摘 要

目前，自动语音识别技术已经进入了一个由实验室到实用化的高速发展时期，基于云计算技术的语音识别系统也在嵌入式平台上得到了较好的应用。然而，大多数实际环境并不能满足基于云的系统要求，如何构建一个基于嵌入式平台的语音识别系统仍是当前语音识别技术研究的主要方向之一。考虑到嵌入式平台和 PC 机性能上的差距，以及不同的语音识别系统对实际噪声环境的适应要求，本文针对嵌入式平台语音识别系统的构建需要，主要从以下几个方面展开研究工作：

第一，广泛了解和分析了语音识别技术的发展过程、技术难点，提出了在特征的噪声鲁棒性和更快速的解码网络上展开研究工作。

第二，对如何构建一个完整的语音识别系统进行分析，对比在语音识别过程中不同层次的主要技术和方法，分析选择最适合当前目的的技术，并分别从信号层、特征层和模型层研究了语音识别的噪声鲁棒性和语音增强的技术及方法。

第三，用基于时域的 GFCC 特征替代传统的频域上的 MFCC 特征。这两种特征都是基于人类听觉感知系统的特征，而用时域分析取代频域分析，用离散余弦变换（DCT）替代快速傅里叶变换（FFT），大大减少了计算量；在同一嵌入式设备上，采用 GFCC 特征的识别任务的实时性更高，速度因此也更快。同时，实验表明，基于时域 Gammatone 滤波的 GFCC 特征在大多数噪音环境下，比 MFCC 具有更强的鲁棒性。

第四，构建了基于加权有限状态转换的解码图来完成对识别的解码操作。将加权有限状态机理论引入语音识别，用加权有限状态转换器构建词图，通过对模型的平滑和压缩处理，对词图的剪枝操作，更够压缩整个系统的大小，并保证识别性能维持在一个较高的水平，解码速度也能相应的提高。

关键词：语音识别； GFCC；鲁棒性；加权有限状态转换器

Abstract

Now the technology of automatic speech recognition has entered a period of rapid development from the laboratory to practical, and a large number of cloud-based speech recognition system achieved good results in the embedded platform. However, most of the physical environment can not offer the requirements of cloud-based system, so it is still the focus of current research that how to build a speech recognition system based on the embedded platform.

Consider the performance gap of embedded platform and PC, as well as the recognition system to adapt the requirements of a variety of actual noise environment, a speech recognition system transplantation on embedded platform is constructed from the following aspects:

First, understanding and analysis extensively of the development of speech recognition technology and the difficulties, then commence the study in the features of noise robustness and faster decoding network.

Second, analysis how to build a complete speech recognition system, contrast the main technology and methods in the different levels of speech recognition processing, and the choose the most suitable technical for current purpose, and respectively, research the technologies and methods of the noise robustness and speech enhancement from the signal layer, the feature layer and the model layer.

Third, we replace the traditional MFCC feature on frequency domain by the GFCC feature of the time domain. These two features all based on the human auditory perceptual system, we can greatly reduce the computation amount through replace the frequency domain analysis by the time domain analysis and using discrete cosine transform(DCT) instead of the fast fourier transform(FFT) ; at the same embedded equipment, GFCC has greater performance in the real-time recognition task, therefore faster than MFCC. Meanwhile, experiments results show that GFCC based on time-domain Gammatone filter is more robust than MFCC in most noisy environments.

The fourth, we construct the decode-graph based on the weighted finite-state transducer to complete the decoding operation. Introduction the weighted finite-state machine theory in ASR, we use weighted finite-state transducer build a word graph. By smoothing and compression the models and pruning the word graph, we can enough to compress the size of the entire system and to ensure the recognition performance is

maintained at a high level, the decoding speech can also be a corresponding increase.

Key Words: Speech Recognition; GFCC; Robust; Weighted Finite-State Transducer

目 录

第 1 章 绪论

在文字产生之前，人类已经开始用语音来进行交流，即使在文明高度发达的今天，语音交流仍然是人类交流最主要的模式。从计算机的发明开始，人们就憧憬着有一天能够实现人与机器的语音信号交流，而不满足于传统的鼠标、键盘的输入，因而语音识别技术的研究应运而生。

语音识别是一门交叉学科，它涵盖了包括信号处理、模式识别、人工智能、生理学、概率统计和随机过程等等在内的大量研究领域。

近二十年来，在语音识别技术领域取得了大量的成果，语音识别技术开始从实验室走向商业应用。未来十年，语音识别技术将大量应用于家电、工业生产、通信服务、汽车电子、消费电子产品、医疗等各个领域，语音识别技术的应用已经成为一个具有高竞争性的高新技术产业。

1.1 语音识别技术研究现状[1]

语音识别的研究最早开始于 1952 年，AT&T 贝尔实验室的的 Davis 等人把语音信号的第一、第二共振峰作为特征参数，实现了第一个可以识别十个英文数字的语音识别系统 Audry System[2]。

20 世纪 50 年代末 60 年代初，随着数字集成电路的出现，语音数字信号处理也因此产生，这是计算机语音识别技术的开端。快速傅里叶变换（FFT，Fast Fourier Transform）在频谱分析中得到广泛应用，人们借此开始研究语音信号的内部本质。

进入 70 年代后，美国国防部高级研究计划署提出了语音理解研究计划并推动了该计划的展开，吸引了众多的工业界和学术界的研究机构，为语音识别领域注入了更多的新鲜血液，这全面推动了语音识别技术的发展。Baum 等人首次系统阐述了隐马尔可夫模型（HHM，Hidden Markov Model），并将其引入语音识别领域。至今为止，HMM 算法仍是语音识别领域最好的算法之一。在这一时期，线性预测参数（LPC，Linear Predictive Coefficient）[3]被提出并与动态时间规整（DTW，Dynamic Time Warping）[4]技术和模式识别[5]方法一起，实现了特定人孤立词语音识别系统。

80 年代，实验室语音识别技术的研究取得巨大突破，研究重点也由孤立词向连续语音识别发展。贝尔实验室 Rabiner 等人不遗余力的对 HMM 模型的研究和推广[6]，使得基于统计概率模型的方法开始在语音识别领域得到广泛应用。1988 年 CMU 采用 VQ/HMM 实现的 Sphinx 系统[7]，是第一个高性能的非特定人连续语音识别系统。

20 世纪 90 年代，随着各种规模的著名语音识别任务的发布和标准数据库的建立，各个研究机构的识别技术有了一个客观比较的平台。在对这些标准数据库的测

试比对取得较好的基础上，IBM、CMU、AT&T 等都将语音识别技术推入了商用领域。其中 IBM 公司推出的 ViaVioce 系统，是具有代表性的汉语大词汇连续语音识别系统，该技术应用于听写机、电话网和语音信息查询服务系统等领域。而剑桥大学推出的 HTK 工具包[8]，也使得研究语音识别的门槛大大降低，大量研究机构的加入掀起了语音识别领域研究的又一波高潮。

进入 21 世纪后，语音识别技术已经广泛应用于商业用途。在半导体技术飞速发展的前提下，嵌入式技术也得到了显著的发展，语音识别不再局限于计算机平台，开始大量进入移动设备领域。从早期的单片机，到后来的 MCU、DSP 和专用语音识别芯片的出现，都为嵌入式语音识别技术的研究和发展提供了平台；而现在，在小型化、高性能的微处理器的普及和云计算服务、无线通信技术的支持下，手机平台的语音识别应用已经得到普及，基于本地语音识别和云计算服务的应用方式开始推广，这其中最成功的例子就是 Apple 公司的 Siri 系统。而随着图形处理器（GPU，Graphics Processing Unit）性能的提高和在某些领域对数字信号处理器（DSP，Digital Signal Processor）的替代，基于深度学习（Deep Learning）的深度神经网络（DNN，Deep Neural Network）[9]也成为当前语音识别最前沿的研究方向之一。

回顾语音识别发展的几十年，可以用“日新月异”来形容：从最初的音素识别到当前的大词汇连续语音识别，各种新技术不断涌现，识别性能不断提升，应用范围不断扩展。但是我们也要清楚的认识到，当前的语音识别技术和我们想象中的还有一定距离，如何真正实现人与机器之间畅通无比的语言交流，推动语音识别技术的全面实用化，将是我们需要面对的困难和研究的方向。

1.2 语音识别系统类型

根据对说话人说话方式的要求，语音识别系统可以分为 3 大类：孤立字（词）识别，关键词检出以及连续语音识别系统。孤立字（词）的识别对象为一个字、词或者是一个短语，对每一个对象都训练出一个模型，并组成词汇表，如“一”、“二”、“开门”等等；关键词检出的识别对象为连续的语音信号，但只对该信号中的某一段或几段信号进行识别；连续语音识别则是对任意的一句话或一段话进行识别。

根据对说话人的依赖程度，语音识别系统可以分为特定人和非特定人语音识别系统。其中，特定人语音识别的训练模型只针对于某一个人，当其他人使用该系统时，需要对这个人重新训练模型才能完成识别任务，这种系统可以应用在某些对安全性要求较高的领域。非特定人语音识别则适用于某一范畴的说话人的识别任务，如英文、中文、方言等等，通过对该范畴内的多个说话人的语音训练出模型，识别对象包括训练模型说话人在内的该范畴的所有说话人。相对于特定人识别系统，非

特定人识别系统更能够满足实际应用的需要，但其需要的训练时间更长、训练资源更多、识别起来也更加的困难。

根据识别词汇量的大小，语音识别系统又可以分为小词汇量、中词汇量、大词汇量以及无限词汇量语音识别系统。

此外，根据语音设备和通道的不同，语音识别系统还可以分为桌面（PC）语音识别、电话语音识别和嵌入式设备（手机、平板、PDA 等）语音识别。

虽然根据分类准备的不同而有各种不同类型的语音识别系统，但是其在基本原理和技术上是相似的。一个简单的语音识别系统原理图如图 1 所示。

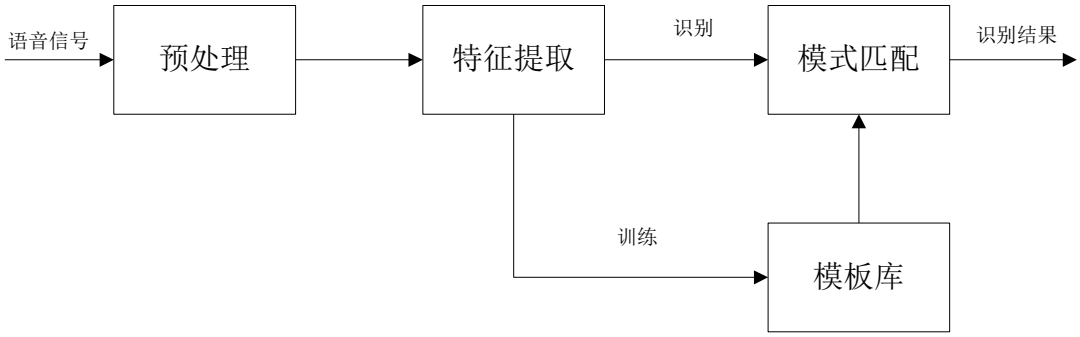


图 1 语音识别基本原理图

1.3 语音识别技术的难点

语音识别的最终目的是让机器能听懂人的语言，真正的实现人机对话。而这却又是十分的困难，主要原因是：

1.语音识别系统的适应性差，对环境依赖性强，要求测试条件和训练条件保持一致，否则系统性能会大大下降；

2.高噪声环境下识别困难，特别是在车载条件下，人的发音变化大，像发音失真、发音速度和音调的改变等等，即所谓的 Lombard 效应；

3.端点检测的不确定性，即使在安静的环境下，语音识别系统一半以上的识别错误来自错误的端点检测；

4.因为汉语的语言特点，使得汉语的语言信息处理更为困难和复杂。包括汉语的字词不分、同音字词、语义的表述等等；

5.语音识别系统从实验室演示到实用化的转换中，还存在着大量的问题，比如识别速度、拒识问题和关键字（词）检测技术等等。

当前语音识别技术的应用可以分为两个发展方向：一个方向是大词汇量连续语音识别系统，其平台为计算机，主要应用于听写机以及与电话、网络结合的语音查询服务；另一个重要的发展方向是在小型化、便携式设备上的应用，如手机、平板电脑、汽车电子设备、智能家电和玩具等等，这些都通过专门的硬件系统实现。以

计算机为平台的语音识别系统计算性能高，存储空间大，工作环境相对安静，系统的识别性能很高，而在这些方面，基于嵌入式设备的语音识别系统则面临着更大的困难：

1.实时性。移动设备对识别任务的实时性要求更高，在相对计算资源受限的情况下，要求计算量小、计算速度快的处理方法。

2.存储空间。即使当前的移动存储技术更先进，但相对 PC 来说，嵌入式设备的存储资源仍较小，这就需要训练模型占用的空间更少。

3.鲁棒性。嵌入式语音识别的应用环境五花八门，需要有很强的语音增强技术，能够减少噪音对识别性能的干扰。

4.自然性。语音识别系统要让用户感觉到是在跟人对话，这就需要系统允许用户以各种自然句式发布命令，这样就要采用有限状态语法网络、对话管理、统计语言模型和关键词检出等技术，来满足用户的自然对话需求。

5.自学习或自适应。包括自动适应用户的口音和说话习惯。这要求对声学模型和语言模型有自适应技术，要求优化模型的架构和管理程序以满足嵌入式系统的需要。

1.4 选题背景及意义

近年来，以手机等为代表、基于可移动嵌入式设备的语音识别技术的研究已经成为一个热点，并且以本地语音识别为主、辅以云计算服务的语音识别方式也进入了市场化阶段；而然，由于各种应用环境中噪声的影响、无线通信网络的限制，如何在性能有限的嵌入式设备上构建一个本地的、噪声鲁棒的、高效的语音识别系统仍是当前研究的重中之重。

在前人对语音识别中噪声鲁棒性技术的研究基础上，本文从语音特征的角度出发，选取用基于 Gammatone 滤波的 GFCC 特征作为语音识别中的特征。实验证明，与传统的 MFCC 特征相比，模拟人类听觉感知系统设计的 GFCC 特征对噪声有更强的区分性，在静音和多种带噪语音的环境中，GFCC 均有高于 MFCC 的识别性能；而在时域上的 GFCC 特征提取与频域上的 MFCC 提取方式相比，计算量更小，能够节省设备资源，更适合于嵌入式语音识别的任务要求。

在 Mohri 等研究者对加权有限状态转换器（WFST，Weighted Finite State Transducer）的先期研究工作的铺垫下，目前主流的大词汇量非特定人连续语音识别系统均采用 WFST 框架。在该理论框架下，语音识别中各层次的模型和知识被转换成 WFST 的形式，并通过加权有限状态机理论中的组合操作，将模型和知识整合成完成的解码网络；而最小化操作又能去除冗余，最大程度的压缩网络的规模。在加

权有限状态机的理论和操作下，我们可以得到一个完整的、高效的、单阶段的 Viterbi 解码静态搜索网络。而通过将其它知识表达成 WFST 并组合到解码网络中的操作，可以解决特定的问题或提高整个系统的识别性能。国外研究结果表明，与传统的两阶段识别系统（2-pass）相比，在优化后的静态网络上的单阶段识别系统（1-pass）更具有竞争力。

综上，本文中构建了一个以 GFCC 为语音识别特征、以 WFST 为理论基础的语音识别系统，实验测试了该系统在噪声环境下的性能，并对其在嵌入式设备上的移植和应用做出分析和总结。

1.5 论文结构安排

本论文主要内容安排如下：

第 1 章为绪论，主要介绍了语言识别技术的发展情况、语音识别系统的分类和当前语音识别技术的难点，特别是在当前语音识别技术向嵌入式系统移植的趋势下的研究方向。

第 2 章介绍了语音识别过程中各处理环节的一些主流技术方法，包括对语音信号的前端处理、语音信号的特征处理方法、声学模型和语言模型的训练与优化处理等等，着重介绍了在本文中所采用的方法并和其它方法作比较分析。

第 3 章详细阐述了 Gammatone 滤波器组的滤波原理，以及基于 Gammatone 滤波的 GFCC 特征的时域提取方法。

第 4 章介绍了加权有限状态机理论以及加权有限状态转换器在语音识别任务中的应用和处理方法，并描述了在 Kaldi 工具包下用加权有限状态转换器构建一个完整的语音识别解码图和对本文中所采用的各层次知识源的组合优化操作。

第 5 章是实验设计和结果分析，设计在 Linux 环境下的一个完整的语音识别解码过程，对 GFCC 和 MFCC 在噪声语音环境下的识别性能进行对比，研究 GFCC 特征的噪声鲁棒性。

第 6 章是总结和展望，对论文的研究工作和结论进行总结，点明优势，指出不足，并提出下一步的研究和工作方向。

第 2 章 语音识别技术与噪声鲁棒性技术研究

语音识别是一门新兴学科，它在发展的过程中不断借鉴和融入其它学科的理论和方法，形成了一门涵盖数字信号处理、声学、生理学、语言学、模式识别、通信理论、计算机科学等多门学科的综合性学科。而在语音识别领域，研究者针对不同的侧重点也进行了一系列研究；提高语音识别系统在噪音环境下的识别性能，增强语音识别系统的噪声鲁棒性，也是语音识别技术中一个重要的研究方向。

在本章中，主要介绍了从前端处理、特征提取到识别的整个语音识别系统的理论和方法；并针对大词汇量连续语音识别任务，着重介绍了本文所构建系统中采用的技术，及与其它技术和方法的分析比较。

2.1 语音识别系统框架

对于不同的识别任务，语音识别系统会不同，但基本技术和处理流程大致上是相同的。一个典型的语音识别系统框架如图 2.1 所示。

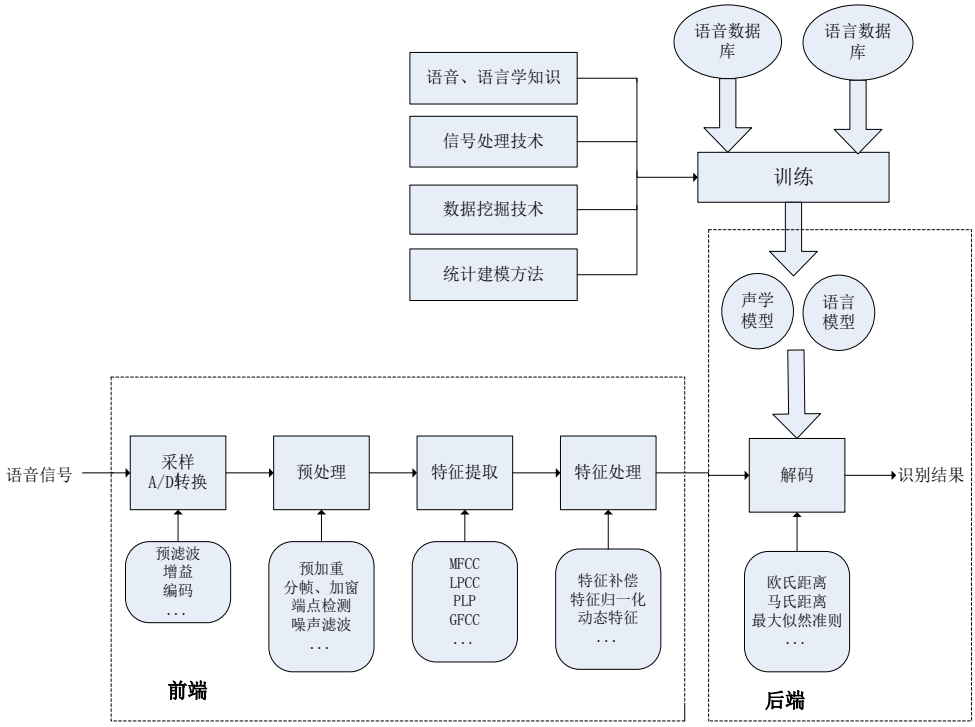


图 2.1 语音识别系统框架

语音信号通过麦克风采集，经过采样和 A/D 转换后由模拟信号转变为数字信号。然后对语音的数字信号进行预加重，分帧，加窗，端点检测和滤波等处理。

预处理过后的语音信号将按照特定的特征提取方法提取出最能够表现这段语

音信号特征的参数，这些特征参数按时间序列构成了这段语音信号的特征序列。

在训练过程中，获得的特征参数通过不同的训练方法获得模型，而后存入模板库；在解码过程中，新采集的语音信号经过处理获得特征参数后，与模板库中的模型进行模式匹配，并结合一些专家知识得出识别结果。

2.2 采集和预处理

2.2.1 采样和量化

在语音信号的采集过程中，麦克风将声音从物理状态转化为模拟的电信号，我们需要把连续的模拟信号转化为时间上离散、但幅值上仍连续的离散模拟信号，这一过程就是采样。在采样过程中，根据采样定理，采样频率 f_s 必须是声音最高频率的 2 倍以上。采样频率越高，数字化后的声波的保真度就越高，但相应的信息的存储量就越大。人耳所能接收到的声音频率范围约为 20Hz~20kHz，通常在 PC 机上的采样频率为 16kHz，嵌入式设备上为 8kHz。

为了便于计算机计算、传输和存储，采样后的信号还要转化为能够用二进制表示的离散值，这一过程就称为 A/D 转换。为了确保系统处理结果的精确度，我们必须保证 A/D 转换具有足够的转换精度。通常采用的方法是均匀量化和脉冲编码调制（PCM，Pulse Code Modulation），当前语音识别中常用 16bit 量化。

2.2.2 预加重、分帧和加窗

对语音信号进行采样处理后，还要进行一些预加重。由于受到口鼻辐射和声门激励的影响，语音信号的高频部分在 800Hz 以上会有-6dB/倍频程的跌落，因此预加重的目的就是提升语音信号的高频部分，使频谱平滑。一般预加重通过一个一阶高通滤波器实现，其表达形式为：

$$H(z)=1-uz^{-1} \tag{2.1}$$

其中 u 值接近于 1，典型取值范围为 0.95~0.97。

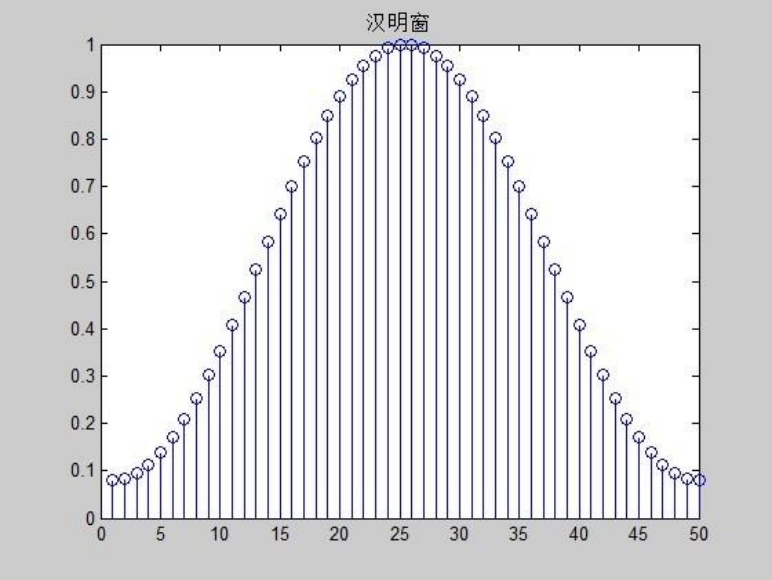
数字化的语音信号是一个不平稳的时变信号，为了便于分析，通常假设语音信号在 10ms~30ms 内是短时平稳的，我们所有的分析工作都是在这个假设基础上进行的。因此，在对语音信号进行分析前，需要对其进行分帧，通常将语音信号的每帧长度设为 20ms，相邻两帧之间有 10ms 的重叠。

为了实现分帧步骤，我们要对语音信号进行加窗操作。不同的窗口选择对语音信号分析的结果会产生影响。最简单的窗函数为矩形窗，即

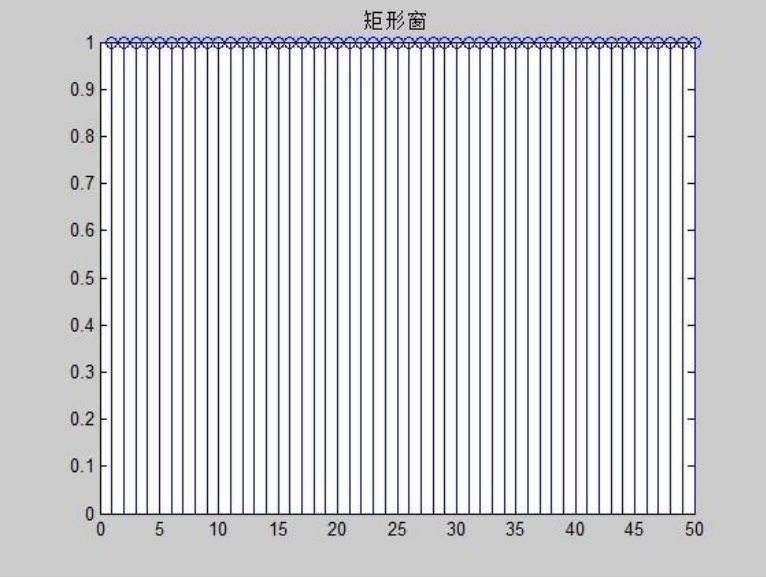
$$w(n)=\begin{cases}1, 0\le n\le N-1\\ 0, \text{其他}\end{cases} \tag{2.2}$$

其中 N 为帧长。通常我们选择的窗函数为汉明窗（Hamming Window），其定义为：

$$w(n)=\begin{cases}0.54-0.46\cos\dfrac{2\pi n}{N-1}, 0\le n\le N-1\\ 0, \text{其他}\end{cases} \tag{2.3}$$



（a） 汉明窗



（b）矩形窗

图 2.2 窗函数波形

选择汉明窗能够减小帧起始和结束处信号的不连续性，避免采用矩形窗带来的 Jibbos 现象，因此在本文的特征提取中，均采用汉明窗。

2.3 语音信号分析方法

2.3.1 语音信号时域分析方法

在信号分析时，最自然最直接的方法就是以时间作为要分析函数的自变量。典型语音信号特征是随时间变化的，本节简单介绍了语音信号基于短时分析的几种时域分析方法。

2.3.1.1 短时能量分析和短时过零率

短时能量分析对语音信号能量的时间变化趋势有一个合理的描述。对信号$\{x(n)\}$的短时能量定义如下

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 = \sum_{m=-\infty}^{\infty} x^2(m)h(n-m) = x^2(n) * h(n) \tag{2.4}$$

其中，$w(n)$为窗函数，$h(n)=w^2(n)$，E_n为从第 n 个点开始的短时能量。

短时能量在对语音信号的分析中的作用：首先能够区分清音和浊音，因为通常情况下浊音比清音具有明显更大的能量；其次能够用来进行端点检测，区分静音段和声音段，或者用来判定声、韵母或连字的分界。

由于对信号的平方运算人为增加了高频信号和低频信号的差距，因此在某些场合可能会造成更大的误差。为了解决这个问题，最简单的方法是用短时平均幅值的变化来表示能量的变化。

短时平均过零率（ZCR，Zero Crossing Rate）是指短时间内信号通过零值的次数，具体于连续信号即其波形通过 x 轴的次数，离散信号即采样符号变化的次数。对于第 n 帧语音信号，其过零率为

$$ZCR_n = \frac{1}{2}\sum_{m=n}^{n+N-1} \left|\mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]\right| \tag{2.5}$$

其中 sgn 是符号函数，即

$$\mathrm{sgn}[x(n)] = \begin{cases} 1, x \geq 0 \\ -1, x < 0 \end{cases} \tag{2.6}$$

短时过零率在一定程度上能够反映频率的高低，浊音的过零率较低，清音的过零率相对较高，因此可以用来初步分析清、浊音。短时过零率容易受到低频的干扰，通常我们在处理中还会加入门限值，即将波形穿过零点的次数改为越过门限值的次数，以此来增强抗干扰能力。

在语音信号处理中，常将短时平均能量和短时平均过零率结合起来进行语音段起始点的检测，即端点检测。当背景噪声较小时，用短时平均能量的方法比较准确，

当背景噪声较大时，采用短时平均过零率能够获得比较精确的结果。

2.3.1.2 短时自相关函数和短时平均幅度差函数

自相关函数是一种描述信号本身的周期性和同步性的便利方法。对于离散的语音数字信号，其自相关函数有如下定义：

$$R(n)=\sum_{m=-\infty}^{+\infty}x(m)x(m+n) \tag{2.7}$$

在信号的第 k 个样本点附近用窗函数截取一帧信号并计算自相关函数，得到这一帧信号的短时自相关函数，其结果如下：

$$R(n)=\sum_{m=k}^{k+N-n-1}x(m)x(m+n) \tag{2.8}$$

短时自相关函数可以用来确定一个浊音的基音周期。若$x(n)$是一个浊音周期信号，根据自相关函数的周期性，其短时自相关函数的周期性与原信号相同。若$x(n)$是一个清音信号，由于清音类似于一个随机噪声，不具有周期性，也就无法确认基音周期。

大量的乘法计算使得短时自相关函数的计算量非常之大。在某些计算资源受限的条件下，为了避免这种情况，常常采用短时平均幅度差函数。对于一个周期为 T 的信号 $x(n)$，对其做如下的差值处理

$$d(n)=x(n)-x(n-m) \tag{2.9}$$

当 $m=0,\pm T,\pm 2T,\cdots$ 时，差值为零，即 m 为周期的整数倍时，短时平均幅度值最小，其函数有如下定义

$$\gamma(\mathrm{n})=\sum_{m=k}^{k+N-n-1}|x(m+n)-x(m)| \tag{2.10}$$

$\gamma(n)$也呈现出周期性，且在周期的整数倍点处具有谷值。$\gamma(n)$在浊音的基音周期上会急剧下降，而在清音中不会有明显的变化。因此，短时平均幅度差函数和短时自相关函数一样，也可用于基音周期的检测，并且比短时自相关函数的计算量更小。

2.3.2 语音信号频域分析方法

频域（Frequency Domain）是描述信号在频率方面的特性时建立的一种坐标系。在频域进行分析，无需求解微分方程，能够间接揭示系统性能和指明改进方向，易于实验分析，在某些非线性系统和抑制噪声的系统上应用广泛。在语音信号分析中，常用的频域分析方法有滤波器组和傅里叶变换的方法。

2.3.2.1 滤波器组方法

用滤波器组的方法分析语音信号的频谱是最早的频谱分析方法之一，并且在现在也经常被使用。当采用宽带带通滤波器时，频率分辨率较低，其与加窗处理中窗口较短时的处理结果相近；采用窄带带通滤波器时，频率分辨率较高，与窗口较长时的处理结果相近。

通常用一组滤波器组对语音输入信号进行滤波处理，分离出输入信号中不同中心频率的分量，再进行各种分析和处理。

2.3.2.2 傅里叶分析

傅里叶变换最早在 1807 年由法国科学家 J. Fourier 提出，并在很多领域得到了广泛应用。而傅里叶分析以傅里叶变换为基础，是语音信号处理中一个非常重要的工具。

我们定义短时傅里叶变换

$$X_n\left(e^{jw}\right)=\sum_{m=-\infty}^{\infty}x(m)\omega(n-m)e^{-j\omega m} \tag{2.11}$$

其中，$\omega(n)$为加窗函数。

为了能够让计算机对数据进行分析，通常用离散傅里叶变换代替连续傅里叶变换。但是随着技术的发展，傅里叶变换的一些局限性也渐渐体现出来：首先，傅里叶变换的时间分辨率为零，不能反映信号在时域上的信息；其次，傅里叶变换是基于信号是平稳的这个假设，而在实际生活中，很多声音信号是非平稳的；最后，傅里叶变换在整个频段内的分辨率都是相同的,不能反映信号在某一频段的某种变化。同时，将声音信号进行频率分析，计算量较大，在对实时性要求高而计算资源又受限的嵌入式设备上也是一个难题。

2.3.3 其他分析方法

为了分析和处理非平稳信号，研究人员提出并发展了一系列信号分析理论：

1．时频分析，即时频联合域分析（JTFA，Joint Time-Frequency Analysisi）。时频分析方法提供了时间域与频率域的联合分布，清楚地描述了信号频率随时间变化的关系。其基本思想是，设计时间和频率的联合函数，用它同时描述信号在不同时间和频率的能量密度或强度。用时频分布来分析信号，能够给出各个时刻的瞬时频率及幅值，并且能够进行时频滤波和时变信号研究。当前主要的时频表示方法有线性时频表示、二次时频表示及其他表示方法。

2.Gabor 变换。当信号处理时窗口函数为高斯函数时，此时的短时傅里叶变换

就称作 Gabor 变换。

3.小波变换（Wavelet Transformation）。小波的振幅振荡是正负相间的，小波理论用多分辨率分析的思想对时频空间进行非均匀划分，时信号在一组正交基上进行分解，是对非平稳信号分析的一种新途径。在语音信号处理中，利用小波变换模拟人类的听感知系统，能够对语音信号进行去噪处理和判断清音与浊音。

2.4 声学特征选择

语音识别发展的过程中，人们研究和发展了很多特征，这些特征在不同应用中起到了不同作用。如何在大量的特征参数中选择出少数具有互补作用的特征参数，既是一个理论问题，又是一个实际应用问题[10]。通常我们将声学特征分为两大类，一类为基于人类发声机理的特征，另一类为基于人耳听觉感知的特征，而这两类具有代表性的特征分别是线性预测倒谱系数（LPCC，Linear Prediction Cepstrum Coefficient）和 Mel 频率倒谱系数（MFCC，Mel Frequency Cepstrum Coefficient）。

2.4.1 线性预测倒谱系数

将语音信号 $x(n)$看作一个全极点模型 $H(z)$在激励 $u(n)$下的输出，该系统的传递函数如下

$$H(Z)=\frac{G}{1-\sum_{k=1}^{p}a_k Z^{-k}} \tag{2.12}$$

其中，G 为增益，a_k 为实数，p 为模型阶数。该模型是一个全极点模型，a_k 和 G 为模型参数。可以定义一个 p 阶线性预测器 $P(Z)$

$$P(Z)=\sum_{k=1}^{p}a_k Z^{-k} \tag{2.13}$$

则经过预测器输出的语音序列 $\tilde{x}(n)$为

$$\tilde{x}(n)=\sum_{k=1}^{p}a_k x(n-k) \tag{2.14}$$

因为 $P(Z)$由 $\{a_k\}$构造成，所以预测器 $P(Z)$为一个最佳预测器，根据最小均方误差准备，此时预测器的误差短时总能量达到最小值。

引入预测误差 $e(n)$，有

$$e(n)=x(n)-\tilde{x}(n) \tag{2.15}$$

定义预测误差逆滤波器 $A(Z)$

$$A(Z)=1-P(Z)=1-\sum_{k=1}^{p}b_k Z^{-k} \tag{2.16}$$

则$e(n)$是输入为$x(n)$时经过滤波器$A(Z)$的输出。当$b_k=a_k$时，有预测误差的 Z 变换恰好等于输入激励信号序列 Z 变换，预测误差序列与声道激励序列一一对应。根据这一性质，由信号$x(n)$开始估计一组线性预测器的系数$\{b_k\}$，即为 LPC 系数。常用来求解 LPC 系数的方法有自相关法、协方差法和格型法等等。

对线性预测分析的合成滤波器

$$H(Z)=\frac{1}{1-\sum_{k=1}^{p}b_k Z^{-k}} \tag{2.17}$$

其冲激响应$u(n)$，根据同态处理方法得到

$$\hat{H}(Z)=\log H(Z) \tag{2.18}$$

$\hat{H}(Z)$可以展开成级数形式

$$\hat{H}(Z)=\sum_{n=1}^{+\infty}\hat{u}(n)Z^{-n} \tag{2.19}$$

其中，$\hat{u}(n)$为$u(n)$的倒谱，所以存在$\hat{u}(n)$为$\hat{H}(Z)$的逆变换。假设$\hat{u}(0)=0$，将上式两边同时对Z^{-1}求导，整理得到

$$\left(1-\sum_{k=1}^{p}b_k Z^{-k}\right)\sum_{n=1}^{+\infty}n\hat{h}(n)Z^{-n+1}=\sum_{n=1}^{+\infty}kb_k Z^{-k+1} \tag{2.20}$$

若上式两边 Z 的各次幂系数相同，则$\hat{h}(n)$和b_k的关系为

$$\begin{cases}\hat{h}(1)=b_1\\ \hat{h}(n)=b_n+\sum_{k=1}^{n-1}\left(1-\frac{k}{n}\right)b_k\hat{h}(n-k), 1\le n\le p\\ \hat{h}(n)=\sum_{k=1}^{p}\left(1-\frac{k}{n}\right)b_k\hat{h}(n-k), n>p\end{cases} \tag{2.21}$$

求出的 LPC 倒谱系数可将其看作是对原信号短时倒谱的近似，因为 LPCC 被认为包含了信号的包络信息。

2.4.2 Mel 频率倒谱系数

MFCC[11]特征是一种基于人类听觉感知特性的特征，模拟了人耳对不同频率的感知程度，其对中低频语音信号较敏感，对高频信息的区分度不大，因而能够从信号的中低频段提取更多的语音信息。

提取一组 MFCC 特征主要有以下几个步骤：

1.首先对输入的语音信号进行预处理，得到分帧和加窗后的时域信号；

2.对时域信号进行快速傅里叶变换（FFT，Fast Fourier Transform），得到语音信号的频率表达；

3.将得到的线性频率转换为 Mel 频率，转换公式如下

$$Mel(f)=2595\times\log_{10}\left(1+\frac{f}{700}\right) \tag{2.22}$$

或者

$$Mel(f)=1125\times In\left(1+\frac{f}{700}\right) \tag{2.23}$$

4.在 Mel 频率轴上构造 M 个三角带通滤波器组，这 M 个三角滤波器在 Mel 频率尺度上是平均分布的，其定义为

$$H_m(k)=\begin{cases}\dfrac{k-f(m-1)}{f(m)-f(m-1)}, f(m-1)\le k\le f(m)\\ \dfrac{f(m+1)-k}{f(m+1)-f(m)}, f(m)<k\le f(m+1)\\ 0,\text{其他}\end{cases} \tag{2.24}$$

其中，m 为滤波器编号$(0\le m\le M-1)$，$f(m-1)$，$f(m)$，$f(m+1)$分别是该滤波器的下限、中心和上线频率。

三角带通滤波器主要有两个目的[12]：一是对频谱进行平滑，消除谐波的影响，突出原始语音的共振峰，因此，以 MFCC 为特征的语音识别系统并不会受到输入语音的音调不同而有所影响；二是降低了信息量。

5.离散余弦变换（DCT，Discrete Cosine Transform）。对每一个滤波器的输出计算其对数能量E_m，并做 DCT 变换

$$C_d=\sum_{m=0}^{M-1}E_m\cos\left[m\left(k-\frac{1}{2}\right)\frac{\pi}{M}\right], d=0,1\cdots,D-1<M \tag{2.25}$$

其中，D 为 Mel 倒谱系数的阶数，其不同分量对应不同的信息。实验表明，最有用的语音信息包含在 MFCC 分量C_1到C_{12}之间，最有用的说话人信息包含在 MFCC 分量C_2到C_{16}之间[13]。

2.4.3 其它特征选择和处理方法

一般除了静态特征（即提取出的基本特征）外，我们还会加入对数能量或由静态特征计算其一阶、二阶差分[14]生成动态特征，并把这些特征拼接起来作为一个多

维的新特征来使用。除了这些方法外，研究者还提出了用特征变换、降维的方法来提高声学特征的区分性[15]，如主分量分析（PCA，Principal Component Analysis）[16]、线性判别分析（LDA，Linear Discriminant Analysis）[17]、异方差线性判别分析（HLDA，Heteroscedastic Linear Discriminant Analysis）[18]等等。

在本文中，我们采用一种基于 Gammatone 滤波的倒谱特征作为静态特征，用 DCT 来去除各维度分量之间的相关性，并拼接其二阶动态特征组成一组特征，这将在第 3 章详细介绍。

2.5 声学模型

在语音识别中，通过特征参数建立并训练语音信号的模型，对识别的语音信号进行模板匹配，是一个语音识别系统中最重要的环节。当前语音识别领域中，常用的主要有基于矢量量化（VQ，Vector Quantization）的识别技术，基于动态时间规整（DTW，Dynamic Time Warping）的识别技术，基于高斯混合模型（GMM，Gaussian Mixture Model）的技术和基于隐马尔可夫模型（HMM，Hidden Markov Model）的技术，而我们将详细介绍最后一种，也是本文中所选用的模型。

2.5.1 隐马尔可夫模型

HMM 最早于 1972 年出现在 Baum 等人的文章中，随后被 CMU 的 Baker 等人和 IBM 的 Bakis、Jelink 等人引入语音识别领域，到 20 世纪 80 年代初美国贝尔实验室的 Rabiner 等人提出将这一方法用于非特定人语音识别，使 HMM 进一步推广开来，并成为公认的最有效的语音识别方法。

HMM 作为语音识别中一种很有效的技术，不仅能用来作为（以音素、音节或词为单位）语音产生的声学模型，而且能作为词法、语法、语义等高层次的语言模型，在很多领域都有应用。

2.5.2 HMM 基本思想

马尔可夫链是马尔可夫随机过程的特殊情况，其状态参数和时间参数都是离散的。而在实际中，观察到的事件与状态并不一一对应，其对应关系通过一组概率分布来描述，这就是 HMM 模型。HMM 是对语音信号的时间序列建立的统计模型，由两个相互关联的随即过程共同描述语音信号的统计特性：一个是用具有有限状态数的马尔可夫链来模拟语音信号统计特性变化的随机过程，它描述状态的转移，另一个随机过程描述状态和观察值之间的统计关系。因此观察者只能看到观测值，不

能直接看到状态，而是通过一个随机过程感知状态的存在，因此这条链为一条“隐”链，而整个模型也称之为“隐”马尔可夫模型，HMM 的组成如图 2.3 所示。

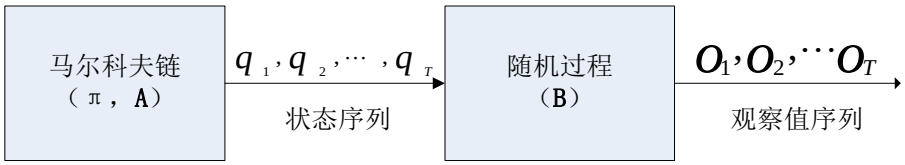


图 2.3 HMM 组成示意图

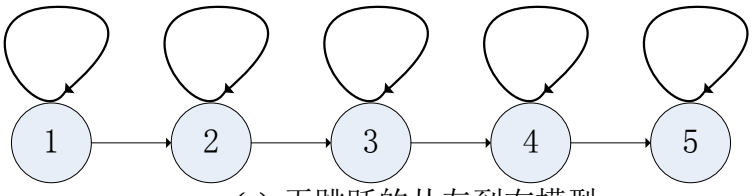
图中，马尔可夫链由初始状态 和状态转移概率矩阵 **A** 描述，产生的输出为状态序列；随机过程由观察值概率矩阵 **B** 描述，产生的输出为观察值序列，T 为观察值的时间长度。因此，一个 HMM 模型可以记为

$$\lambda = (\pi, A, B)$$

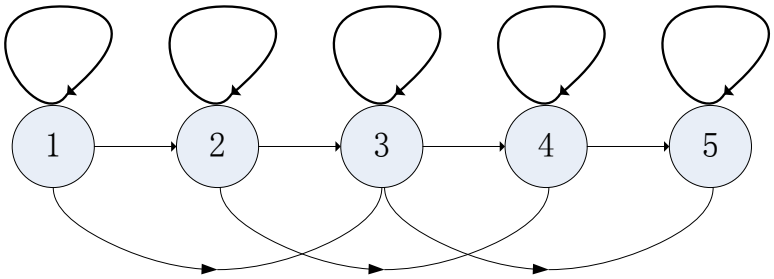
2.5.3 HMM 类型

几种典型的 HMM 模型如图 2.4 所示。图（a）和（b）是两种典型的从左到右 HMM 模型结构，其特点是：1.这种模型结构的状态转移矩阵 ***A*** 被限制为上三角，因此其状态转移被加以适当的限制；2.这种模型的拓扑结构包含了时间信息，因为前面状态的输出观察值必定在后面状态的输出观察值之前，于是使得模型能适应语音的时序性；3.该模型的初始状态始终在第一个状态，并且认为多套训练样本是相互独立的，因此稍加修改即可得到训练算法。图（c）可以从任意状态出发，并在下一刻到达任一状态，其 ***A*** 矩阵无零值。

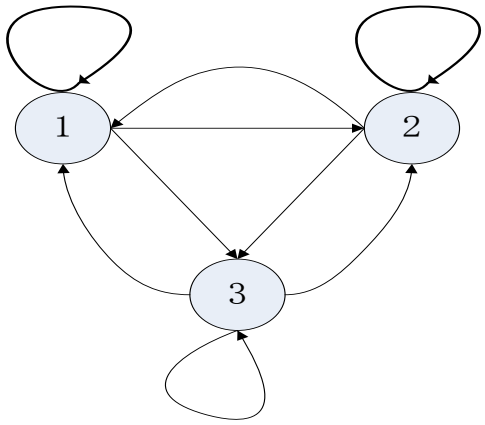
实验表明，状态数目超过 5 对识别率没有改善[19]，很多实验也认为 5~6 个状态的 HMM 足够满足孤立词识别的需要，而对音素或声、韵母，一般 2~3 个状态就比较合适了。



(a) 无跳跃的从左到右模型



（b）有跳跃的从左到右模型



（c）全连接模型

图 2.4 典型 HMM 结构

2.5.4 HMM 训练

HMM 训练的算法有多种，最经典的就是 Baum-Welch 算法[20]。该算法用于解决 HMM 参数估计的问题，用数学方式描述即在已知一个观察序列 $O=o_1,o_2,\cdots o_T$，确定一个 HMM 模型 $\lambda=(\pi,A,B)$ 使 $P(O|\lambda)$ 最大。

定义前向变量

$$\alpha_t(i)=P(o_1,o_2,\cdots o_T,q_t=i|\lambda)$$

和后向变量

$$\beta_t(i)=P(o_1,o_2,\cdots o_T,q_t=i|\lambda)$$

所以

$$P(O|\lambda)=\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b(o_{t+1})\beta_{t+1}(j),1\le t\le T-1 \qquad (2.26)$$

由于训练序列有限，所以得不出估计 λ 的最佳方法。Baum-Welch 算法的思想即是用递归的方法，求出概率 $P(O|\lambda)$ 的极大值，从而得到 λ。

定义

$$\xi_t(i,j)=P(q_t=i,q_{t+1}=j|O,\lambda)$$

该式表示在给定的训练序列 O 和模型 λ 的条件下，HMM 模型在 t 时刻处于状态 i，t+1 时刻处于状态 j 的概率为 $\xi_t(i,j)$。根据全概率公式可推导得

$$\xi_t(i,j)=\frac{\alpha_t(i)a_{ij}b(o_{t+1})\beta_{t+1}(j)}{P(O|\lambda)} \qquad (2.27)$$

那 HMM 模型在 t 时刻处于状态 i 的概率为

$$\gamma_t(i)=P(q_{t=i}|O,\lambda)=\sum_{j=1}^{M}\xi_t(i,j)=\frac{\alpha_t(i)\beta_{t+1}(i)}{P(O|\lambda)} \qquad (2.28)$$

因此 $\sum_{t=1}^{T-1}\gamma_t(i)$表示从状态 i 转移出去的次数的期望，$\sum_{t=1}^{T-1}\xi_t(i,j)$表示从状态 i 转移到状态 j 的次数的期望，可以推导出

$$\bar{\pi}_i = \gamma_1(i) \tag{2.29}$$

$$\bar{a}_{ij} = \sum_{t=1}^{T-1}\xi_t(i,j) \Big/ \sum_{t=1}^{T-1}\gamma_t(i) \tag{2.30}$$

$$\bar{b}_{jk} = \sum_{\substack{t=1 \\ O_k=V_k}}^{T}\gamma_t(j) \Big/ \sum_{t=1}^{T}\gamma_t(j) \tag{2.31}$$

我们将这组公式称为重估公式，由此求得一组新参数和一个新的模型$\bar{\lambda}$，重复这个过程直到$P(O|\bar{\lambda})$收敛，此时就得到最终的模型 $\bar{\lambda}$。

Baum-Welch 算法是收敛的，但只能收敛到局部最优，所以对模型的初始化很重要，但这一问题至今没得到有效解决。同时，该算法是基于最大似然估计（MLE，Maximum Likelihood Estimation）的，即假定语音观察序列确实由该模型产生，如果这个假设条件不成立，就不能达到预期效果。

针对这个问题，研究者提出了基于最大互信息（MMI，Maximum Mutual Information）准则[21]的估计方法。对于训练序列 $\boldsymbol{O}$ 和模型 ，互信息定义为

$$I(\lambda,O) = \log\frac{P(O,\lambda)}{P(O)P(\lambda)} \tag{2.32}$$

通过变换得到

$$I(\lambda,O) = \log P(O|\lambda) - \log\sum_{\lambda'}P(O|\lambda')P(\lambda') \tag{2.34}$$

MMI 准则即为求出一个 λ 使得 $I(\lambda,O)$最大。

虽然从理论上可以证明在模型准确的情况下它不可能比 MLE 更好[22]，但在实际中，由于 HMM 与语音的产生过程有一定的偏差，MMI 比 MLE 的效果更好。因此本文中 HMM 和语言模型的训练均采用 MMI 的估计方法。

2.5.5 Viterbi 解码

连续语音识别中，对给定的声学特征序列 $x^T=(x_1,x_2,\cdots,x_T)$，解码的基本任务是找到最大似然的词序列 $w^{*N}=(w_1^*,w_2^*,\cdots,w_N^*)$，使得

$$w^{*N} = (w_1^*,w_2^*,\cdots,w_N^*) = \arg\max_{w^N}\Pr(x^T|w^N) \tag{2.35}$$

因此，一个好的搜索（解码）算法要求：

（1）准确性。有效地利用各种知识源，使识别结果尽可能的准确。

（2）高效率。尽量快地得出识别结果，理想情况是语音流一经输入立刻得到

识别的文本结果（即实时）。

（3）低消耗。尽量少地占用系统资源，包括需要的内存空间、硬盘空间、CPU 等。

从搜索算法的基本基本思路上，可分为两大类：

（1）时间不同步搜索策略：思路为深度优先搜索，代表算法有 A*算法等。

（2）时间同步搜索策略：思路为广度优先搜索，代表算法有 Viterbi 算法、帧同步算法等。

对 A*算法，其优点是当评估函数选择合适的情况下，算法非常高效，且需要的资源较少；缺点则是评估函数的选择非常困难。因此，在连续语音识别中，一般采用最多的是 Viterbi 解码算法。

Viterbi 算法中，在每一个时刻，都将路径队列中的所有路径在搜索空间内扩展到下一个时刻。在所有可能到达的状态上都保存一条（或多条）似然得分最高的路径，这就形成了下一时刻的路径队列，再继续全部扩展；到达最后一个时刻后，选择所有刚好到达词、词组或句子边界的路径中得分最高的作为输出结果。

将 Viterbi 算法公式化，若一个 left-to-right 的无跳跃 HMM 模型共有 L 个状态，则其初始化为

$$\Phi_1(j)=b_j(y_1),1\le j\le L \tag{2.36}$$

对随后的每个时刻进行递推计算

$$\Phi_t(j)=\max_{1\le i\le L}\left[\Phi_{t-1}(i)\cdot a_{ij}\right]\cdot b_j(o_t),2\le t\le T,1\le j\le L \tag{2.37}$$

则搜索结束时得到的最优分数为

$$Score=\max_{1\le i\le L}\left[\Phi_T(i)\right] \tag{2.38}$$

图 2.5 所示为一个 Viterbi 解码过程的示意图。

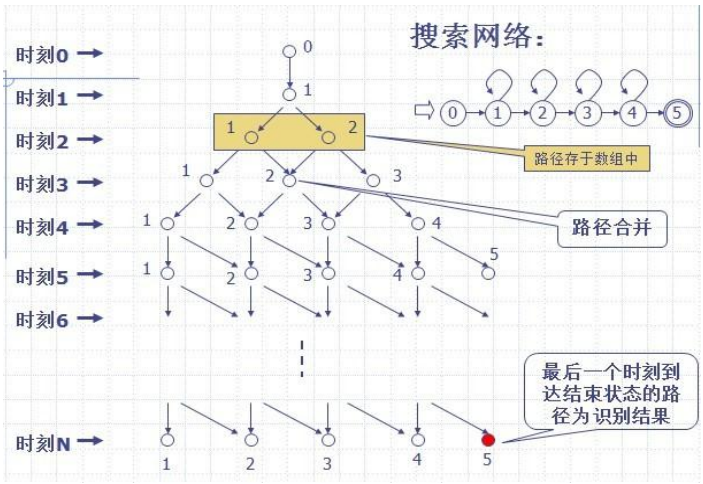


图 2.5 Viterbi 解码示意图

Viterbi 算法的优点是思路简单，容易实现，只需要计算概率似然得分即可，并可搜索到全局最优；缺点是需要进行全搜索，计算复杂度极大，效率低下。为了在保留 Viterbi 算法实现简单的前提下，尽可能地提高搜索效率，研究者改进算法，提出了 Beam（束）搜索概念。

Beam 搜索，顾名思义，就是不对所有路径进行扩展，而是只对一部分（一束）最可能的或者是得分最高的路径进行扩展。对 t 时刻，有

$$Score_{\max}(t)=\max_{path}\left[Score_{path}(t)\right],t=1,2,\cdots,T \tag{2.39}$$

令 b 为 Beam 宽度，在 t 时刻有 $b(t)=f\cdot Score_{\max}(t)$ 其中，f 为 Beam 系数且小于 1。

则对任一条路径 p，得分为 $Score_p(t)$，如果 $Score_p(t)\geq b(t)$，则扩展路径 p（即沿着路径 p 继续搜索），否则删除路径 p。

如果概率得分用对数表示，即 $\log\left[Score_p(t)\right]$，则上述过程修正为，如果

$$\log\left[Score_p(t)\right]-\log\left[Score_{\max}(t)\right]\geq\log[f] \tag{2.40}$$

则扩展路径 p，否则删除路径 p。图 2.6 为 Viterbi-Beam 解码过程的示意图。

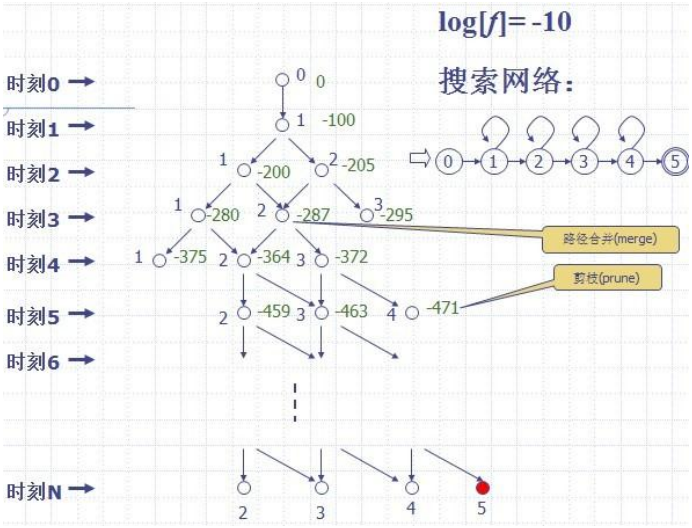


图 2.6 Viterbi-Beam 解码示意图

在 Beam 系数的选择上，从图 2.7 中可以看出，随着 f 增大，识别效率会升高，但识别性能会下降；所以 f 是考虑性能和效率的折中点，通常要通过多次实验给定。

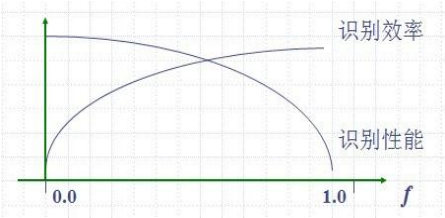


图 2.7 Beam 系数选择

大量实验证明，当 f 选择比较好，可以非常明显地提高搜索效率；因为不是对所有路径都扩展，所以得到的结果不能保证是全局最优，但对搜索性能的影响可以忽略不记。因此，本文实验中解码图采用的也是 Viterbi-Beam 解码的方法。

2.5.6 HMM 算法的实现问题

在 HMM 实现的过程中，还会遇到几个问题：

1.初始模型的选取。在用 Baum-Welch 算法训练 HMM 参数时，选取一个好的初始模型会使最后求得的局部极大值接近全局最大值，因此，对不同形式的 HMM 通常会采用不同的初值选取方法。典型的 HMM 参数估计过程如图 2.8 所示。

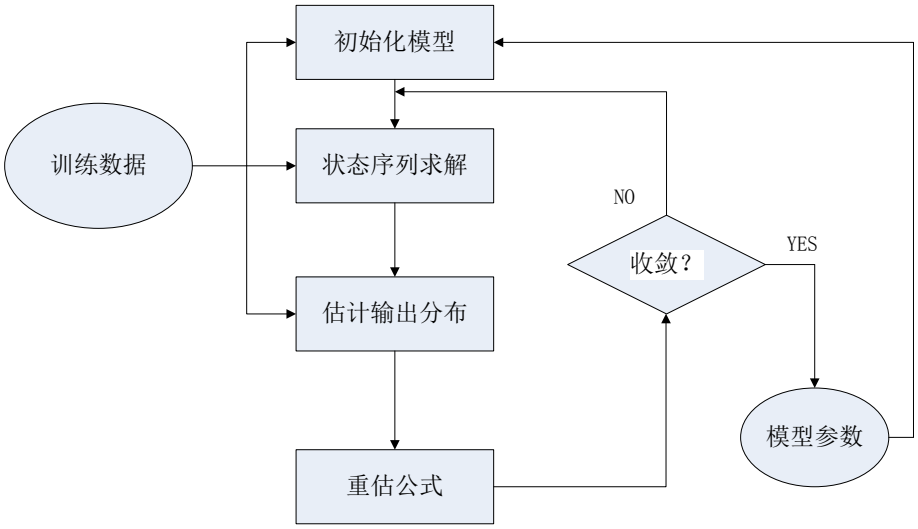


图 2.8 典型 HMM 参数估计过程图

2．多个观察序列的训练。在实际对 HMM 训练的过程中，常常用到多个观察序列，此时就要对 Baum-Welch 算法的重估公式进行修正。

3.数据下溢问题。在 Baum-Welch 算法中有对前向变量和后向变量的递推计算，其值小于 1，因此在递推过程中，这两个变量都将迅速趋近于零，这就是数据下溢问题。为了解决这个问题，通常加入比例系数对算法进行修正。

4.模型的选取。这在本章节的第 2 小节中已经有过详细论述。

2.6 语言模型

当若干词组成的一个序列合乎语法时，这个序列才能算是一个句子，因此，人们在语音识别中引入了语言模型来实现这种约束。当前的语言模型主要有基于句法的语言模型和基于统计的语言模型两大类。

句法语言模型（Syntactic Language Model），也称确定性语言模型（Deterministic Language Model）或形式语言模型（Formal Language Model），是人工对人类语言的内在规律总结出一套形式上可以推理和扩展的文法，对识别结果中不符合文法的结果进行排除。这种方法在某些识别任务中能够获得很好的效果。

基于统计的语言模型对大量文本中的词的出现频率及其出现条件进行统计。通常我们将统计语言模型与声学模型结合起来完成识别任务，这可以降低因为声学模型的不合理带来的拒识率。

目前在大词汇量连续语音识别中常用的是 N-Gram 语言模型[23]，对中文而言，我们称之为汉语语言模型（CLM，Chinese Language Model）[24]。对一个句序列 $S = w_1, w_2, \cdots w_Q$，其出现的概率 $P(S)$ 为

$$P(S) = P(w_1, w_2, \cdots w_Q) \tag{2.41}$$

将其转换为条件概率形式为

$$P(S) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1 w_2)\cdots P(w_Q \mid w_1 w_2 \cdots w_{Q-1}) \tag{2.42}$$

在这种情况下，精确计算出 $P(S)$ 几乎是不可能的事情，因此假设其中第 N 个词的出现只于前面 $N-1$ 个词相关，与其它词都不相关，整句的概率就是各个词出现概率的乘积，而这些概率都可以通过统计的方法从文本中获得。通常系统中采用的为 Bi-Gram（N=2）和 Tri-Gram（N=3）。

一个语言模型的质量的评价指标通常用语言模型复杂度（Perplexity）来表述，其定义为词序列概率的几何平均的倒数

$$PP = \left[p\left(w_1^Q\right)^{-\frac{1}{Q}} \right] = \left[\prod_{q=1}^{Q} p\left(w_q \mid w_{\max[q-N+1,1]}^{q-1}\right) \right]^{-\frac{1}{Q}} \tag{2.43}$$

当复杂度越低，说明语言模型对当前词的预测确定程度越高。因此对语言模型的训练一般把训练语句的复杂度最小化作为目标，而实现这个目标，先要对训练语句中的词频进行统计，以此计算出语言模型的参数。而在词表很大而训练的数据不是充分多的时候，就会发生有些词的序列的概率很小或者没有出现过的情况。为了解决这些问题，就需要用到 discouting[25]和 back-off[26]等一些技术。discouting 的方法是将一些训练集中的词序列概率压缩，将多出的分配给未出现在词表中的词序组合；back-off 的方法则是将未出现的、长词序列用段词序列来替代，这点在我们训练语言模型的时候将会具体说明。

2.7 噪声鲁棒性技术

当前的语音识别系统在实验室环境下已经能够获得足够高的识别率，但导致语

音识别技术的广泛应用的一个重大障碍就是实际环境中各种噪声的干扰，这些干扰包括各种环境的背景噪声、语音信号的采集和传输过程中的信道噪声，还有因为说话人的情绪变化所引起的发音变异问题。

2.7.1 噪声与信噪比

通常在语音信号处理中，我们将噪声分为加性噪声（Additive Noise）和信道噪声（Channel Noise）两大类。

加性噪声按随时间的变化可以分为稳态噪声（Stationary Noise）和非稳态噪声（Non-stationary Noise），而非稳态噪声又有瞬态的、周期性起伏的、脉冲的和无规则的噪声之分。

在加性噪声中，噪声与信号的关系是相加的，即不管有没有信号，噪声都存在。若用 $x(i)$ 表示含噪语音信号，$s(i)$ 表示纯净语音信号，$n(i)$ 表示噪声信号，则加性噪声中这三者的关系为

$$x(i)=s(i)+n(i)$$

信道噪声，又称为卷积噪声或乘性噪声，其随着信号的存在而存在，当信号消失后，信道噪声也随着消失，其叠加关系为

$$x(i)=s(i)*n(i)$$

通过同态变换，信道噪声可以变换为加性噪声。

为了度量语音信号受到噪音污染的程度，我们引入了信噪比（SNR，Signal to Noise Ratio）的概念，其定义如下

$$SNR=10\log_{10}\left(\frac{P_{signal}}{P_{noise}}\right) \tag{2.44}$$

或

$$SNR=20\log_{10}\left(\frac{A_{signal}}{A_{noise}}\right) \tag{2.45}$$

信噪比的单位为分贝（dB），式中 P 为功率，A 为幅值。

2.7.2 信号空间噪声鲁棒技术

信号空间的去噪技术一般又称为语音增强，这部分处理是在时域空间上，发生在特征提取之前。语音增强技术主要从人类的听觉感知出发，考虑人对增强后的语音信号的听觉感受的变化，取决于人的主观感受。

在语音增强中，谱减法（SS，Spectral Subtraction）[27]是最早采用也是最经典的

方法之一，其基本思想是直接从带噪语音信号的频谱中减去噪声的平均频谱，实现上分为噪声更新和噪声消除两个步骤。在噪声更新时，首先要进行端点检测（VAD，Voice Activity Detection），然后根据帧信息来估计噪声频谱；在噪声消除时，用带噪语音信号的频谱减去估计出来的噪声频谱，就得到了干净语音信号频谱的估计。使用谱减法能够较好的抑制原始语音信号中的噪声，但处理过程中的非线性操作会残留音乐噪声[28]。

语音增强的另一种主要方法是维纳滤波（WF，Wiener Filtering）。在这种方法中，需要设计一个最佳线性滤波器，使该滤波器输入的估计纯净语音信号与我们期望的纯净语音信号的均方误差最小。

2.7.3 特征空间噪声鲁棒技术

特征空间的噪声鲁棒技术应用在特征提取之后，其目的主要是降低识别系统的识别错误率，因而很少考虑到人耳听觉系统对某些语音信号畸变的自动纠正。

在特征空间的去噪技术中，最常采用的是倒谱均值归一（CMN，Cepstral Mean Normalization）。在 CMN 中，假设信道的线性传输函数是时不变的，而信道产生的噪声一般是卷积噪声，将时域的卷积运算变换为倒谱域的求和运算，此时通过减均值的处理就可以将信道产生的噪声去除。

在本文中，采用的是 CMN 的一种扩展方法，即倒谱均值方差归一（CMVN，Cepstral Mean and Variance Normalization）[29]，通过同时对均值和方差进行归一化处理，不仅能够去除信道的影响，对加性噪声的抑制也有很好的效果。

对一组特征提取后的特征参数序列 $X=\{x_1,x_2\cdots x_T\}$，其均值和标准差为

$$\mu=\frac{1}{T}\sum_{t=1}^{T}x_t,\sigma=\sqrt{\frac{1}{N}\sum_{t=1}^{T}(x_t-\mu)^2}$$

则 CMVN 可以定义为

$$\hat{x}_t=\frac{x_t-\mu}{\sigma} \tag{2.46}$$

2.7.4 模型空间噪声鲁棒技术

模型空间的噪声鲁棒技术通常是通过修改声学模型来实现对模型的补偿。通常在语音识别中改变 HMM 模型的拓扑结构，会导致模型的参数不能够精确的表示，降低模型的准确性，因此大多数模型补偿的方法并不改变 HMM 的拓扑结构

最大似然线性回归（MLLR，Maximum Likelihood Linear Regression）[30]的方法是通过对噪声数据的统计分析得出经验，在此基础上对纯净语音输出的均值和方差

进行线性变换来对带噪数据进行拟合，这种方法简单而有效，因而本文中也采用了这种方法。

最大后验（MAP，Maximum a Posteriori）[31]估计准则是一种基于贝叶斯（Bayesian）理论的自适应方法。假定待估参数为一随机变量且服从某种先验分布，MAP 的实质是将先验知识与从自适应数据中得到的知识结合起来，当自适应数据无限多时，MAP 估计等价于 MLE 估计，当缺少自适应数据时，相当于没有进行自适应处理，即参数值等于初值，这一点也是 MAP 最大的缺陷。

2.7.4.1 区分性训练

在贝叶斯决策理论中，首先必须对条件概率的概率分布建模，但是在实际环境中，问题会变得异常复杂，而我们又不得不选择一些较为简单的数学函数，因为这样才能够实现处理，这样的矛盾使得我们在实际中无法用简单的数学模型来表述十分复杂的数据分布，因而也就使得假设的声学模型是一个错误的模型。其次，即使我们在这种假设的错误模型上进行处理，我们也无法得到充分的数据来对模型参数进行估计，特别在语音识别中，现实中存在着大量的语音变体（如说话人差异、口音差异、环境差异等），导致我们的训练数据总是稀疏的。

基于这种种原因，理论最优的方法在实际中几乎不可能实现，因此 MLE 估计也就不能得到最优分类。正因如此，使用区分性训练的方法就能够在诸多限制条件下提高声学模型的区分能力。

常用的区分性训练方法有基于最大互信息（MMI，Maximum Mutual Information）准则的训练方法和基于最小分类错误（MCE，Minimum Classification Error）准则的方法。本文中采用了前者来对 HMM 进行模型补偿。

2.7.4.2 最大互信息准则

Lalit R.Bahl 在 1986 年提出了用 MMI 准则估计 HMM 参数的方法[32]。令 W 为语音信息中的随机变量，O 为观察序列的随机变量。根据信息论的理论，可以描述为信息 W 被编码为 O。在已知 O 的情况下求出信息 W 即是语音识别中的解码过程，在该条件下，描述 W 平均不确定性度量的条件熵为

$$H(W|O)=-\sum_{W,O}P(W,O)\log P(W|O)=-E\cdot\log P(W|O) \qquad (2.47)$$

我们的目的就是要降低不确定度，使解码器的判决更准确。在实际声学建模的过程中，由于 $P(W|O)$ 不可知，我们用一个参数化声学模型 来得到该后验概率的一个估计 $P_{\Lambda}(W|O)$，则有

$$\begin{aligned} H_{\Lambda}(W|O) &= -E\cdot\log P_{\Lambda}(W|O) \\ &= -\sum_{W,O}P(W,O)\log P_{\Lambda}(W|O) \\ &= -\sum_{W,O}P(W,O)\log\frac{P_{\Lambda}(W|O)}{P(W|O)}-\sum_{W,O}P(W,O)\log P(W|O) \\ &\geq -\sum_{W,O}P(W,O)\left[\frac{P_{\Lambda}(W|O)}{P(W|O)}-1\right]+H(W|O) \\ &= H(W|O) \end{aligned} \tag{2.48}$$

根据上面的推导过程可以得到，$H_{\Lambda}(W|O)$是$H(W|O)$的一个最大值，因此对$H(W|O)$的精确估计就是求得$H_{\Lambda}(W|O)$的最小值。

根据互信息的定义有

$$I(W;O)=H(W)-H(W|O) \tag{2.49}$$

令$H(W)$为一已经的固定值，对声学模型Λ有

$$I_{\Lambda}(W;O)=H(W)-H_{\Lambda}(W|O) \tag{2.50}$$

因此将$H_{\Lambda}(W|O)$的最小值转变为求$I_{\Lambda}(W;O)$最大值的问题，即求最大互信息量，在这一过程中，MMI 准则等价于条件最大似然准则（CML，Conditional Maximum Likelihood）[33]。

在实际训练的过程中，由于训练数据量的限制，我们通常采取将式（2.47）的期望改为对所有训练语料求和的方法，即式（2.41）可改写为

$$\hat{H}_{\Lambda}(W|O)=-\frac{1}{N}\sum_{n=1}^{N}\log P_{\Lambda}(W_n|O_n) \tag{2.51}$$

其中$\hat{H}_{\Lambda}(W|O)$为$H_{\Lambda}(W|O)$的估计。

在基于 HMM 的模型中，上式的求和就是参考序列的后验概率，通过计算正确序列与所有可能序列的似然比可求得

$$P_{\Lambda}(W_n|O_n)=\frac{P_{\Lambda}(O_n|W_n)P(W_n)}{\sum_{W'\in M}P_{\Lambda}(O_n|W')P(W')} \tag{2.52}$$

因此，MMI 准则又可看作是所有训练语料正确模型序列的最大化后验概率，即

$$\mathrm{F}_{\mathrm{MMI}}=\frac{1}{N}\sum_{n=1}^{N}\log\frac{P_{\Lambda}(O_n|W_n)P(W_n)}{\sum_{W'\in M}P_{\Lambda}(O_n|W')P(W')} \tag{2.53}$$

2.8 小结

本章主要阐述了一个完整的语音识别系统的主要组成部分和主要的噪声鲁棒性

技术。对构建一个完整的语音识别系统，介绍了在语音信号的分析处理、声学特征的选择和声学模型的训练、语言模型的训练上的主要方法和常见问题；而对语音识别实际应用中需要面对的噪声鲁棒性问题，分别从信号层、特征层、模型层介绍了一些主要的鲁棒性方法；针对本文中大词汇量连续语音识别任务的要求，对具体采用的技术和方法进行了详细分析。

第 3 章　基于时域 Gammatone 滤波的 GFCC 特征

与 MFCC 类似，本文中研究的基于 Gammatone 的倒谱系数（GFCC，Gammatone Frequency Cepstrum Coefficient）也是一种模拟人类听觉系统响应特征的语音特征提取方法。人类的听觉系统是一个高度复杂敏感的系统，对不同频率的信号分量有不同形式的响应，这种响应是非线性的，这种非线性可以通过一组 Gammatone 滤波器实现[34]。

本章节主要介绍了 Gammatone 滤波的原理和 GFCC 的提取方法。

3.1 等效矩形带宽

针对人耳基底膜的滤波特性，假设有一组滤波器，每个滤波器的中心频率 f_c 和等效矩形带宽（ERB，Equivalent Rectangular Bandwidth）都不相同，它们之间的关系为

$$ERB(f_c) = f_c/Q + B_0 \tag{3.1}$$

其中，Q 为渐进因子，B_0 为最小带宽。

在研究 ERB 的过程中，研究者对 ERB 与中心频率 f_c 之间的关系做出了不断的修正，1961 年 Zwicker[35]提出

$$ERB_1 = 25 + 75\left(1 + 1.4 f_c^2\right)^{0.69} \tag{3.2}$$

1983 年 Moore 和 Glasberg[36]修正 ERB 为

$$ERB_2 = 6.23 f_c^2 + 93.39 f_c + 28.52 \tag{3.3}$$

1990 年 Glasberg 和 Moore[37]又对 ERB 做出修正

$$ERB_3 = 24.7\left(1 + 4.37 f_c\right) \tag{3.4}$$

在本文中，取 $Q = 9.26449$，$B_0 = 24.7\,\text{Hz}$，则根据式（3.1）可以得出式（3.4）。

3.2 时域 Gammatone 滤波

3.2.1 Gammatone 滤波器组

Gammatone 滤波器（简称 GF）最早由 Aertsen 和 Johannesma 提出[38]，其为一组非均匀重叠的带通滤波器组，它的时域阶跃响应可以表示为

$$g(t) = at^{n-1}e^{-2\pi bt}\cos\left(2\pi f_c t + \phi\right)u(t) \tag{3.5}$$

其中，f_c 为滤波器的中心频率，ϕ 为相位，通常取 $\phi = 0$；a 为增益常数；n 为滤波器的阶数，通常设置为小于等于 4；b 为带宽相关的参数，它与 ERB 的关系可

以表示为

$$b = 1.019ERB(f_c) \tag{3.6}$$

根据上式和式（3.4），b 与 f_c 的关系可以表示为

$$b = 1.019 \times 24.7 \times (4.37 \times f_c / 1000 + 1) \tag{3.7}$$

图 3.1 所示为覆盖带宽从 100Hz 到 8000Hz，16 通道的滤波器的频率响应曲线。

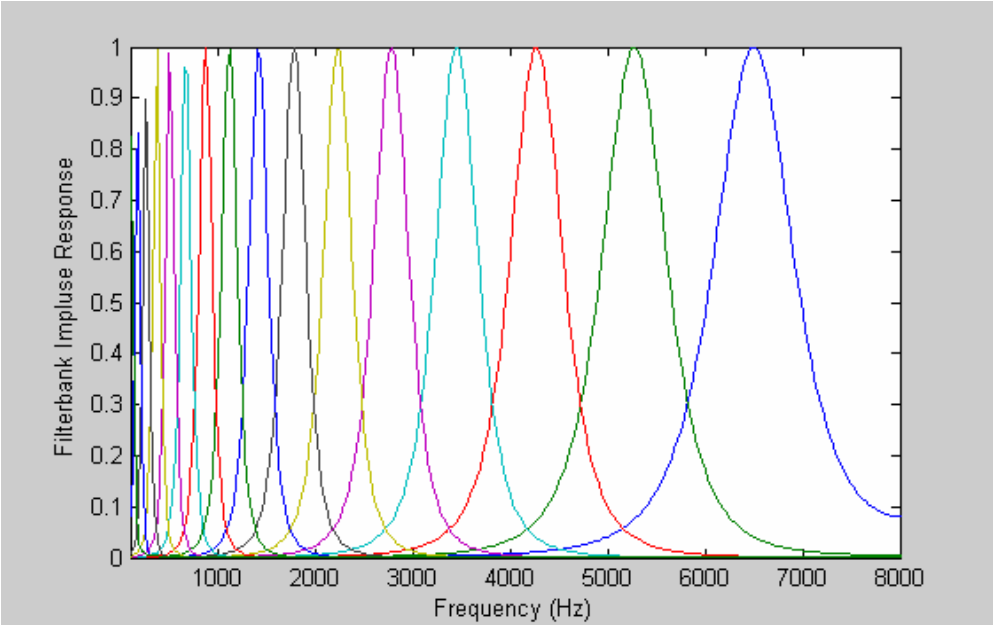


图 3.1 Gammatone 滤波器频率响应

3.2.2 带宽和中心频率

因为是基于人耳的听觉系统设计，每一个 Gammatone 滤波器的滤波带宽都由听觉临界频带决定，即 ERB，其与中心频率的关系为式（3.4）。对临界频带划分数量为 M，其与频率 f 之间的微分关系为

$$dM = \frac{\Delta M}{\Delta f} df = \frac{1}{\Delta f / \Delta M} df = \frac{1}{ERB(f)} df \tag{3.8}$$

则这两者的积分表达为

$$M = \int_0^{f_c} \frac{1}{ERB(f)} df \tag{3.9}$$

对一组数量为 N 的 Gammatone 滤波器组，其覆盖频率范围为 f_L ~ f_H，相邻两个滤波器之间的重叠百分比为ν，根据式（3.9）有

$$M = \int_{f_L}^{f_H} \frac{1}{ERB(f)} df \tag{3.10}$$

联合式（3.1）和式（3.10）可得

$$M=\int_{f_L}^{f_H}\frac{Q}{f+QB_0}df=QIn\frac{f_H+QB_0}{f_L+QB_0} \tag{3.11}$$

因为$M=N\cdot\nu$，所以有

$$N=\frac{Q}{\nu}In\frac{f_H+QB_0}{f_L+QB_0} \tag{3.12}$$

对第 n 个滤波器，$1\le n\le N$，其中心频率 $f_c(n)$ 为

$$f_c(n)=-QB_0+(f_H+QB_0)\exp^{-\frac{n}{N}In\frac{f_H+QB_0}{f_L+QB_0}} \tag{3.13}$$

带宽 $ERB(n)$ 为

$$ERB(n)=24.7[1+4.37f_c(n)]$$

3.2.3 时域分析

文献[39]中提出了对 Gammatone 滤波的时域分析方法。多个不同中心频率的 Gammatone 滤波器构成了一个 Gammatone 滤波器组，与基于快速傅里叶变换（FFT，Fast Fourier Transform）的短时谱分析类似，经过 Gammatone 滤波器组的信号代表了原始信号在不同频率分量上的响应特征。观察式（3.6）的结构，它由两个部分组成：波形包络 $at^{n-1}e^{-2\pi bt}$ 和频率 f_c 的调幅 $\cos(2\pi f_c t+\phi)$。通过傅里叶分析，$g(t)$ 的频率域表达如下

$$G(f)=\frac{a(n-1)!}{2(2\pi b)^n}\left\{\left[\frac{j(f-f_c)}{b}+1\right]^{-n}+\left[\frac{j(f+f_c)}{b}+1\right]^{-n}\right\} \tag{3.14}$$

在 f_c/b 足够大的情况下，$[j(f+f_c)/b+1]^{-n}$ 趋近于零，则式（3.9）可以改写成下面的形式

$$G(f)\approx\frac{a(n-1)!}{2(2\pi b)^n}[1+j(f-f_c)/b]^{-n} \tag{3.15}$$

令 $s=j2\pi f$，GF 的拉普拉斯变换表示为

$$G(s)=\frac{a(n-1)!}{2}[s-(j2\pi f_c-2\pi b)]^{-n} \tag{3.16}$$

其 Z 变换表示为

$$G(z)=\frac{a(n-1)!}{2}(1-e^{j2\pi f_c-2\pi b})^{-n} \tag{3.17}$$

令 $A(z)$ 为一个元素，其表达式为

$$A(z)=\frac{1}{1-e^{j2\pi f_c/f_s-2\pi b/f_s}z^{-1}} \tag{3.18}$$

$G(z)$ 可以看作是 n 个 $A(z)$ 递归应用的串联。$A(z)$ 与中心频率 f_c 有关，因此 $G(z)$ 也

与 f_c 有关。考虑当 $n=4$ 时，有如下表达

$$\hat{G}(z)=\frac{3a}{1-4mz^{-1}+6m^2z^{-2}-4m^3z^{-3}+m^4z^{-4}} \tag{3.19}$$

其中 $m=e^{-2\pi b/f_s}$ 。

3.3 GFCC 特征提取

在特征提取前，首先对每一通道的 Gammatone 滤波信号进行预加重，本文中采用的预加重函数为

$$H(z)=1+4e^{-2\pi b/f_s}z^{-1}+e^{-2\pi b/f_s}z^{-2} \tag{3.20}$$

其中，b 由式（3.7）定义，f_s 为采样频率。经过预加重的信号被分成有重叠的语音帧序列。在本文的处理方法中，采样信号率为 16kHz，帧长为 400 个点，相邻帧重叠 160 个点，即每一帧的长度为 25ms，帧移为 10ms。对每一帧信号求均值，得到该通道的平均帧能量。

对每一帧时刻，Gammatone 滤波器在各个通道上的平均帧能量组成了该帧的向量表达，并通过离散余弦变换（DCT）去除向量间的相关性，得到 GFCC 特征。为使数值处理上更加稳定，帧向量在进行 DCT 之前先对其进行对数压缩，其数学表达为

$$F(n,v)=\left(\frac{2}{M}\right)^{0.5}\sum_{i=1}^{M}\left\{\frac{1}{3}\log\left[\bar{y}(n,i)\cos\left[\frac{\pi v}{2N}(2i-1)\right]\right]\right\} \tag{3.21}$$

其中 M 为通道数，本文中 $M=32$，n 为通道编号，范围从 0 到 31，v 为特征维数。与 MFCC 的特征维数选取类似，我们需要选择出最能表征语音信息特征的分量。当 $v>13$ 时，$F(\cdot,v)$ 的大多数数值会接近 0，因此我们选取前 12 个元素，并将 C0 分量作为第 1 维特征，即生成一组 13 维的静态 GFCC 特征向量。

为了取得更好的识别效果，通常会对静态特征进行一些处理，生成一些新的特征向量。我们在此对静态特征进行差分计算，将静态特征与差分运算后生成的向量拼接成一组新的特征，这些拼接的特征又被称作 GFCC 的动态特征。在本文中，我们选取一阶差分和二阶差分作为动态特征，其数学表达如下：

$$\Delta F(n,v)=\frac{\sum_{k=1}^{K}k\left[F(n+k,v)-F(n-k,v)\right]}{2\sum_{k-1}^{K}k^2} \tag{3.22}$$

$$\Delta\Delta F(n,v)=\frac{\sum_{k=1}^{K}k\left[\Delta F(n+k,v)-\Delta F(n-k,v)\right]}{2\sum_{k-1}^{K}k^2} \tag{3.23}$$

其中 K=2，即差分窗口长度为 5。

3.4 本章小结

本章中详细介绍了 Gammatone 滤波的原理及其在时域中的实现方法，对语音信号在时域上进行 DCT 变换并提取出 GFCC 的方法，相对于传统的频域 FFT 变换的方法，其计算量更小，因此识别系统的识别速度也就更快。而 GFCC 对噪声的鲁棒性，我们将通过第 5 章中的实验结果进行详细分析。

第 4 章 基于 WFST 的语音识别解码方法

当前很多主流的大词汇量语音识别系统都能够通过加权有限状态转换器（WFSTs，Weighted Finite-State Transducers）来构建。在加权有限状态机的理论下，语音识别中的各种模型，如 HMM，发音词典，多元语法语言模型都能够转换成加权有限状态转换器的形式，并根据加权有限状态转换器理论中的组合、最小化操作，将这些转换器组成一个完整的静态搜索网络。

与传统的语音识别解码网络相比，经过组合和最小化后的 WFST 解码网络，能够大大降低网络的规模，并且降低了解码过程中的时间和空间复杂度，保存了全局最优路径。同时，单阶段识别系统（1-pass）与传统的两阶段识别系统（2-pass）相比，识别速度更快；而在 WFST 的组合操作中，我们可以选择加入更多的知识源，这样能够提高某些特定识别任务或特别环境的识别性能。因此，采用 WFST 构建的语音识别解码系统，是一种快速的、高效的语音识别系统。

本章主要介绍了加权有限状态转换器的理论和构建 WFST 解码图的具体实现。

4.1 加权有限状态机定义

在加权有限状态机理论中，加权有限状态接收器（WFSAs）和加权有限状态转换器（WFSTs）都以半环代数结构来表示。文献[40]中，作者对其作出了定义。

一个半环代数结构 K 包含一个数值集合 **K**，两个基本操作 $\oplus$ 和 $\otimes$，两个基本单位 $\bar{0}$ 和 $\bar{1}$，可以写为 $(K,\oplus,\otimes,\bar{0},\bar{1})$。表 4.1 给出了几种半环代数结构的表达。

表 4.1 几种半环代数结构表达

半环代数结构	集合	$\oplus$	$\otimes$	$\bar{0}$	$\bar{1}$
Boolean	$\{0,1\}$	$\vee$	$\wedge$	0	1
Probability	R_+	+	×	0	1
Log	$R\cup\{-\infty,+\infty\}$	$\otimes_{\log}$	+	$+\infty$	0
Tropical	$R\cup\{-\infty,+\infty\}$	min	+	$+\infty$	0
String	$\Sigma^*\cup\{\infty\}$	$\wedge$	•	∞	ε

4.1.1 加权有限状态接收器

在语音识别中，各种模型（如 HMMs）就是 WFSAs 的一个特例。对一个定义在半环代数结构 **K** 上的 WFSA，其表达为 $A=(\Sigma,Q,E,i,F,\lambda,\rho)$，其中，$\Sigma$ 为符号字母集，Q 为有限状态集，E 为有限传递集且有 $E\subseteq Q\times(\Sigma\cup(\varepsilon)\times K\times Q)$，$i$ 为初始状态且 $i\in Q$，F 为结束状态集且 $F\in Q$，λ 为初始权值，ρ 为结束权值函数。

对一个传递 $t=(p[t],l[t],w[t],n[t])\in E$，通过一段弧连接初状态或前一状态 $p[t]$ 和末状态或下一状态 $n[t]$，其中 $l[t]$ 为符号，$w[t]$ 为权值且通常表示为概率或对数概率。

对 A 的一条连续的传递路径 $t_1,t_2,\cdots,t_n$，有 $n[t_i]=p[t_{i+1}]$，当输入为空时，用空符号 ε 来表示转移的符号。对于一条成功的路径 $\pi=t_1,t_2,\cdots,t_n$，其初始状态为 i，结束状态为 f 且 $f\in F$。则路径 π 的符号就是其传递过程中所有状态的符号的联合，即

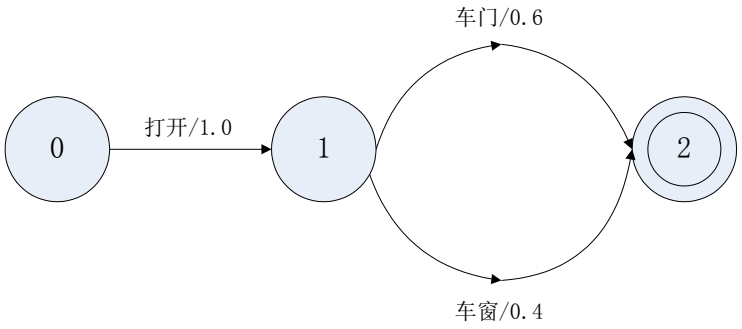
$$l(\pi)=l[t_1]\cdots l[t_n] \tag{4.1}$$

其初始权值和终止权值函数与 π 的关系为

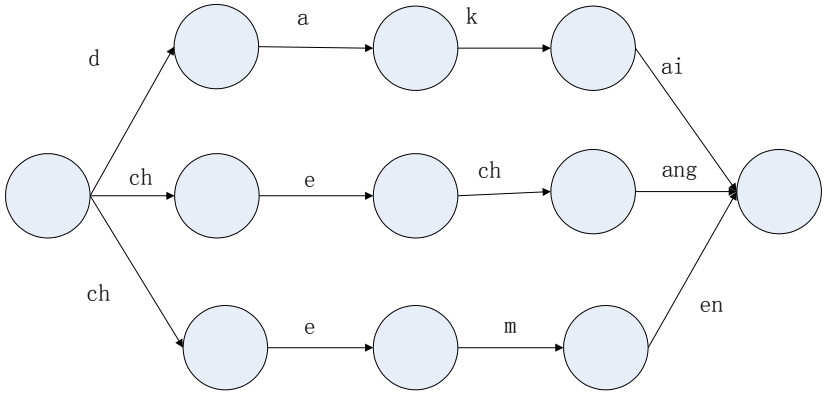
$$w[\pi]=\lambda\otimes w[t_1]\otimes\cdots\otimes w[t_n]\otimes\rho(n[t_n]) \tag{4.2}$$

如果存在成功的路径 π，则有其符号序列 $x=l[\pi]$。对所有的成功路径 π，计算其整个权值 $w[\pi]$ 来最终确定符号序列 x。

我们通过图 4.1 给出了 WFSAs 的一些简单的例子。



（a） 语言模型



（b）发音词典

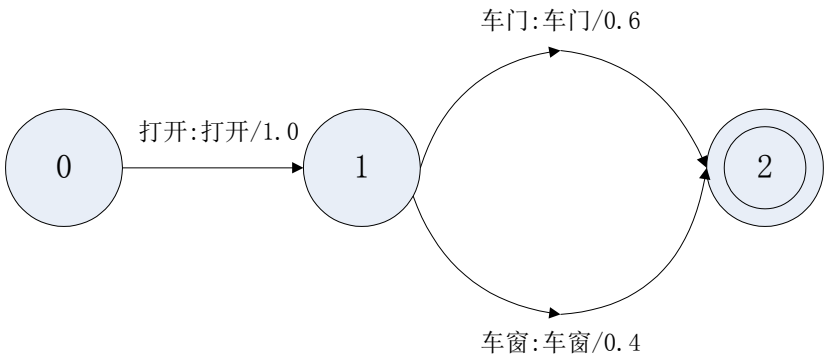
图 4.1 简单的 WFSA 结构

在图 4.1（a）中，0 为初态，2 为终态，每一条路径都代表了一个合法的句子（如打开车门，打开车窗），在每一个传递中每个词出现的权值的乘积都是整个句子出现的概率（如打开车门的概率为 0.6，打开车窗的概率为 0.4）。图 4.2（b）则是发音词典的 WFSA 表示，我们也可以对每一个音素加上一个权值来确定他们出现的概率。

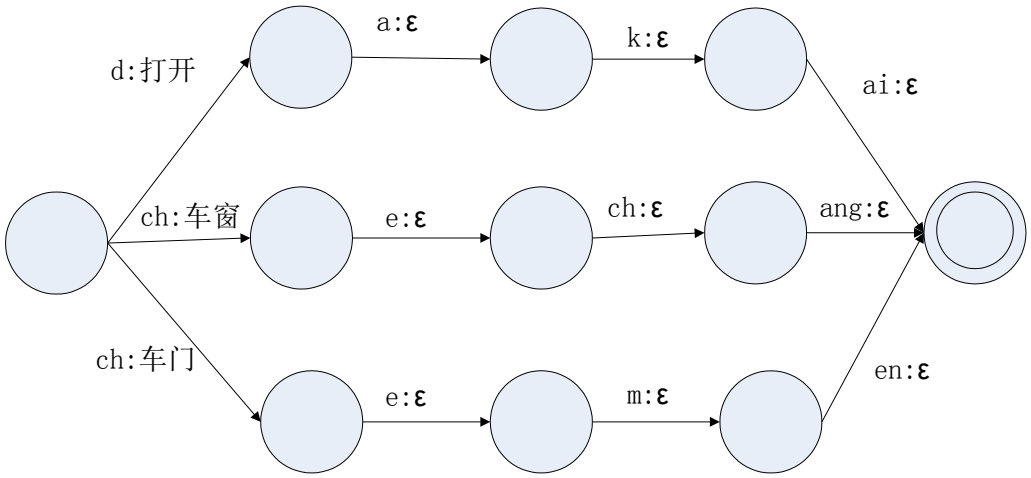
4.1.2 加权有限状态转换器

WFSTs 与 WFSAs 相比，用一对变量(i,o)来替换单个的转移符号，其中 i 为输入符号，o 为输出符号。即我们定义一个在半环结构 **K** 上的 WFST 为 $T=(\Sigma,\Omega,Q,E,i,F,\lambda,\rho)$，与 WFSA 有所不同，其中 为输入的字母表，$\Omega$ 为输出的字母表，有限状态传递集 $E\subseteq Q\times(\Sigma\cup\{\varepsilon\})\times(\Omega\cup\{\varepsilon\})\times K\times Q$。

对 WFST 的一个传递 $t=(p[t],l_i[t],l_o[t],w[t],n[t])$，与 WFSA 类似，可以用一段弧来表示其传递过程，但与 WFSA 不同的是，用一组词表的输入 $l_i[t]$ 和输出 $l_o[t]$ 这一过程替代了原本的 $l[t]$。



（a） 语言模型



（b）发音词典

图 4.2 简单的 WFST 结构

图 4.2 为 WFST 的一个例子，图 4.2（a）为单词序列在语言规则限制下映射到单词序列的语言模型，数字为映射过程中的权重关系，一条路径上的权重的乘积即为该句子出现的概率，如“打开车门”出现的概率为$0.6(=1.0\times0.6)$，“打开车窗”出现的概率为$0.4(=1.0\times0.4)$。图 4.2（b）则是单词“打开”、“车门”和“车窗”的音素映射序列，除了初始单词的传递外，其他音素的输出词表都是ε，当采用输出词表的方式来对单词进行编码，就能够将相同的单词的发音转换器联合起来而不会丢失单词的特定标记（ID）。

4.2 加权转换器处理

在基于有限状态理论的语音识别任务中，通常要求实时解码器能够组合并优化这些转换器。解码器首先在词典中搜寻单词的发音并在语法中替换它们[41]，然后用文本中的每一个音素去识别错误的文本相关的模型，最后生成一个 HMM-级的转换器去替换之前的转换器。在这一过程中，对转换器之间有组合、确定化、最小化等操作，而这也是其不同于传统语音识别解码器的特点。

4.2.1 组合（Composition）

对两个转换器T_1和T_2，它们的组合可以表示为$T=T_1\circ T_2$，且存在路径使序列u映射到序列w，其中第一条路径在T_1中从序列u映射到序列v，第二条路径T_2中从序列v映射到序列w，其组合后的路径中的权值可以由T_1和T_2中相关路径的权值通过$\otimes$计算得到。

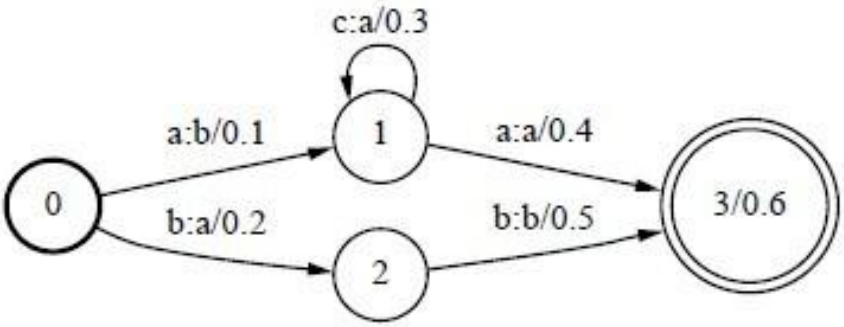
对一个组合$T=T_1\circ T_2$，其必须满足以下三个条件：

（1）T 的一对初始状态必须是T_1和T_2的初始状态；

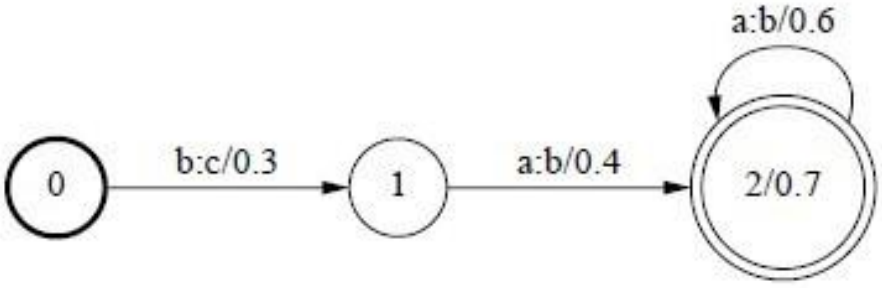
（2）T 的一对结束状态必须是T_1和T_2的结束状态；

（3）存在传递 t 从(a,b)到(a',b')，且对传递t_1有从状态 a 到状态a'，对传递t_2有从状态 b 到状态b'。

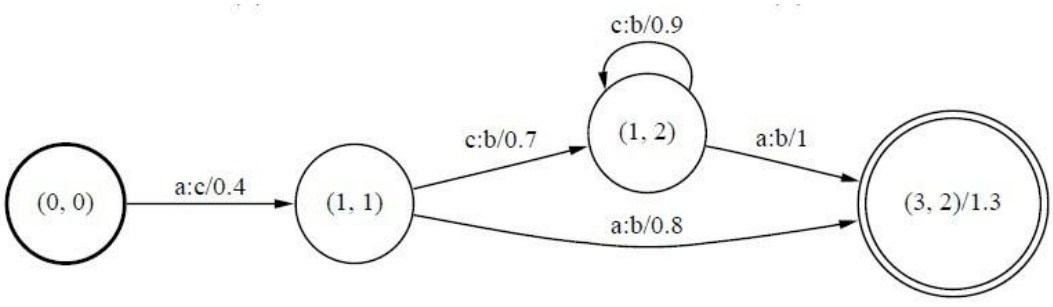
在实际中，ε符号会出现在T_1的输出序列和T_2的输入序列中，这将会对组合操作产生影响，因此需要在这一过程中引入ε过滤器[42]。



（a）



（b）



（c）

图 4.3 WFST 组合

图 4.3 举出了一个在 Tropical 半环上两个简单的 WFST 组合的例子，图 4.3（a）和 4.3（b）为两个简单的转换器，分别是T_1和T_2，图 4.3（c）是它们组合后的结构。在T_1中，路径$0\to1\to3$将输入符号序列“a c a”映射到序列“b a a”，在T_2中，路

径$0\rightarrow1\rightarrow2$将输入符号序列“a c a”映射到“c b b”。

T 中路径的权值为T_1和T_2对应路径上的权值进行$\otimes$操作后的结果，通常为各自对应的对数概率之和。图 4.4 即为图 4.2 中两个 WFST 的组合。

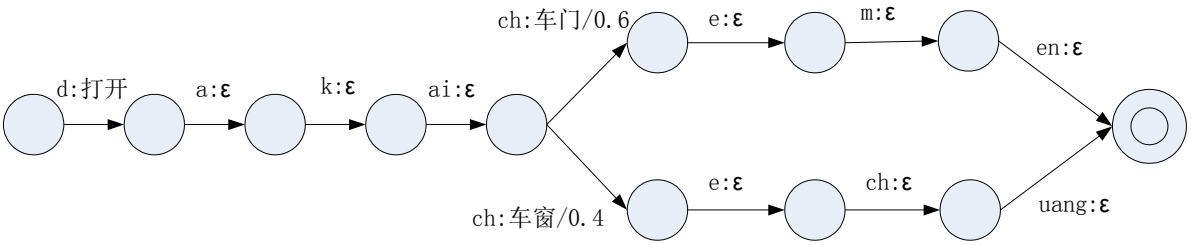


图 4.4 WFST 组合实例

4.2.2 确定化（Determinization）

当一个 WFST 的每一个状态的输入只有一个传递且输入不为ε时，我们称这个转换器是确定的或有序的。图 4.5 给出了一个非确定性的 WFSA，在状态 0 时，对词表 a 有多条传递，因此是不确定的。

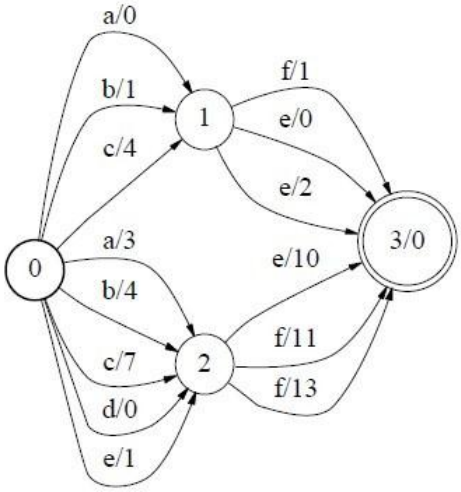


图 4.5 不确定性 WFSA

对加权有限状态机进行确定化操作，能够减少冗余度。对一个确定化以后的加权有限状态机来说，每一个给定的输入序列只有至多一条路径与之对应，因此降低了处理过程中的时间和空间复杂度。图 4.6 给出了一个确定化操作后的加权有限状态机。

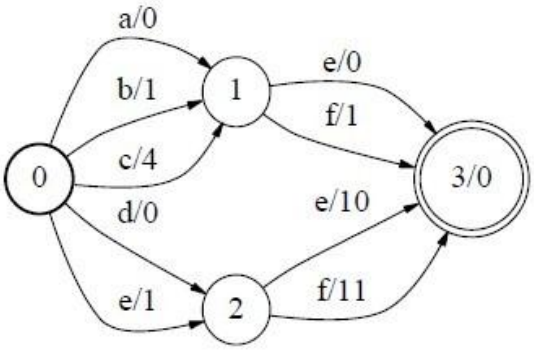


图 4.6 确定化后的 WFSA

在图 4.5 中，输入序列“ab”有两条路径，其权值和为$\{0+0=0,3+10=13\}$，则经过确定化后在图 4.6 中对应的输入序列为“ab”，权值和取其最小值，即为 0。

我们对图 4.4 进行确定性操作，得到图 4.7 的一个 WFST。

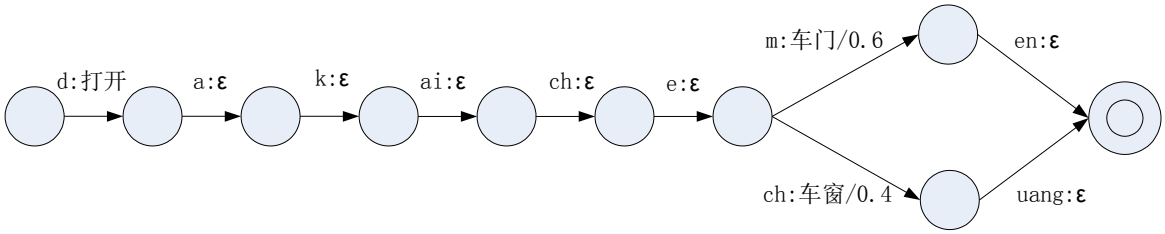


图 4.7 确定化后的 WFST

4.2.3 最小化（Minimization）

确定化后的加权有限状态机能够通过最小化进行进一步的优化。采用经典的最小化算法[43-44]，能够使任意确定性的有限状态机得到优化。同样，任意确定化的加权有限状态机也能够进行最小化处理的优化[45]。最小化处理后的加权有限状态机与处理前的确定化加权有限状态机是等效的，且在所有确定化的加权有限状态机中，其状态数和传递弧的数量都是最少的。

对一个确定化的加权有限状态机，我们可以把一组符号-权值(a,w)看成一个单独的符号，进而把该加权有限状态机当作一个无权值的有限状态机，这样就能用经典的最小化算法进行优化。但实际上，由于符号-权值(a,w)基本互不相同，经典的最小化算法对加权有限状态机是无效的。

为了解决这个问题，我们需要用改进后的最小化算法来进行处理。改进后的算法分为两个步骤：首先，对所有传递弧进行权值前推；然后，将每一组符号-权值看作一个单独的符号，并采用经典的最小化算法来进行优化。

权值前推即对权值进行重新分配。我们考虑在 Tropical 半环结构上的权值前推

方法，同样的方法也能运用在其他的半环结构上。

一个加权有限状态机通过权值的重新分配能够生成有限个与其等效的加权有限状态机。假设加权有限状态机 A 只有一个结束状态 f_A，令 $V:Q\to R$ 是任意状态的势函数。对每个传递弧权值的更新可以表示为

$$w[t]\leftarrow w[t]+(V(n[t])-V(p[t])) \tag{4.3}$$

则最终的权值可以表示为

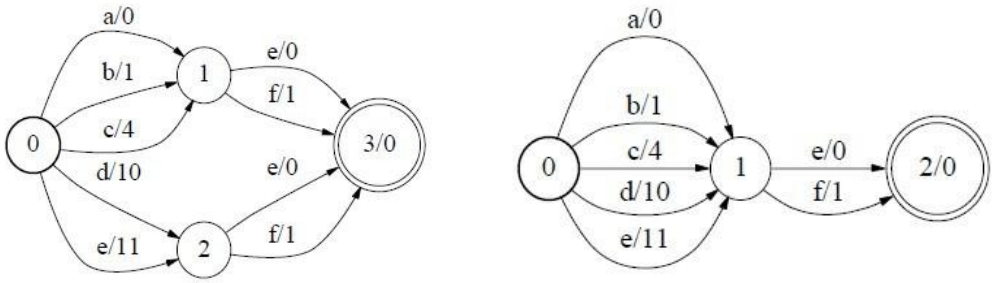
$$\rho(f_A)\leftarrow \rho(f_A)+(V(i_A)-V(f_A)) \tag{4.4}$$

很容易看出，在这种权值重新分配的方法下，每一个从初状态到结束状态路径上的势函数是先增后减的，最终结果是保持不变，即

$$(V(f_A)-V(i_A))+(V(i_A)-V(f_A))=0 \tag{4.5}$$

因此，对一条完整的路径来说，权值的重新分配并不影响其权值之和，且经过权值分配后的自动机与最初的自动机是一种等价的关系。而为了将加权有限状态机 A 中的权值尽可能的往初始状态进行前推，我们需要选择一个特殊的势函数，该势函数必须满足其每个状态到结束状态的路径权值都是最小的。因此，在进行权值前推的步骤后，每个状态到结束状态的最小路径权值都是 0。

图 4.8 给出了一个改进后最小化处理的完整的例子。图 4.8（a）是对图 4.6 中的 WFSA 进行权值前推后的结果，图 4.8（b）则是对（a）进行最小化后得到的优化后的 WFSA。



（a） （b）

图 4.8 最小化 WFSA

4.3 知识源的 WFST 表示

对语音识别中的几种知识源及其网络，我们有以下几种 WFST 的表达形式：

（1）语言模型，WFST 中用 *G* 来表示；

（2）发音词典，用 *L* 表示；

（3）上下文相关因素模型，用 *C* 来表示；

（4）隐马尔可夫模型，用 *H* 来表示；

（5）L 和 G 组成的单音素网络，用 $L \circ G$ 表示；

（6）C、L 和 G 组成的 C-Level 搜索网络，用 $C \circ L \circ G$ 表示；

（7）HMM 构成的 H-Level 搜索网络，用 $H \circ C \circ L \circ G$ 表示。

4.3.1 语言模型（G）

WFST 既可以表示基于规则的语言模型，也可以表示基于统计的语言模型。而在大词汇量连续语音识别任务中，常采用的是 back-off 多元文法语言模型。以一个简单的 back-off bi-gram 语言模型为例，如图 4.9，每个单词都对应一个状态 w_i，对于一个 bi-gram 组合 $w_1 w_2$，其传递弧从状态 w_1 到 w_2，该传递弧对应的权值即为单词 w_1 到 w_2 的转移概率 $p(w_2 \mid w_1)$，这个概率通过对训练数据的统计得到。而对于一个未出现在训练数据中的 bi-gram 组合 $w_1 w_3$，我们则用 back-off 的方法来实现：w_1 先退回到零元文法状态 b，再从 b 转移到 w_1。所以有 w_1 到 w_3 的转移概率为

$$\rho(w_3 \mid w_1) = \beta(w_1)\rho(w_3) \tag{4.6}$$

其中 w_3 为 w_3 的零元文法概率，$\beta(w_1)$ 为 w_1 的 back-off 权值。

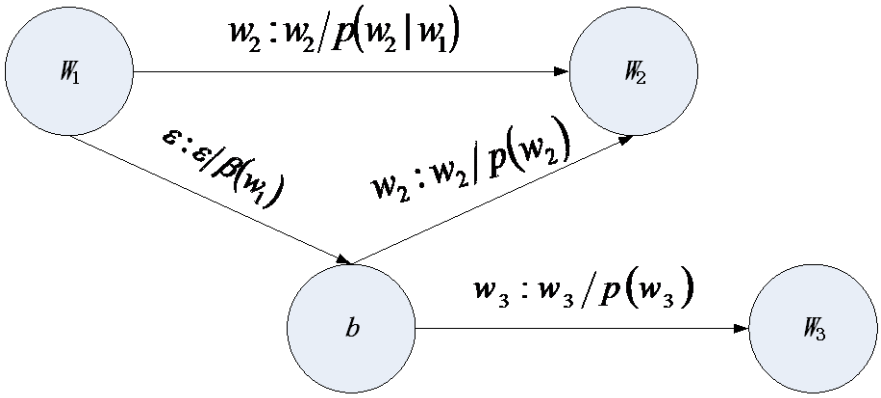


图 4.9 简单的 back-off bi-gram 语言模型

在更高阶的 N-gram 中，我们也采用 back-off 方法的语言模型，当某一事件的频率小于 K 时，用 n-1 元语法来代替 n 元语法，如式（4.6）所示。

$$P(x_n \mid x_1 \cdots x_{n-1}) = \begin{cases} (1-\alpha(f(x_1 \cdots x_{n-1})))\dfrac{f(x_1 \cdots x_n)}{f(x_1 \cdots x_{n-1})}, & \text{当} f(x_1 \cdots x_n) > K \\ \alpha(f(x_1 \cdots x_{n-1}))P(x_n \mid x_2 \cdots x_{n-1}), & \text{当} f(x_1 \cdots x_n) \le K \end{cases} \tag{4.6}$$

其中 α 为归一化因子。

4.3.2 发音词典（L）

对发音词典，首先构造它的克林闭包（Kleene closure），即通过在发音词典的每

个结束状态和初始状态之间连接一个ε跳转来实现。这样得到的发音词典 L 能够使词表中的任何单词序列和与之对应的发音匹配起来。因此，组合$L \circ G$就完成了从音素到单词序列的映射。

对于发音词典中出现的多音字词，我们有两种处理方式：

（1）对多音字的区分，我们引入了音调，即在字的音素序列的结尾引入数字 0-4 来区分发音的轻声和一至四声；

（2）对多音词的区分，在词的音素序列的结尾引入辅助符号$\#_0$， $n = 0,1\cdots$，以此来区分多音词。

采取对多音字词区分的处理方法，其目的是为实现对 WFST 的确定化操作。

4.3.3 上下文相关音素模型（C）

在组合$L \circ G$中，我们只对上下文无关的音素进行了转换。因此，我们还需要将上下文无关的音素序列转换成上下文相关的音素序列，即完成$C \circ L \circ G$的组合。图 4.10 给出了一个上下文相关的三音素 WFST。

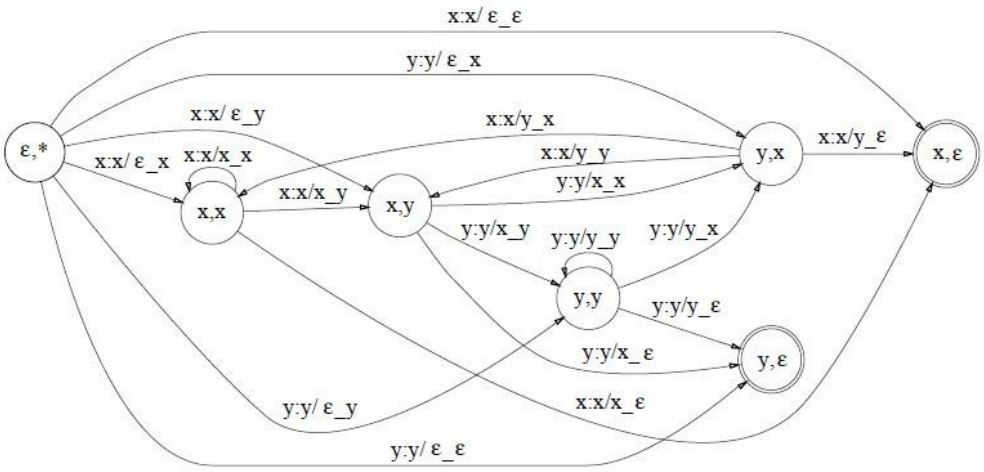


图 4.10 上下文相关三音素 WFST

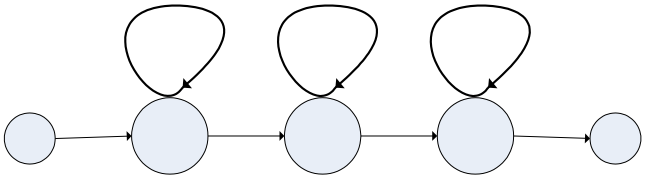
图 4.10 中，我们用 *phone/left context_right context* 的形式来表示上下文相关的音素，每个状态都包含前一个音素 *a* 和后一个音素 *b*，用符号(a,b)表示，ε 表示一个音素序列的开始或者结束，*表示下一个音素不可知。考虑两个上下文无关的音素 *x* 和 *y*，对音素序列 *xyx*，其通过状态$(\varepsilon,*)(x,y)(y,x)(x,\varepsilon)$被 WFST 映射成序列 *x/*$\varepsilon$*_y y/x_x x/y_*$\varepsilon$。通常，当存在 *n* 个上下文无关的音素时，三音素结构会生成一个含$O(n^2)$个状态和$O(n^3)$个传递弧的转换器，四音素结构生成的转换器则会包含$O(n^3)$个状态和$O(n^3)$个传递弧。

在实际应用中，因为文本聚类常用于缓解数据稀疏带来的影响，因此 *n* 个音素

常常共享相同的 HMM 模型，在这种情况下，对上下文相关的 WFST 进行确定化和最小化操作后的效果就变得非常明显。

4.3.4 声学模型（H）

语音识别中常用的 HMM 模型通常为 left-to-right 的链状模型结构，因此，HMM 模型很容易表示成 WFST 的形式。在生成 H 的过程中，通常不会在状态中加入自回弧，而是在解码的时候自动加入。图 4.11（a）是普通音素模型的 3 状态拓扑结构，每个状态带有自回弧且状态之间没有跨越，图 4.11（b）则是其拓扑结构的文本表示方法。



（a） 音素 HMM 模型拓扑结构

```
<Topology>
<TopologyEntry>
<ForPhones> 1 2 3 4 5 6 7 8 </ForPhones>
<State> 0 <PdfClass> 0
<Transition> 0 0.5
<Transition> 1 0.5
</State>
<State> 1 <PdfClass> 1
<Transition> 1 0.5
<Transition> 2 0.5
</State>
<State> 2 <PdfClass> 2
<Transition> 2 0.5
<Transition> 3 0.5
</State>
<State> 3
</State>
</TopologyEntry>
</Topology>
```

（b）拓扑结构的文本表达

图 4.11 3 状态 HMM 音素模型的拓扑结构

4.4 WFSTs 的优化

在将各层知识源表达成 WFST 的形式后，我们需要对组合起来的 $H \circ C \circ L \circ G$ 结构的搜索网络进行优化，与单个的 WFST 处理方法类似，这也分为两个步骤，即确定化操作和最小化操作。

4.4.1 确定化

对组合后的 WFST 进行确定化操作，主要目的是为了清除冗余路径，因此能够大幅度的减少识别时间。同时，这样的处理还能够提升对 WFSTs 组合的效率，减少 WFST 存储空间的大小。

对从音素序列到单词的 WFSTs 组合 $L \circ G$，最明确的问题就是同音字的存在，处理方法在章节 4.3.2 中已详细说明。但即使没有同音字，在整个输入的音素序列没有被完全扫描之前，输出词序列的第一个词是不可知的，这种不受控制的输出延迟就导致了 $L \circ G$ 不是可确定化的。

为解决这种问题，就引入了辅助音素符号 $\#_0$，将其作为每个语言学意义上的词的结束符号，而其他的辅助符号 $\#_1 \cdots \#_{k-1}$ 则用来区分多音字。若多音字的最大阶数为 P，则至多需要 P 个辅助符号。我们用 $\tilde{L}$ 来表示加入这些辅助符号后的发音词典。

相应的，上下文相关音素模型 C 的所有传递路径中也要包含这些新的辅助符号。为进一步在上下文相关音素级和分布级进行确定化操作，每个新加入的辅助符号都必须映射为一个独特的上下文相关级的符号。所以我们在 C 的每个状态中都加入了自回弧，将每个辅助的音素符号都映射到一个新的上下文相关的辅助音素符号上，而这个扩展后的 C 我们则记为 $\tilde{C}$。

同样的，每个上下文相关的辅助音素符号都要映射到一个新的独特的分布式标记上，所以我们在 H 初始状态的输入和输出词表中加入 P 个自回弧来实现这种映射过程，而新得到的 H 我们就记为 $\tilde{H}$。

在得到新的 $\tilde{L}$，$\tilde{C}$，$\tilde{H}$ 后，我们就要对这些 WFST 重新进行组合和确定化。首先是组合得到 $\tilde{L} \circ G$ 并对其进行确定化操作，用 $\det(\tilde{L} \circ G)$ 来表示。原始 $\tilde{L} \circ G$ 的某些状态可能有 V 个传递路径，V 是单词的个数，而得到的 $\det(\tilde{L} \circ G)$ 的每个状态最多只有音素个数的传递路径，因此确定化操作就大大降低了传递网络的复杂度。

然后将 $\tilde{C}$ 和 $\det(\tilde{L} \circ G)$ 组合并进行确定化操作得到 $\det(\tilde{C} \circ \det(\tilde{L} \circ G))$，最后在组合 $\tilde{H}$ 并确定化后得到 $\det(\tilde{H} \circ \det(\tilde{C} \circ \det(\tilde{L} \circ G)))$。在完成整个确定化操作后，整个网络中每一个状态对同一 HMM 模型状态至多只有一条传递路径。

4.4.2 最小化

在进行最小化操作之前，我们还要擦除引入这一过程中的辅助符号，用 ε 替换，这一操作用 π_ε 来表示，而我们可以用下式来表达这一过程：

$$N = \pi_\varepsilon \det(\tilde{H} \circ \det(\tilde{C} \circ \det(\tilde{L} \circ G))) \tag{4.7}$$

式中，N 为进行整合后的识别网络。

移除辅助符号后，再对 N 进行权值最小化操作，这个操作可以在多种半环结构

上实现，且大大减少了网络的复杂度。而在 log 半环结构上进行权值前推操作对标准 Viterbi 解码网络的剪枝操作具有重大影响，能够使权值和大的路径在剪枝操作中不会被清除。

在对整合后的 WFSTs 结构进行优化后，这些 WFSTs 构成了一个特殊的确定性的、最小的 WFST 网络，且任意状态的转移路径权值概率之和为 1。

4.5 本章小节

将加权有限状态转换器理论引入语音识别中，利用加权有限状态转换器的组合操作，可以将不同的知识源加入解码网络之中，提高整个语音识别系统的识别性能；而最小化操作可以大大减小整个网络的冗余度，降低整个搜索网络在时间上和空间上的复杂度，提高解码、识别的速度。在第 5 章中，将创建在 Kaldi 环境中加权有限状态转换器的解码图，并以此为系统完成大词汇量连续语音识别任务中的噪声鲁棒性实验。

第 5 章 系统设计和实验结果

整个实验都在 Linux 环境下进行，为研究 GFCC 在噪声环境下的鲁棒性，还进行了以 MFCC 为特征的对比实验。实验的最终目的是完成对系统的嵌入式移植，因此充分考虑了嵌入式设备的存储和计算能力，对某些实验参数进行修改和压缩，以期能够在嵌入式设备上实现。

5.1 语音数据库

本次实验所有数据均来自 863 大规模连续语音识别计划所录制的标准普通话语音数据集。该数据集有 38 名女性和 38 名男性共 76 个说话人，采样频率为 16kHz，采样精度为单声道 16 bits，wav 格式，PCM 编码。选取其中的 32 名女性 32 名男性共 64 人 42 小时时长数据作为训练集，2 名女性 2 名男性共 4 人作为测试集。

实验中的噪声数据来自于两部分：

（1）背景说话人噪声（Babble noise）、汽车噪声（Volvo 340 noise）和坦克噪声（Leopard M109）来自标准噪声库 Noisex-92[46]，其在 1992 年由 Institute for Perception-TNO 和 Speech Research Unit 联合录制而成。该噪声库的噪声采样率都为 19.98kHz，因此使用工具 Cool Edit Pro 将其进行降采样到 16kHz。

（2）其它人工噪声，通过 SoX[47]工具合成。同时，SoX 工具还用于语音/噪声信号的混合。生成和混合的语音采样率均为 16kHz，16bit 量化，PCM 编码。

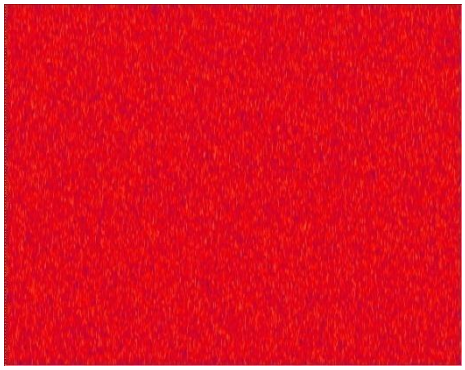
对于各种噪声，我们通过改写 SoX 参数将其处理成 0dB、5dB、10dB、15dB、20dB、25dB、30dB 一共 7 个噪声级别并与纯净的语音信号进行混合，以此来测试 GFCC 和 MFCC 在不同噪声等级下的识别性能表现。

5.2 噪声分析

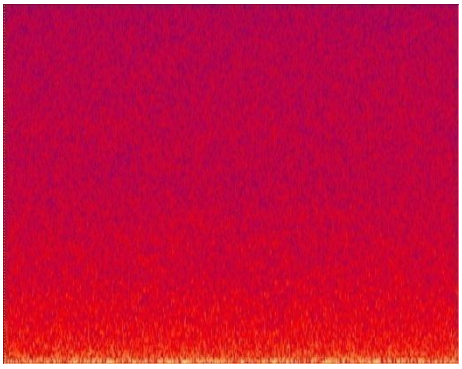
通过对信噪比为 0dB 时各种噪声的能量分布图的分析，可以将测试噪声分为以下三大类。

（1）白噪声、粉红噪声和褐色噪声。

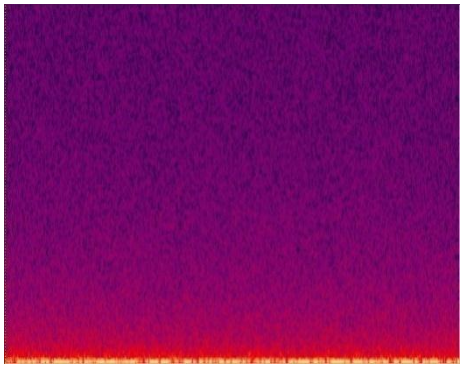
这三种噪声本质上都是一种宽频谱信号，见图 5.1，其中横向为时间轴，纵向为频率轴，频率分布为 0Hz~8000Hz。



（a）白噪声



（b）粉红噪声



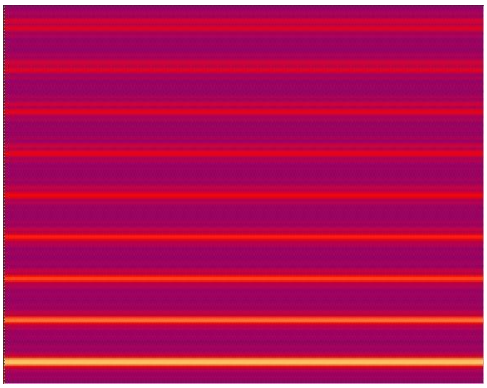
（c）褐色噪声

图 5.1 白噪声、粉红噪声和褐色噪声能量分布图

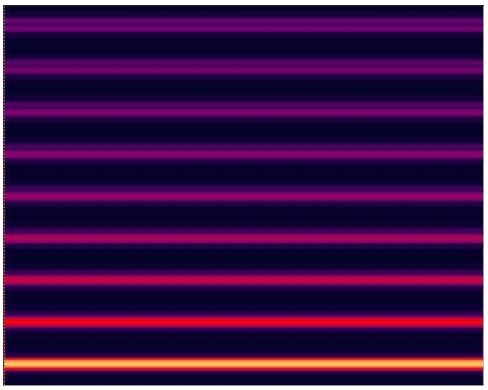
可以直观的发现，白噪声的能量在整个坐标域上都是均匀分布的；粉红噪声的能量主要集中在中低频段，随着频率增大，能量下降为$1/f$；褐色噪声的能量主要集中在低频段，随着频率增大，能量下降为$1/f^2$。

（2）矩形波、三角波、梯形波、锯齿波和指数波噪声。

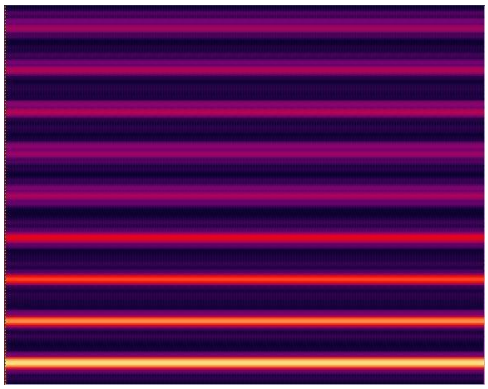
这几种波形都属于窄带信号，大部分能量都集中在一个很窄的频率范围内，如图 5.2 所示。



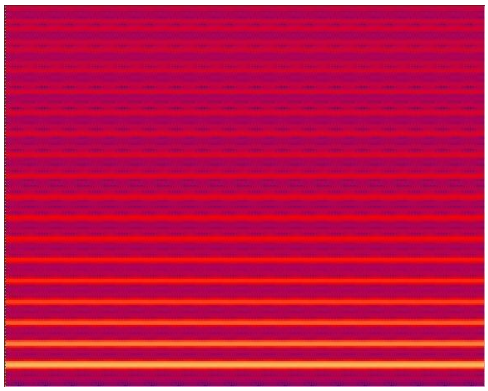
（a）矩形波噪声



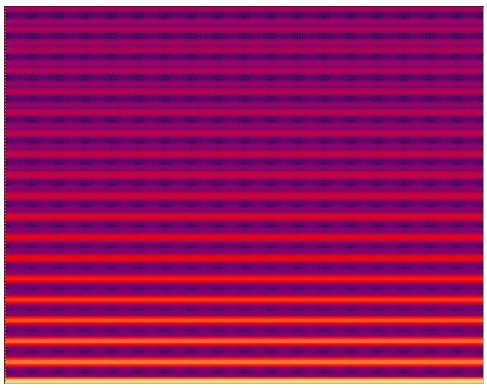
（b）三角波噪声



（c）梯形波噪声



（d）锯齿波噪声



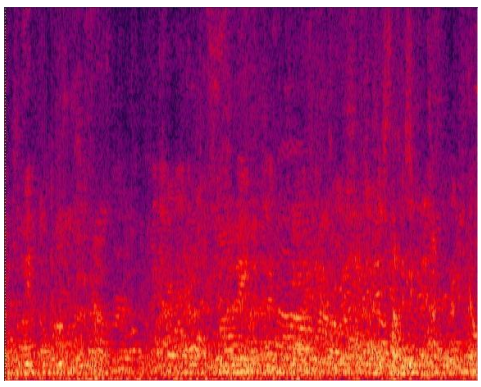
（e）指数波噪声

图 5.2 几种窄带噪声能量分布图

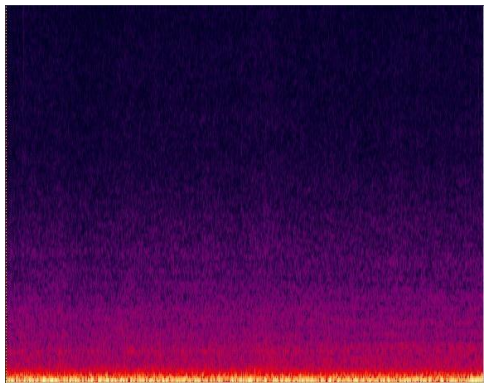
从图 5.2 中可以看到，除了正弦波外，其他几种噪声的能量每隔一段频率就有一个很集中的能量分布在很窄的频率段内，且在其他频率各自还有不同强度的能量分布。

（3）背景说话人噪声、汽车噪声和坦克噪声。

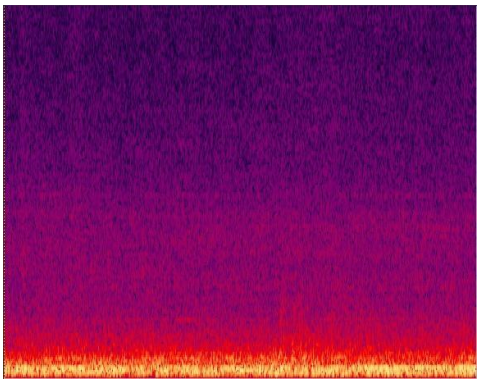
这三种噪声都是通过现场设备采集录制获得，其能量分布图如图 5.3 所示。



（a）背景说话人噪声



（b）汽车噪声



（c）坦克噪声

图 5.3 几种现场噪声的能量谱图

从图 5.3 中可以看到，这三种噪声的能量分布都具有很强的随机性，特别是背景说话人噪声的能量部分，无论在时间上还是在频率上，都是随机的；汽车噪声和坦克噪声由于内部环境相对较安静，因此大部分能量都集中在低频区域。

5.3 实验设置

5.3.1 声学模型训练

本实验中的声学模型为基于 left-to-right 3 状态的上下文相关音素（tri-phone）HMM 模型。模型中含有 218 个单音素（包括静音），其中元音音素对应带音调韵母。

声学模型的训练采用基于最大互信息量的区分性训练（bMMI，见章节 2.5.4），训练环境为 Kaldi 工具包，训练后的最终模型中含有约 4000 个共享状态，15000 个高斯分布。

在训练环节的特征处理中，我们将基于说话人的倒谱均值方差归一（CMVN）技术用于对特征进行信道规整，最大似然线性变换（MLLT）技术用于消除特征各维度之间的相关性。

5.3.2 语言模型训练

语言模型为 2 万词表的 back-off 三元统计文法模型，由中文 Gibbytes 数据库训练而成。

为降低数据稀疏带来的影响，我们需要对训练好的语言模型进行平滑处理。采用 Witten-Bell discounting 方法用于模型平滑，其基本思想是把在训练样本中出现过的事件的概率适当减小，把减小得到的概率密度分配给训练语料中没有出现过的事件，换句话来说，即修改训练样本中事件的实际计数，使训练样本中不同事件的概率之和小于 1，剩余的概率量分配给未见概率。同时，我们还采用了低通 interpolation

的方法进行模型平滑。

为适应嵌入式平台的移植要求，减小语言模型占用的存储空间，我们对语言模型进行了剪枝操作，剪枝至最小概率为10^{-7}，经过剪枝处理后，将整个语言模型压缩到 25Mb 左右。

采用 SRILM 工具包用于语言模型的训练[48]。

5.3.3 创建解码图

本文实验中使用的 FST 工具包为 OpenFst[49]，并在 Kaldi 环境[50]环境下构建一个完整的 WFST 解码图系统。

1.准备两个最初的符号文本 words.txt 和 phones.txt。其中 words.txt 是包括符号ε（即$\#_0$）；phones.txt 是不包括ε的，但在生成 L.fst（即发音词典 L 的 WFST）后，新生成的 phones_disambig.txt 中会包含对多音字消除歧义的辅助符号。

2.准备发音词典 L。用于训练的 L 是不含有辅助符号的，而用于生成解码图的 L 则包含辅助符号。

（1）将不含有辅助符号的 L 转换成 WFST 的形式，其指令如图 5.4，其中静音的输出概率是 0.5。

```
scripts/make_lexicon_fst.pl data/lexicon.txt 0.5 SIL | \
  fstcompile --isymbols=data/phones.txt --osymbols=data/words.txt \
  --keep_isymbols=false --keep_osymbols=false | \
   fstarcsort --sort_type=olabel > data/L.fst
```

图 5.4 无辅助符号 L 的 WFST 转换

（2）将加入辅助符号后的 L 转换成 WFST 的形式，见图 5.5。

```
phone_disambig_symbol=`grep \#0 data/phones_disambig.txt | awk '{print $2}'`
word_disambig_symbol=`grep \#0 data/words.txt | awk '{print $2}'`

scripts/make_lexicon_fst.pl data/lexicon_disambig.txt 0.5 SIL  | \
   fstcompile --isymbols=data/phones_disambig.txt --osymbols=data/words.txt \
   --keep_isymbols=false --keep_osymbols=false |   \
   fstaddselfloops  "echo $phone_disambig_symbol |" "echo $word_disambig_symbol |" | \
   fstarcsort --sort_type=olabel > data/L_disambig.fst
```

图 5.5 含辅助符号 L 的 WFST 转换

在完成这一步后，L 中的每次词序列和其音素序列将呈现一一对应的关系，而我们将得到一个如图 5.6 的发音词典。

```
的 d e0
的 d i4
的 d i2
在 z ai4
和 h e4
和 h uo2
和 h uo4
和 h u2
和 h e2
```

图 5.6 发音词典

3.准备语言模型 G。首先从 WFST 移除这些辅助符号；然后确保在语言模型中没有词表之外的单词出现；接着在句子的开始和结尾移除非法的序列；最后用辅助符号$\#_0$替代ε。

4.组合 L 和 G。对组合后的 LG 移除符号ε，并进行最小化操作。在最小化操作中，移除权值前推的处理方法，以此来保留随机性。

5.组合 CLG。

（1）首先生成 C 的 WFST，见图 5.7。在 C 中，存在$N-1$个状态，对应所有可能出现的音素。

```
fstmakecontextfst --read-disambig-syms=$dir/disambig_phones.list \
--write-disambig-syms=$dir/disambig_ilabels.list data/phones.txt $subseq_sym \
  $dir/ilabels | fstarcsort --sort_type=olabel > $dir/C.fst
```

图 5.7 生成 C 的 WFST

（2）动态组合 C，生成 CLG.fst，见图 5.8。

```
fstcomposecontext  --read-disambig-syms=$dir/disambig_phones.list \
                   --write-disambig-syms=$dir/disambig_ilabels.list \
                   $dir/ilabels < $dir/LG.fst >$dir/CLG.fst
```

图 5.8 组合 CLG

（3）减少上下文相关输入符号的数量。在生成 CLG.fst 后，我们选择对生成的 WFSTs 进行压缩，能够达到 5%-20%的压缩量。

6.组合 HCLG。

（1）生成 H 的 WFST，调用见图 5.9。此时生成的 WFST 我们命名为 Ha.fst，因为此时其结构中并没有自回弧的存在。

```
make-h-transducer --disambig-syms-out=$dir/disambig_tstate.list \
   --transition-scale=1.0  $dir/ilabels.remapped \
   $tree $model  > $dir/Ha.fst
```

图 5.9 生成 H 的 WFST

（2）生成缺少自回弧的 HCLGa.fst，见图 5.10。

```
fsttablecompose $dir/Ha.fst $dir/CLG2.fst | \
 fstdeterminizestar --use-log=true | \
 fstrmsymbols $dir/disambig_tstate.list | \
 fstrmepslocal  | fstminimizeencoded > $dir/HCLGa.fst
```

图 5.10 无自回弧的 HCLG

（3）在 HCLGa.fst 中加入自回弧，生成 HCLG.fst，如图 5.11。自回弧尺度是一个对数概率，可表达为 self-loop-scale $*\log(P)$，而对该状态的所有其它对数转移概率，则是 self-loop-scale $*\log(1-P)$。

```
add-self-loops --self-loop-scale=0.1 \
  --reorder=true $model < $dir/HCLGa.fst > $dir/HCLG.fst
```

图 5.11 含自回弧的 HCLG

5.3.4 特征提取

在实验中，需要对语音数据分别提取其 GFCC 特征和 MFCC 特征。为体现出对比性，对 MFCC 和 GFCC 的提取均采用 25 毫秒的帧长和 10 毫秒的帧移，特征向量包括 13 维的倒谱系数和一阶、二阶差分，共计 39 维的特征向量。

GFCC 特征提取采用时域 GFCC 提取工具包[51]，并对其做出修改以满足 Kaldi 工具包的调用。MFCC 特征提取用 Kaldi 工具包来完成。

完成特征提取后，需要分别对 MFCC 和 GFCC 为特征的解码任务训练不同的解码图，这个任务也是用 Kaldi 工具包来完成。

5.4 实验结果

对 GFCC 和 MFCC，分别在纯净语音数据和噪声语音数据下进行了对比实验，评价识别性能的指标为词错误率（WER，Word Error Rate）[52]。

5.4.1 纯净语音对比实验

用 32 男 32 女共 64 人数据作为训练集，2 男 2 女 4 人数据作为集进行实验。实验结果见表 5.1。

表 5.1 纯净语音实验结果

	MFCC	GFCC
	WER：17.20%	WER：15.82%
测试集	共 27207 词，识别错误 4679 词，47 插入错误,1013 删除错误，3619 替代错误	共 27207 词,识别错误 4303 词，41 插入错误,956 删除错误,3306 替代错误

在纯静语音的测试数据下，GFCC 比 MFCC 有更好的识别效果，词错误率降低了 1.38%。

5.4.2 带噪语音对比实验

用 32 男 32 女共 64 人的纯净语音数据作为训练集来训练生成解码图，用 2 男 2 女共 4 人的纯净语音数据混入噪声后作为测试集。为方便分析对比，按噪声分类将 GFCC 和 MFCC 的噪声鲁棒性测试结果分为表 5.2、表 5.3 和表 5.4。

表 5.2 分类一：带噪语音实验结果

Noise	Feature	WER(%)						
		30dB	25dB	20dB	15dB	10dB	5dB	0dB
白噪声	MFCC	18.29	19.65	27.07	34.46	49.14	73.51	93.81
	GFCC	**17.22**	**18.12**	**21.20**	**29.25**	**45.55**	**71.37**	**91.27**
粉红噪声	MFCC	17.37	18.10	19.56	24.17	35.56	64.37	91.54
	GFCC	**16.36**	**17.08**	**18.62**	**23.33**	**35.29**	**63.41**	**87.57**
褐色噪声	MFCC	17.06	17.09	17.16	17.20	17.42	18.40	**21.88**
	GFCC	**15.75**	**15.79**	**15.83**	**16.29**	**16.97**	**18.12**	22.26

表 5.3 分类二：带噪语音实验结果

Noise	Feature	WER(%)						
		30dB	25dB	20dB	15dB	10dB	5dB	0dB
矩形波噪声	MFCC	19.62	22.53	27.22	34.39	44.81	**57.88**	**74.48**
	GFCC	**17.97**	**20.17**	**24.16**	**31.35**	**43.09**	60.21	79.94
三角波噪声	MFCC	18.91	21.36	25.25	31.40	39.39	**48.83**	**58.41**
	GFCC	**17.22**	**19.18**	**21.94**	**27.40**	**36.73**	49.41	61.86
锯齿波噪声	MFCC	19.61	23.08	29.97	40.60	55.86	71.63	90.13
	GFCC	**17.37**	**19.37**	**22.91**	**30.13**	**44.99**	**67.42**	**88.64**
梯形波噪声	MFCC	19.88	22.91	28.89	37.53	**48.45**	**60.54**	**71.95**
	GFCC	**17.99**	**20.41**	**25.25**	**34.67**	50.08	67.61	82.48
指数波噪声	MFCC	17.61	19.10	22.76	31.23	45.01	58.34	70.92
	GFCC	**16.34**	**17.12**	**18.62**	**21.80**	**30.84**	**49.09**	**67.78**

表 5.4 分类三：带噪语音实验结果

Noise	Feature	WER(%)						
		30dB	25dB	20dB	15dB	10dB	5dB	0dB
背景说话人噪声	MFCC	17.26	17.78	18.93	22.20	33.98	61.88	87.83
	GFCC	**16.07**	**16.35**	**17.62**	**20.80**	**30.04**	**55.00**	**83.53**
汽车噪声	MFCC	17.15	17.35	17.54	17.65	18.10	18.34	**19.02**
	GFCC	**15.90**	**16.03**	**16.32**	**16.67**	**17.25**	**18.18**	19.35
坦克噪声	MFCC	17.24	17.41	17.61	18.08	20.64	26.48	37.88
	GFCC	**15.85**	**16.20**	**16.57**	**17.30**	**18.44**	**21.73**	**29.81**

从 3 个表中数据可以看出，不论是 GFCC 还是 MFCC，噪声的引入都会严重影响语音识别的正确率；在多种带噪语音的识别结果中，GFCC 的识别性能普遍高于 MFCC。

从表 5.2 数据和图 5.1（a）（b）可以看出，对于白噪声和粉红噪声，由于其频带分布宽，语音的谐振信息在各个频段都受到干扰和破坏，因而对语音识别性能的影响极为显著。当 SNR 值低至 10 时，语音信息受到破坏的程度相当明显，语音识别的错误率会急剧增大；当 SNR 值低到 0 时，语音信息几乎完全被噪声所淹没，语音识别的错误率接近 100%；而在 SNR 值大于 10 时，我们还能从语音信号中提取出较多的声音信息，识别率也在一个较能接受的水平，在这一 SNR 值区间，GFCC 的识别性能明显要高于 MFCC。

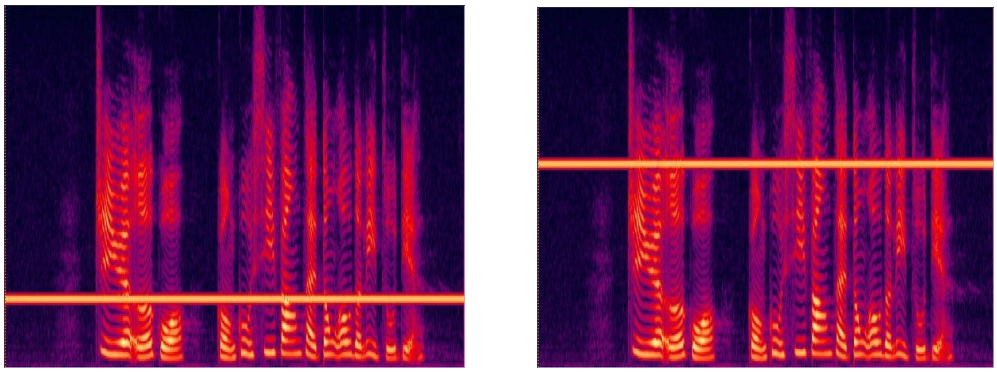
对于窄带噪声（见图 5.2），其主要的能量分布都集中在某几个很窄的频率区间内。从表 5.3 的数据可以看出，在 SNR 较高时，GFCC 都有优于 MFCC 的性能；当 SNR 下降（10dB 以下）时，对能量集中的频带较密集的噪声（如锯齿波和指数波），可近似为能量全频段分布的噪声，在这种条件下，GFCC 的性能仍高于 MFCC；而对能量集中的频带较稀疏的噪声（如矩形波、三角波和梯形波），当 SNR 值很低时，噪声对某一频段内的声音信息破坏严重，对 GFCC 造成了很大的影响，此时 MFCC 的性能略高于 GFCC（即使两者的识别性能都不能满足应用要求）。

对于比较典型的环境噪声，如背景说话人噪声，更接近于实际应用的环境，从表 5.4 数据可以看出，GFCC 的性能是明显高于 MFCC 的；对于汽车噪声和坦克噪声（见图 5.3（b）（c）），其噪声的能量主要集中在低频段，在进行特征提取的过程中，我们已经对这一频段进行了过滤处理，因此即使在 SNR 值很低的情况下，GFCC 和 MFCC 依然可以得到 60%以上的识别率，且 GFCC 的性能要优于 MFCC。

从噪声角度分析，当噪声对信号的中高频部分破坏比较严重时，语音识别系统受到的影响比较大，这是因为 MFCC 对高频依赖于发现共振峰信息本身的敏感性，而 GFCC 更多的是模拟人耳的听觉特性。为验证这个想法，本文设计了 GFCC/MFCC 不同频段抗噪的对比实验。

5.4.3 不同频段抗噪对比实验

为验证不同频段、不同强度的噪声对 MFCC 和 GFCC 性能的影响，我们利用正弦波噪声能量分布单一且集中的特性，将不同频段的正弦噪声与纯静语音信号混合，以分析 GFCC 和 MFCC 对不同频率噪音的抵抗能力。混合后的语音能量分布如图 5.12 所示。



（a）1500Hz （b）4500Hz

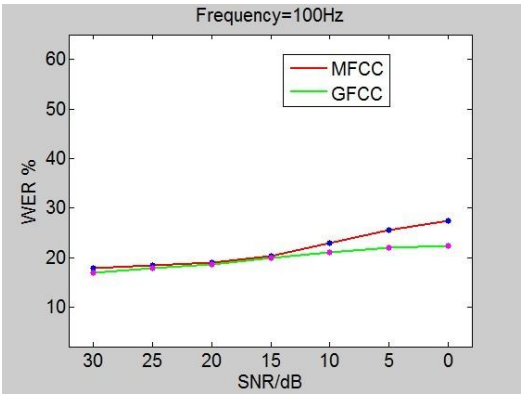
图 5.12 SNR=0dB 时不同频率正弦带噪语音能量分布

考虑在进行特征提取的过程中，我们将滤波器组的下限中心频率设置为 80Hz，上限中心频率设置为 5000Hz，以此为依据，我们调整合成参数，令正弦波的频率分

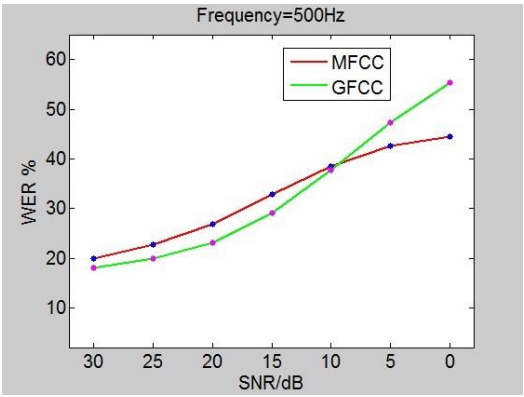
别为 100Hz、500Hz、1000Hz、1500Hz、2000Hz、2500Hz、3000Hz、3500Hz、4000Hz、4500Hz，共设置了 10 组实验，结果见表 5.5 和图 5.13。

表 5.5 不同频率正弦波噪声实验结果

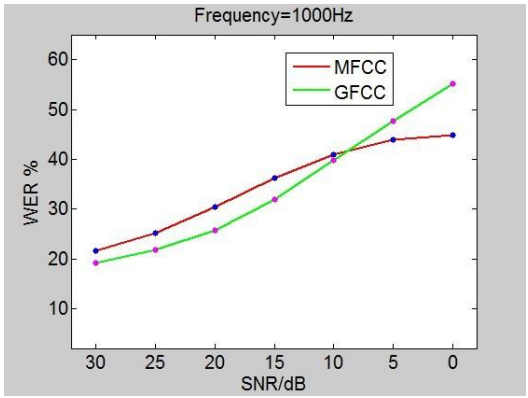
Frequency	Feature	WER(%)						
		SNR=30dB	25dB	20dB	15dB	10dB	5dB	0dB
100Hz	MFCC	17.90	18.50	18.97	20.33	22.88	25.50	27.35
	GFCC	**16.96**	**17.93**	**18.58**	**19.98**	**21.04**	**21.95**	**22.43**
500Hz	MFCC	19.87	22.66	26.88	32.90	38.43	**42.61**	**44.40**
	GFCC	**18.00**	**19.94**	**23.17**	**29.09**	**37.70**	47.25	55.27
1000Hz	MFCC	21.55	25.09	30.33	36.13	40.88	**43.83**	**44.78**
	GFCC	**19.22**	**21.72**	**25.79**	**31.95**	**39.77**	47.58	55.08
1500Hz	MFCC	20.27	22.65	**25.73**	**29.24**	**31.43**	**33.05**	**34.00**
	GFCC	**18.75**	**21.45**	26.10	32.19	42.32	55.02	67.00
2000Hz	MFCC	20.51	22.31	**25.12**	**27.82**	**30.36**	**32.73**	**34.16**
	GFCC	**19.36**	**21.88**	26.29	33.94	43.06	52.91	62.86
2500Hz	MFCC	20.79	**23.40**	**27.04**	**30.61**	**35.16**	**36.64**	**35.70**
	GFCC	**19.63**	26.54	31.87	37.48	41.79	45.66	50.11
3000Hz	MFCC	18.32	**18.97**	**19.69**	**20.61**	**21.10**	**21.16**	**21.52**
	GFCC	**16.45**	19.49	28.63	32.47	34.79	37.76	40.24
3500Hz	MFCC	**18.44**	**20.98**	**23.71**	**24.73**	**25.41**	**25.93**	**26.10**
	GFCC	21.91	24.23	28.14	32.27	35.10	37.31	39.71
4000Hz	MFCC	**18.52**	**19.18**	**21.53**	**22.64**	**22.97**	**23.52**	**23.97**
	GFCC	23.09	24.37	28.21	29.11	35.07	37.10	39.05
4500Hz	MFCC	**19.04**	**22.02**	**23.17**	**23.27**	**23.71**	**24.04**	**24.28**
	GFCC	19.90	23.17	28.92	31.34	33.04	35.04	36.26



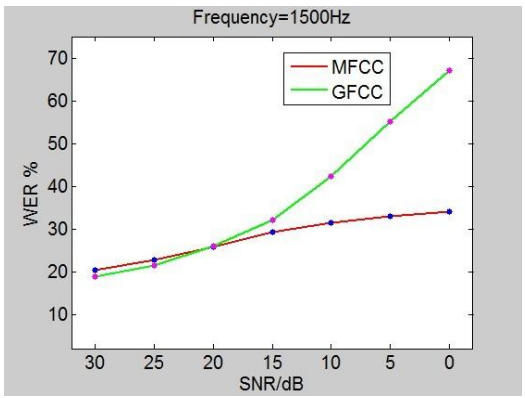
（a）100Hz



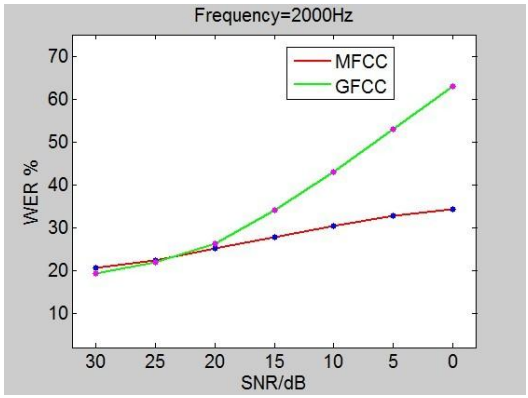
（b）500Hz



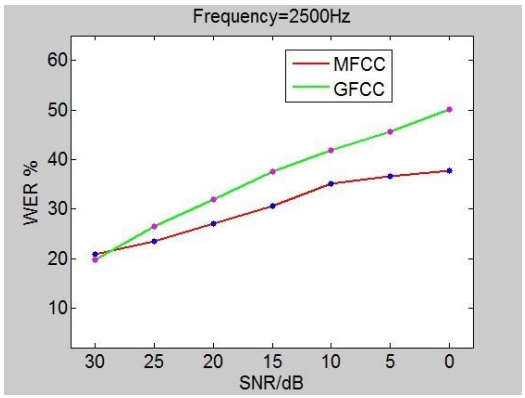
（c）1000Hz



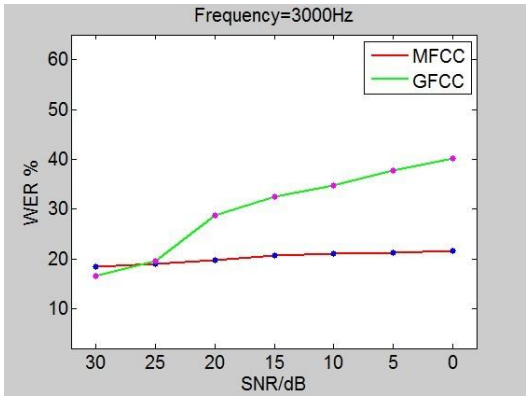
（d）1500Hz



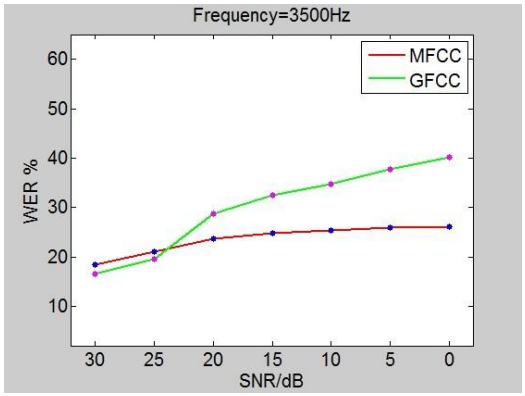
（e）2000Hz



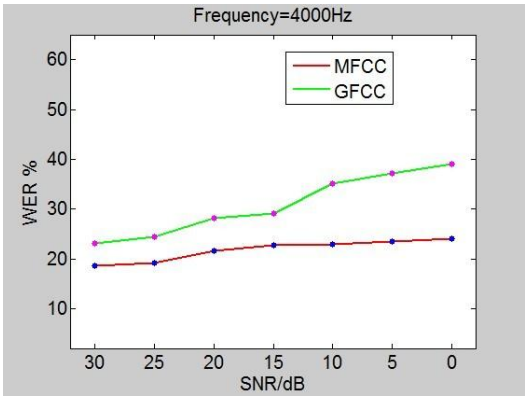
（f）2500Hz



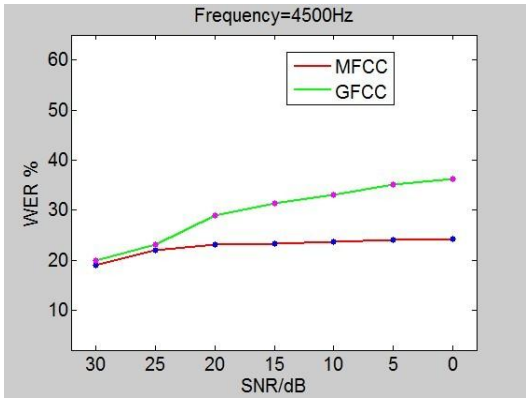
（g）3000Hz



（h）3500Hz



（l）4000Hz



（m）4500Hz

图 5.13 不同频率正弦波噪声实验结果曲线图

从实验结果中可以看出，当存在较低频的噪声时（100~1000Hz），GFCC 较 MFCC 可得到更好的识别性能；而当噪声频率比较高时，在 SNR 较大的环境下，GFCC 有近似或略优于 MFCC 的识别性能，而在 SNR 较小的环境下，GFCC 与 MFCC 相比有较大的差距。结合章节 5.4.2 中的实验数据，我们认为这可能要归因于人说话频率主要集中在 300Hz~700Hz 之间，而人耳对 2000Hz~5000Hz 的频率范围感受力最强。由于 GFCC 倾向于模拟人耳听觉系统的频率响应，因而中高频的噪音对特征的破坏相对比较严重，对识别结果的影响较大。

基于这种分析，在进行特征提取时，我们要考虑对 MFCC 和 GFCC 设计不同的前端滤波器，通过对不同频段的噪声信号进行降噪处理来提高各自的抗噪性能。

第 6 章 总结与展望

6.1 工作总结

本文针对嵌入式语音识别实际应用中的具体情况，主要做了以下两方面工作：

第一，对语音识别系统在实际噪声环境中识别性能下降的问题，研究用噪声鲁棒的特征 GFCC 代替传统的 MFCC 特征。由于 Gammatone 滤波器组的滤波特性，能够较好的滤过低频和高频噪声中对识别性能影响较大的能量，因此 GFCC 对噪声具有较好的鲁棒性。实验证明，在相同信噪比的噪声测试语音条件下，GFCC 比 MFCC 具有明显更高的识别率。同时，基于对嵌入式移植的考虑，采用了时域提取 GFCC 和 DCT 运算，相比于频域 MFCC 提取和 FFT 运算的方法，前者的计算量更小，相应的计算速度更快、实时性更高，因此更适合被用于嵌入式语音识别平台的移植。

第二，采用基于加权有限状态机的思想，用加权有限状态转换器作为基本单位，构建一个加权有限状态解码网络。相比于传统的基于 HMM 模型的 Viterbi 解码方法，我们可以通过对各个知识源的加权有限状态转换器进行组合、最小化等操作，压缩解码网络，并通过剪枝的方法消除小概率传递路径的影响、通过平滑处理解决未出现在训练文本中的识别词概率分配问题。

6.2 未来展望

本文只是对 GFCC 的噪声鲁棒性及 GFCC 和加权有限状态转换解码图的嵌入式可移植性进行了一个初步研究，在整个基于嵌入式平台的语音识别系统上，我们还有很多工作要做：

1.本文对 GFCC 特征只做了 CMVN 一个较简单的特征变换。针对嵌入式语音识别系统的噪声应用环境，在对含噪语音信号进行特征处理时，还可以借鉴对 MFCC 进行降噪处理的技术，加入更多的鲁棒性方法，来减小噪声对识别性能的干扰。

2.将时域分析的方法应用在嵌入式语音识别任务中，是一个很好的思路。直接在时域进行语音信号处理，避免了传统的将时域信号变换到频域后再处理这一费时费力的过程，可以大大提高嵌入式语音识别的实时性，这对嵌入式语音识别任务来说是很重要的一点。因此可以在对 GFCC 的时域分析方法上进行进一步探索，在保留足够性能的基础上，进一步减少计算量，加快识别速度。

3.通过对 GFCC 和 MFCC 在多种噪声环境下的比对试验结果分析，可以发现 GFCC 和 MFCC 对不同频段能量分布的噪声的性能各有不同，考虑可以通过特征融

合的方法，对 GFCC 各维度的特征分量进行分析，去劣存优，提高 GFCC 在某些噪声环境下的鲁棒性。

4.对加权有限状态转换解码图的研究还有待深入，可以对特定的识别环境，在模型训练时加入更多的约束条件，优化解码网络，从而提高识别的速度和识别结果的准确性。

致 谢

衷心感谢李银国老师、清华大学语音和语言技术中心的郑方老师、王东老师几年来给我的悉心指导与关怀，这三位老师无论在学习上还是生活上都给予了我莫大帮助，并且在和这三位老师的相处交流中，他们无论是在学术还是品德上，都让我十分敬佩，他们的教诲也会让我终身受益。在此，谨向两位老师致以最诚挚的谢意。

此外，感谢清华大学语音和语言技术中心的所有同学给予我的帮助和鼓励，特别感谢别凡虎和刘超同学给予我的帮助和建议。感谢重庆邮电大学汽车电子实验中心的程安宇老师、徐洋老师对我学习生活的关心和照顾。感谢杨丽坤、储文、万吉权等同学对我的包容和与照顾。

感谢我的父母家人，你们对我的支持和爱是我前进的动力。

最后，衷心感谢在百忙之中评阅论文和参加答辩的各位专家、教授！

硕士期间从事的科研工作

主要从事的科研工作

“核高基”国家科技重大专项——汽车电子控制器嵌入式软件平台研发及产业化（2009ZX01038-002-002-2）

重庆市科委自然科学基金项目——面向车载智能终端的语音识别控制技术与算法研究（CSTC2012JJA60002）

清华大学国家实验室项目——面向移动设备的语音 QA 系统研究（042003107）

发表的论文

1.欧阳希子，李银国，郑方.基于 Gammatone 特征的鲁棒语音识别研究.第十二届全国人机语音通讯学术会议（NCMMSC 2013）.（EI 会议，已录用）

2.李银国，欧阳希子，郑方.语音识别中听觉特性的噪声鲁棒性分析.清华大学学报（自然科学版）.（EI 期刊，已录用）

参考文献

[1] Furui S. 50 years of progress in speech and speaker recognition[C]//10th International Conference on Speech and Computer-SPECOM, Patras, Greece. 2005: 1-9.

[2] Davis K H, Biddulph R, Balashek S. Automatic recognition of spoken digits[J]. The Journal of the Acoustical Society of America, 1952, 24: 637.

[3] Itakura F, Saito S. A statistical method for estimation of speech spectral density and formant frequencies[J]. Electronics and Communications in Japan, 1970, 53: 36-43.

[4]Vintsyuk T K. Speech discrimination by dynamic programming[J]. Cybernetics and Systems Analysis, 1968, 4(1): 52-57.

[5]Velichko V M, Zagoruyko N G. Automatic recognition of 200 words[J]. International Journal of Man-Machine Studies, 1970, 2(3): 223-234.

[6]Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Proceedings of the IEEE, 1989, 77(2): 257-286.

[7]Lee K F, Hon H W, Reddy R. An overview of the SPHINX speech recognition system[J]. Acoustics, Speech and Signal Processing, IEEE Transactions on, 1990, 38(1): 35-45.

[8] Young S, Evermann G, Gales M, et al. The HTK book[J]. Cambridge University Engineering Department, 2002, 3.

[9]Yu D, Deng L, Dahl G. Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition[C].Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning. 2010.

[10]杨大利, 徐明星, 吴文虎. 语音识别特征参数选择方法研究[J]. 计算机研究与发展, 2003, 7.

[11] Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences[J]. Acoustics, Speech and Signal Processing, IEEE Transactions on, 1980, 28(4): 357-366.

[12] 张智星.语音信号处理与辨识[M].

[13]甄斌, 吴玺宏, 刘志敏, 等. 语音识别和说话人识别中各倒谱分量的相对重要性[J]. 北京大学学报 (自然科学版), 2001, 37(3): 371-378.

[14] Furui S. Speaker-independent isolated word recognition using dynamic features

of speech spectrum[J]. Acoustics, Speech and Signal Processing, IEEE Transactions on, 1986, 34(1): 52-59.

[15] 刘聪. 声学模型区分性训练及其在 LVCSR 系统的应用[D].中国科学技术大学,2010.

[16] Jolliffe I T. Principal component analysis[M]. New York: Springer-Verlag, 1986.

[17] Hunt M J, Lefebvre C. A comparison of several acoustic representations for speech recognition with degraded and undegraded speech[C]//Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on. IEEE, 1989: 262-265.

[18] Kumar N, Andreou A G. Investigation of silicon auditory models and generalization of linear discriminant analysis for improved speech recognition[D]. Johns Hopkins University, 1997.

[19] Rabiner L R. On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition[J]. The Bell System Technical Journal, 1983, 62(4): 1075-1105.

[20] Baum L E. An equality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes[J]. Inequalities, 1972, 3: 1-8.

[21] Woodland P C, Povey D. Large scale discriminative training of hidden Markov models for speech recognition[J]. Computer Speech & Language, 2002, 16(1): 25-47.

[22] Nadas A. A decision theorectic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood[J]. Acoustics, Speech and Signal Processing, IEEE Transactions on, 1983, 31(4): 814-817.

[22]Jelinek F. The development of an experimental discrete dictation recognizer[M].Informatik-Anwendungen—Trends und Perspektiven. Springer Berlin Heidelberg, 1986: 109-117.

[24] 秦健. N-gram 技术在中文词法分析中的应用研究 [D]. 中国海洋大学, 2009.

[25] Katz S. Estimation of probabilities from sparse data for the language model component of a speech recognizer[J]. Acoustics, Speech and Signal Processing, IEEE Transactions on, 1987, 35(3): 400-401.

[26] Ney H, Essen U, Kneser R. On structuring probabilistic dependences in stochastic language modelling[J]. Computer Speech and Language, 1994, 8(1): 1-38.

[27] Van Compernolle D. Increased noise immunity in large vocabulary speech recognition with the aid of spectral subtraction[C] Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'87. IEEE, 1987, 12: 1143-1146.

[28] Berouti M, Schwartz R, Makhoul J. Enhancement of speech corrupted by acoustic noise[C].Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'79. IEEE, 1979, 4: 208-211.

[28]Viikki O, Laurila K. Cepstral domain segmental feature vector normalization for noise robust speech recognition[J]. Speech Communication, 1998, 25(1): 133-147.

[30] Leggetter C J, Woodland P C. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models[J]. Computer speech and language, 1995, 9(2): 171.

[31] Gauvain J L, Lee C H. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains[J]. Speech and Audio Processing, IEEE Transactions on, 1994, 2(2): 291-298.

[32] Bahl L, Brown P, De Souza P, et al. Maximum mutual information estimation of hidden Markov model parameters for speech recognition[C]//Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86. IEEE, 1986, 11: 49-52.

[33] Valtchev V. Discriminative methods in HMM-based speech recognition[M]. University of Cambridge, 1995.

[33]Patterson R D, Moore B C. Frequency Selective in Hearing, Chapter Auditory Filters and Excitation Patterns as Representations of Frequency Resolution[C]. Academic Press Ltd.,London,1986.123-177.

[35] Zwicker E. Subdivision of the audible frequency range into critical bands (Frequenzgruppen)[J]. The Journal of the Acoustical Society of America, 1961, 33: 248.

[36] Moore B C J, Glasberg B R. Suggested formulae for calculating auditory‐filter bandwidths and excitation patterns[J]. The Journal of the Acoustical Society of America, 1983, 74: 750.

[37] Glasberg B R, Moore B C. Derivation of Auditory Filter Shapes from Notched-noise Data [J]. Hearing Research, 1990, 47 (1 - 2): 103 - 108.

[38] Aertsen A, Johannesma P I M, Hermes D J. Spectro-temporal receptive fields of auditory neurons in the grassfrog[J]. Biological Cybernetics, 1980, 38(4): 235-248.

[39] Qi Jun, Wang Dong, Jiang Yi, et al. Auditory Features based on Gammatone Filters for Robust Speech Recognition [C].ISCA, 2012.

[40] Mohri M, Pereira F, Riley M. Weighted finite-state transducers in speech recognition[J]. Computer Speech & Language, 2002, 16(1): 69-88.

[41] Ortmanns S, Ney H, Eiden A. Language-model look-ahead for large vocabulary speech recognition[C]//Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on. IEEE, 1996, 4: 2095-2098.

[42] Mohri M. On some applications of finite-state automata theory to natural language processing[J]. Natural Language Engineering, 1996, 2(1): 61-80.

[43] Aho A V, Hopcroft J E. Design & Analysis of Computer Algorithms[M]. Pearson Education India, 1974.

[44] Revuz D. Minimisation of acyclic deterministic automata in linear time[J]. Theoretical Computer Science, 1992, 92(1): 181-189.

[45] Mohri M. Finite-state transducers in language and speech processing[J]. Computational linguistics, 1997, 23(2): 269-311.

[46] Varga A P, Steeneken H J M, Tomlinson M, Jones D. The NOISEX-92 study on the effect of additive noise on automatic speech recognition [R]. Technical Report, Speech Research Unit, Defense Research Agency, Malvern, UK, 1992.

[47] SoX sound exchange[2013.2.1]. http://sox.sourceforge.net/Main/HomePage.

[48] Stolcke A. SRILM-an extensible language modeling toolkit[C].Proceedings of the international conference on spoken language processing. 2002, 2: 901-904.

[49] Allauzen C, Riley M, Schalkwyk J, et al. OpenFst: A general and efficient weighted finite-state transducer library[M]//Implementation and Application of Automata. Springer Berlin Heidelberg, 2007: 11-23.

[50] Povey D, Ghoshal A, Boulianne G, et al. The Kaldi speech recognition toolkit[C].Proc. ASRU. 2011.

[51] http://homepages.inf.ed.ac.uk/v1dwang2/public/tools/index.html

[52] GB/T 21023-2007.中文语音识别系统通用技术规范[S].