究工作中，人们利用语音识别领域中常用的系统鲁棒性技术、统计建模技术、模式分类技术、假设检验方法等对系统中的每一组成部分都进行了优化，以期达到最优的识别效果。

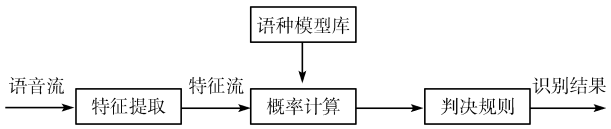


图 2　典型的语种识别系统框图

图 2 中各部分简述如下：

1. 特征参数的选择

对于语种识别系统而言，选择什么样的特征参数是很重要的。通过假定语音信号在短时内是平稳的，运用数字信号分析方法如线性预测、傅里叶变换等求得其频谱，但这些方法没有考虑人的听觉特性。从听觉机理出发，可以较好地表达语音特征的如美尔标度频率倒谱系数，美尔标度感知线性预测系数等要优于简单的线性预测系数。近年来对倒谱特征进行长时的移动差分处理，取得了较好的效果，提高了系统的鲁棒性。在特征提取模块中，要首先对语音信号进行前端预处理，比如：有效语音端点检测技术、说话人聚类技术、语音增强技术等，这是为了解决实际应用中不可避免的环境噪声干扰。

2. 目标语种建模

目标语种建模是指对于每一种待识别的目标语言，预先用该语言的语音数据，在经过特征参数提取后，建立统计模型。在目标语种建模方面，有基于高斯混合模型建模和基于区分性分类器建模的声学层建模方法；另外还有一种方法是采用音子识别器对训练数据进行解码，用解码得到的音子串建立对应语种的语法模型，这种方法也是目前大部分研究机构的主流技术。本文统一将基于音子识别器的语种识别系统命名为基于语法建模的语种识别系统。相比较声学建模系统来说，语法建模系统受语音中环境噪声的影响相对较小，系统鲁棒性较好。

3. 判决规则

语种识别系统的输出一般都是根据输入的语音，对各个目标语种相对应的声学模型进行匹配打分，这些得分都是通过将测试语音同模型库进行似然概率计算或者进行距离度量所得到，这些得分正是判决的依据。同说话人识别系统类似，语种识别系统的判决模块采用的是基于假设检验的思想，系统的性能通过门限来调节。在此过程中，语种识别系统可能发生两类错误：一是待识别目标语种被错误拒绝；二是非目标语种被错误接受。

语种识别技术经过了几十年的研究，尤其是近十年美国国家标准和技术局（National Institute of Standards and Technology, NIST）组织的语种识别技术评测，使得技术本身有了长足的进步，并且已经在国家安全保密和军队国防建设上做辅助人力的工作，在民用方面也在基于内容的音品信息检索方面得到了应用。目前自动语种识别技术离人们的实际需求仍存在着较大的差距，尤其是在复杂信道情况下（如手机、IP 网络、短波广播等），仍然存在着许多很具有挑战性的难题。

三、说话人识别技术

说话人识别是以话音来自动辨认、获取和验证说话人身份的技术，又常称为话者识别或声纹识别。说话人识别技术的研究始于 20 世纪 60 年代，经过 40 多年的研究，得到了长足的发展，并已经出现了一些成熟的应用。与语音识别不同，说话人识别不需要识别所说的内容信息，只需要识别出说话人的身份，因此说话人识别可以按照应用任务分为说话人辨认和说话人确认两种。说话人辨认判断某段语音是若干说话人中哪一个说话人说的，是个多选一的问题。而说话人确认是确定某段是不是某一个特定人所说，给出的判决只能是真或假，是个二元判决问题。另外按照识别对象的不同，说话人识别还可以分为文本固定、文本无关和文本提示三类。文本固定的说话人识别系统正确率最高，但说话人注册和验证时都必须按照规定的固定内容发音。文本无关的说话人识别系统对说话内容没有任何限制，说话人可以自由讲话，也可以说方言或者外语，其

难度远远超过了文本固定的说话人识别任务，但应用范围也得到了很大的扩展，因此文本无关的说话人识别技术仍是目前的研究重点。

说话人识别技术是集声学、语音学、语言学、计算机、信息处理和人工智能等诸领域的一项综合技术。说话人识别的实现依据是每个人都有自己独一无二的发音特征，首先每个人的发音器官存在着差异，例如在声带和声管上的差异，另外每个人的发音习惯也有差异，包括方言、口音、土语、节奏和常用词等。这些发音习惯和发音器官的差异都以复杂的形式反映在说话人语音的波形中，使得每个人的语音都带有强烈的个人色彩。说话人识别的原理就是提取这些反映这些差异的声学特征，建立相应的数学模型并进行识别。

一般说来，说话人识别主要包括两个阶段，即训练阶段和识别阶段。训练阶段，根据话者集中的每个说话人的训练语料，经特征提取后，建立各说话人的模板或模型。说话人模型的建立方法主要有：模板匹配法（Template Matching）、神经网络（Neutral Netowrk）、隐含马尔可夫模型（Hidden Markov Models），其中常见的是使用单状态 HMM 模型，即 Gaussian Mixture Models（GMMs）。由 GMM 方法扩展的背景模型（UBMs）+GMMs 方法是现在文本无关说话人识别的主流方法。识别阶段，由待识别人说的语音同样经特征提取后，与系统训练时产生的模板或模型进行比较。在说话人辨认中，取与测试语音相似度最大的模型所对应的说话人作为识别结果；在说话人确认中，则通过判断测试音与所声称说话人的模型之间的相似度是否大于某一判决门限，做出确认与否的判断。因此说话人辨认和说话人确认仅在判决策略上有所不同。

四、基于哼唱的音乐检索技术

在生活中你可能会碰到这样的场景：嘴边哼起一首歌曲，却怎么也想不起来歌名和歌词。此时你可能就会想，有没有什么方法和途径，仅仅哼唱一段旋律就能够查出这是哪一首歌。基于内容的音乐检索是近年来一个越来越受到关注的研究方向。哼唱检索就是一种基于内容的检索方式，它不再需要用户输入歌名、歌手、歌词等文本信息来检索，只要哼唱歌曲中的某一个片段，根据哼唱的旋律就可以在系统中有效检索所想要查找的歌曲。

基于哼唱的音乐检索系统采用的技术方法都不尽相同，但系统在总体上都包含旋律提取、旋律搜索和旋律数据库三个部分，如图 3 所示。对于输入的一段哼唱语音，旋律提取模块对输入语音进行一系列时域和频域上的处理，得到所需要的旋律特征。其中最重要的是提取出哼唱语音的基频值，以及切分得到哼唱旋律中的音符序列，并以恰当的方式表示旋律特征。以此旋律特征在旋律数据库中一一定位旋律片断可能出现的位置，然后计算哼唱旋律与歌曲旋律的相似度，最后给出旋律最为相似的歌曲列表作为识别结果输出。

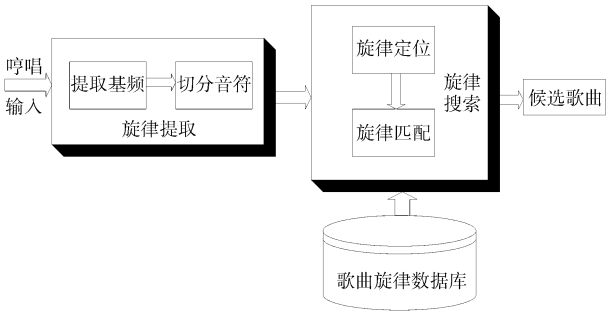


图 3　哼唱检索系统的整体框架

五、音频识别和检索技术的实际应用

目前，音频信息识别与检索技术已经越来越广泛地应用到了人们日常生活中的各个领域。运用音频识别与检索技术可实现对非结构化多媒体信息的自动检索，从而彻底改变目前只能对文本进行检索的现状；在国家信息安全领域，可以用来自动过滤、监测互联网或电信网上的特定语音信息或者特定说话人的信息；在电信的信息自动查询服务中，利用自动语音识别技术可以使用户用语音来代替传统的按键实现自动查询，提供个性化、趣味化的语音增值服务；此外可以用语音命令控制各种设备操作，替代原来的手工操作，例如在一些工作环境恶劣、对人身有伤害的地方（如地下、深水、辐射及高温等）或手工难以操作的地方，均可通过语音发出相应的控制命令，让设备完成各种工作。总之，随着技术的不断成熟以及计算机硬件水平的发展，可以预见，在不久的将来，该技术将迅速走进大众的生活，它将改变人们学习、工作和生活娱乐的方式，从而产生巨大的经济效益和社会效益。

（中国科学院声学研究所　100080）