密级：内部★　　2　年

嵌入式语音识别系统中噪声与无关语音的消除和实现

Noise and Interruptive Speech Rejection for the Embedded Speech Recognition System

（申请清华大学工学硕士学位论文）

培 养 单 位 ： 电子工程系

学　　科 ： 信息与通信工程

研　究　生 ： 初 伟

指 导 教 师 ： 刘 加 教 授

二〇〇七年六月

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容。

本人保证遵守上述规定。

（保密的论文在解密后应遵守此规定）

作者签名：_____　　导师签名：_____

日　　期：_____　　日　　期：_____

摘 要

语音是人与人之间最为自然的交流方式，也是最有效人机交互方法之一。如何去除语音识别系统中或语音信息检索系统出现的干扰因素，使得语音技术可以更广泛、更有效地应用到现实生活中，是当前语音研究人员的工作重点。

我们研究工作主要分为两部分。一部分为孤立词语音识别系统中的干扰噪声拒识算法以及算法的片上实现，另一部分为连续音频流中的非语音音频移除。

针对孤立词语音识别系统中的干扰噪声拒识这一目标，我们采用了隐含马尔科夫模型对噪声进行声学建模，并且采用了与正常语音类似的线性识别网络对噪声解码，最后对网络输出的识别结果进行了置信度分析来确认系统输入是否为干扰噪声。

我们的干扰噪声拒识模块在正常语音与干扰噪声组成的测试集上获得了低于2.5%的系统等错点。最后，我们将干扰噪声拒识模块整合到了片上孤立词识别系统中，使得原有的只能接受集内词条作为输入的孤立词语音识别系统也能快速而准确地拒识干扰噪声输入，增强了原有孤立词识别系统的实用性。

针对连续音频流中的非语音音频移除这一目标，我们设计并实现了我们自己的多步语音/非语音分类器。主要设计思想是先将通过BIC分段算法得到若干相同性质的音频段，然后使用GMM分类器判断BIC分段算法得到的同质音频段的类型，最后采用后处理算法调整音频类型转换点的位置，以降低分类结果的帧错误率。

我们的多步语音/非语音分类器在共计7小时的中央广播电视台新闻联播的数据库上，获得了低于1.2%的总体帧错误率。最后，我们将我们的多步语音/非语音分类器添加到关键词检索系统的前端，使得原有的只能在纯净语音流下工作的关键词检索系统也能接受广播音频流作为输入，拓展了系统的应用范围。

关键词：语音识别　干扰噪声　拒识　置信度　硬件实现　音频分类　BIC　GMM

Abstract

Speech is the most natural communication way between human beings. Speech is also one of the most efficient human-machine interaction means.

Our research work can be mainly divided into two parts. One part is the rejection of interruptive noise appeared in the isolated speech recognition system and the on-chip implementation of the algorithm. The other part is the non-speech audio data removal for the continuous audio stream.

To achieve the goal of the rejection of interruptive noise appeared in the isolated speech recognition system, we utilize Hidden Markov Models in the acoustic modeling of the noise, decode the interruptive noise using a linear recognition net which is similar to that of speech, perform confidence measure analysis on the recognition results generated from the net to judge whether the system input is a interruptive noise finally.

Our interruptive noise rejection module obtained a system equal error rate lower than 2.5% on the test set which is composed of normal speech and interruptive noise. At last, we integrate the interruptive noise rejection module into the on-chip isolate speech recognition system. The isolate speech recognition system which could only take in-vocabulary words as input previously can now reject the interruptive noise input fast and accurately. Therefore, the practicability of the original isolate speech recognition system is enhanced.

To accomplish the goal of removal of the non-speech audio data from the continuous audio stream, we design and implement our own multi-pass speecn/non-speech discriminator. The major idea is first to segment the audio stream into several homogeneous audio segments through BIC segmentation algorithm, and then to employ a GMM classifier to judge the types of the segments obtained from the BIC segmentation algorithm, and finally to adopt a post-processing procedure to adjust the positions of the audio type change points so as to decrease the frame error rate of the discrimination results.

Our multi-pass speech/non-speech discriminator achieved an overall frame error rate which is smaller than 1.2%. Finally, we contribute our multi-pass speech/non-speech discriminator to the keywords spotting system as a front-end. The previous keywords spotting system which could only act under pure speech stream can now take broadcast news as input. Therefore, the application of our current advanced keyword spotting system is extended.

Keywords: Speech Recognition Interruptive Noise Reject Recognition Confidence Measure Hardware Implementation Audio Classfication BIC GMM

目 录

第 1 章 引言

1.1 语音识别前后端处理技术的意义与发展

1.1.1 语音识别的意义与发展现状

语音是人与人之间最自然、最基本、最有效的交流通信方式。在文字出现之前，语言和声音就作为人类日常生活的主要交流手段而存在了。可以说，人类社会今天的繁荣昌盛离不开语音通信。

1945年，随着世界上第一台电子计算机ENIAC 在美国的宾夕法尼亚大学摩尔学院电子工程系的诞生，人类社会的发展步入了一个崭新的时代。如何更好的促进人类与计算机的交流，使计算机更好地为人类社会服务，已经成为科研工作者面临的新挑战。人机语音通信，作为人机交互的有机组成部分，也因此得到了广泛关注。

人机语音通信包括自动化语音识别（Automatic Speech Recognition）和语音合成（Text-To-Speech Syntheis）两个主要方面。自动化语音识别的目标是让机器“听懂”人的语音，而语音合成的目标是让机器“说好”人的语音。

如何使得计算机识别以及理解人类日常生活中的语音是一项艰巨并富有挑战性的任务。因为人类不但能够听到传输到耳朵的声音，而且能够理解语音的内容。这也是为什么人类为什么在嘈杂环境下仍然能听懂对方说的语音内容。但是人类能够理解语音内容是基于对世界的广泛认识上的，这就是计算机语音识别任务如此困难的原因之一。

1971年，美国国防部先进研究项目局（Defense Advanced Research Projects Agency，DARPA）投资1500万美元启动了一项为期5年的语音识别项目，从此拉开了语音识别研究的序幕。美国电话及电报公司（AT&T）、BBN技术公司、卡耐基梅隆大学（CMU）、国际商用机器公司（IBM）、林肯实验室（Lincoln Labs）、麻省理工学院（MIT）以及斯坦福研究院（Stanford Research Institute，SRI）的许多研究人员参与到了这个研究项目中，并做出了重要的贡献。其中卡耐基梅隆大学开发的两套语音识别系统Harpy 和Hearsay-II，达到了预期的效果。

1988年，来自卡耐基梅隆大学李开复坚持采用隐含马尔可夫模型（Hidden Markov Model）的框架，成功地开发出世界上第一个大词汇量非特定人连续语音识别系统SPHINX [1]。这标志着语音识别研究的主要途径由基于经验的方法转到了基于统计的方式。

在20世纪90年代，是基于统计方法的语音识别技术的推广期。世界各国的研究人员搭建并发布了许多语音识别系统。1992年，剑桥大学（Cambridge University）的熵实验室（Entropy Research Laboratory）发布了隐含马尔可夫模型工具包（Hidden Markov Model Toolkit，HTK）1.3 [2]；并在1995年和2000年发布了HTK 2 和HTK 3 [3]。卡耐基梅隆大学分别在1992年，1996年，2004年分别完成了Sphinx-2 [4]，Sphinx-3 [5], Sphinx-4 [6]。

时至今日，语音识别仍然还有一段很长的路要走。从表 1.1中我们可以看出对于简单的数字串识别，机器还要经过40年才能达到人的水平。由于数字串中不包含高层次的内容信息，所以我们仍然要改进和完善最基本的机器语音识别技术。特别地，我们从表中可以看出，比起数字串识别这类简单识别任务，机器语音识别或许会在复杂任务中的识别性能更早地达到人类的识别能力，例如说话人相关的听写任务[7]。口语（sponetaneous）语音识别以及噪声鲁棒语音识别将是未来的两个最主要的研究难题。

表 1.1 当前机器语音识别能力与人的语音识别能力的比较

识别任务	机器识别的错误率 (%)	人类识别的错误率 (%)	预计机器追赶人类需要的时间 (年)
日常口语识别	30	4	19
数字串识别	0.7	0.009	41
字母识别	5	1	15
书面语音识别	3	0.09	11

1.1.2 语音识别前后端处理技术的意义与应用现状

由于语音识别系统的输入并不可能总是理想的具有高信噪比的语音数据流，所以在将音频数据输入系统之前，对其作一定的预处理去除干扰因素，使得处理后的音频流更加符合系统的要求，是很有必要，也是很有实际意义的。例如，研究工作者通常对信噪比（Speech Noise Ratio, SNR）比较低的输入语音应用语音

增强（Speech Enhancement）算法[8]，使之在特征提取（Feature Extraction）之前就有比较好的信噪比。常用的语音增强算法主要有基于谱分析的谱减法（Spectral Substraction）[9]、最小均方误差估计法（Minimum Mean Square Error ）[10]等。基于谱分析的语音增强技术对于加性噪声（Additive Noise）有比较好的移除效果，但是对于卷积噪声（Convolutional Noise）的去除效果不佳。近年来，许多研究人员将目光投向了盲信号分离（Blind Source Separation）技术[11] [12]，即使用盲解卷积（Blind Deconvolution）的方法来实现语音信号与干扰信号的剥离，取得了一定的效果。

对于实际的在线语音识别系统，由于系统需要时刻检测用户是否处于说话状态。所以就必须对输入的音频数据流作语音端点检测（Voice Active Detection，VAD），将音频流中的语音段提取出来作为语音识别系统的输入。如何找到一种语音端点检测算法，在低信噪比的情况下，快速而准确地找到语音非语音的转换点，是当前研究的重点。

同样在语音识别的前端对语音提取特征之后，开始维特比译码（Viterbi Decoding）之前，可以对特征作预处理，例如倒谱均值减（Cepstrum Mean Subtraction，CMS）[13]处理可以从特征谱中移除噪声对语音的干扰；声道长度归一化（Vocal Tract Length Normalization，VTLN）[14]可以从特征谱中移除不同说话人的声道特性对后续识别的影响。实验结果表明，特征预处理技术对于识别率的提高也是有很大帮助的。

为了最小化系统在训练与识别阶段的失配，可以利用第一阶段的解码结果对系统的模型或特征作无监督自适应（Unsupervised Adaptation）。科研工作者在无监督自适应领域作了大量的工作，其中最大后验概率自适应（Maximum a Posteriori，MAP）[15]和最大似然度线性回归自适应（Maximum Likelihood Linear Regression，MLLR）[16]都可以利用系统输出的对前期训练得到的模型中的参数进行无监督的调整。特别的，受束缚的最大似然度线性回归自适应（Constrained MLLR）[17] 还可以对语音特征作相应变换达到减小训练识别失配的目的。当前自适应技术可以利用特定的说话人的信息，将说话人无关（Speaker Independent，SI）的识别任务通过调整语音识别系统中模型参数和特征矢量转化为说话人相关（Speaker Dependent，SD）的识别任务，这样就大大减小了由于说话人变化对语音识别系统的影响，从而提高了语音识别系统的性能。

在语音识别系统得到识别结果之后，这些后验概率（posterior probability）和对数似然度（log likelihood）实际上反映了识别结果的可信程度，即置信度（Confidence Measure）。近年来，有很多科研工作者利用置信度分数来引导自适应训练[18]、集外词（Out of Vocabulary，OOV）拒识[19] 以及识别结果可信度分析[20]，取得了比较好的效果。

1.2 论文的选题依据与意义

1.2.1 孤立词识别系统中的干扰噪声拒识

进入21世纪，语音识别的研究重点是建立实用的、稳健的语音识别系统。其中一个关键点就是如何应对用户的集外词输入问题，即在识别过程中，用户的话音里出现了训练字典里没有出现过的词汇。随着计算机计算能力的提高，语音识别系统训练数据量增加，产生的字典也将随之增大，这在一定程度上缓解了集外词输入问题。但是由于任何语言的词汇都是持续增长的，所以无论计算机训练数据多么庞大，得到的字典也无法完全覆盖一种语言的所有词汇，也就是说。因此集外词问题将一直存在。在海泽灵顿的报告中指出，即使是采用了华尔街日报（Wall Street Journal）这样大的数据库训练得到超过100,000词的字典进行识别，集外词的比率仍然超过1%，而含有集外词的句子在所有测试句子中的比率将超过17% [21]。

除了集外词，无关语音输入也会影响语音识别系统正常工作。无关语音输入指的是用户在使用语音识别系统时，用户不经意间产生的干扰突发噪声。例如：清嗓子、嘴吹气、鼻子哼气、鼻腔长音、咂嘴、叹气、咳嗽、拍手、拍桌子、关节敲桌子等声音。如果语音识别系统在用户输入了干扰噪声的情况下仍然输出识别结果，那么无论识别结果是什么都会是不合理的。合理的情况应该是语音识别系统在用户输入了干扰噪声时能够准确快速地发现干扰输入，拒识该语音输入，并及时对终端用户作出提示或引导用户的下一次输入。而且语音识别系统还应当保证不会将用户说出的正常语音词条或句子误判为无关的干扰噪声。传统的语音识别系统大多针对用户配合的朗读语音进行处理。而当前语音识别系统的发展方向是在保证低的字错误率（Word Error Rate）的前提下，尽可能提高系统的人机交互质量，提高系统的人性化程度，最大化用户的舒适度，真正体现

科技以人为本的思想。

在用户配合的情况下，当前的中小词汇量的孤立词语音识别系统可以取得令人满意的识别结果。但是实际的语音识别系统每时每刻都在面临着用户输入的无关语音的考验。所以，在尽可能地保留集内输入语音的前提条件下，有效地据识无关语音，有着很强的实际意义。

随着海量存储技术的发展，人们得以随意的在存储媒质上储存各式各样的音频内容。例如日常对话、演讲报告、电视和广播节目中的音频数据。虽然语音是人与人之间最自然、有效的交流手段，但是如何快速并准确地在以指数速度爆炸性增长的海量音频数据中找到用户感兴趣的内容仍然不是一件容易的事情。因此，语音文件检索（Spoken Document Retrieval）作为自动化语音识别的一个重要应用，逐渐引起了科研工作者的兴趣。美国国家标准局（National Institute of Standards and Technology，NIST）早在1997年就开始将语音文件检索评测加入到文本检索会议（Text REtrieval Conference，TREC）评测中[22]。

虽然在提供高信噪比的朗读语音（例如播音员的新闻播报）作为系统输入的前提下，当前的语音信息检索和识别系统都能够获得比较高的识别正确率，但是大多数系统输入音频内容更多的是混合音频数据（Hybrid Audio Data）。这些混合音频数据不但包括了纯净的单人朗读语音，还包括了噪声环境下的多人对话，音乐场景、干扰杂音等音频内容[23]。因此，原有的针对纯净朗读语音的相关算法已经不可能满足工作在混合音频数据上的关键词检索系统或连续语音识别系统的要求。21世纪的语音信息检索和识别系统迫切需要高性能的混合音频分类技术。这种混合音频分类技术可以在去掉音频数据中的冗余信息的同时，保留对语音数据检索有用的信息。

事实上，广播音频数据中语音、音乐、背景噪声或其他干扰音频无论在时域或是频域都有着极强的混淆性。如何找到一个合理的框架来对这些易混数据作高效快速分离，这涉及到了模式分类领域中的很多理论。而且如何找到一系列能够高的区分度的特征，这也会涉及到信号处理领域的许多方法。所以说，本论文的非语音移除处理前端不但有很强的实际应用价值，也有很强的理论意义。

1.3 论文的目标与研究方向

1.3.1 硬件平台上的针对孤立词识别系统的干扰噪声拒识

本论文的一部分工作目的就是解决实际的语音识别系统面对含有突发噪声的情况。本论文要完成的硬件平台上的针对孤立词识别系统的干扰噪声拒识模块应当具有如下功能：

- 正确辨识出突发噪声和语音，对其中包含的集内发音做出正确的响应；对于无关语音发音，系统要进行有效的拒识。

- 最终的干扰噪声拒识算法在保证较高的无关语音拒识率的前提下，占用尽可能低的内存和消耗尽可能低的中央处理器的运算资源，以满足实时率的要求。

- 将最终无关语音拒识算法通过定点化整合到片上孤立词语音识别系统中，成为稳健孤立词识别系统的一个有机组成部分。

1.3.2 连续广播音频流中的非语音移除

本论文的一部分工作的目的就是构建一个稳健的高性能语音信息检索系统的前端。这个高性能语音信息检索系统的前端处理模块应当具有如下功能：

- 将音乐、背景噪声这样的无关干扰信息从海量的广播音频数据中快速而准确地移除。

- 在移除无关音频信息的同时，应尽可能地可靠地保留有用的语音信息段。

- 将最终的非语音干扰信息移除模块添加到关键词检索平台上，成为稳健关键词检索系统的一个高性能的前端模块。

1.4 论文各部分的主要内容

- 本论文在第二章介绍了我们的孤立词识别基线系统使用的语音识别引擎的基本识别算法，描述了特征提取的流程，介绍了声学模型基本单元的训练和识别网络的构造。
- 我们在第三章介绍并分析了干扰噪声的类型与特点，然后提出了对干扰噪声建模的方法，并提出了我们自己的干扰噪声拒识的系统框架，并对系统中的各个部分作详细说明，最后给出系统参数、训练测试数据库以及实验性能及分析。
- 我们在第四章介绍干扰噪声拒识模块的片上实现。
- 我们在第五章介绍并分析常用的语音/非语音分类方法的优点和不足，然后提出我们自己的语音/非语音分类器的系统框架，并对系统中的各个部分作详细说明，最后给出系统参数、训练测试数据库以及实验性能及分析
- 我们在第六章总结本论文的工作，阐述本论文的创新点，并对未来工作进行展望。

第 2 章 孤立词语音识别系统

本章主要介绍孤立词识别识别系统中特征提取、声学模型结构，声学模型训练、识别网络结构以及解码算法，并在本章给出现有的两阶段孤立词语音识别系统的性能。

2.1 特征提取

在自然界中，语音信号是以模拟的形式存在的，要用数字信号处理器对语音信号进行处理，首先需要通过A/D变换的方法，将其转换成为数字信号。但是，数字语音信号中包含了大量的冗余信息，因而还需要通过特征提取的方法提取其中的关键特征信息作为语音识别特征矢量。

目前基于统计模型的语音识别系统中常用的特征主要有三种：线性预测倒谱系数（Linear Predictive Cepstral Coefficient，LPCC）[24] [25]、Mel频标倒谱系数（Mel Frequency Cepstral Coefficient，MFCC）[26]和感知线性预测系数（Perceptual Linear Predictive Coefficient，PLPC）[27]。

MFCC特征是一种模拟了人耳的听觉特性的特征矢量，是Davis和Mermelstein在八十年代初期首先应用到语音识别领域当中的。MFCC参数识别性能、抗噪性能、稳健性能都远远高于之前出现的各种特征参数。近二十年以来，它已经成为基于隐含马尔科夫模型（Hidden Markov Model，HMM）语音识别系统中最常用的特征矢量。另外，MFCC特征也非常适用于以数字信号处理器为核心运算器件的嵌入式语音识别系统中。而且其数据精度完全能够达到大多数语音识别系统的任务要求。因此，我们最终选用了MFCC特征作为系统的特征。

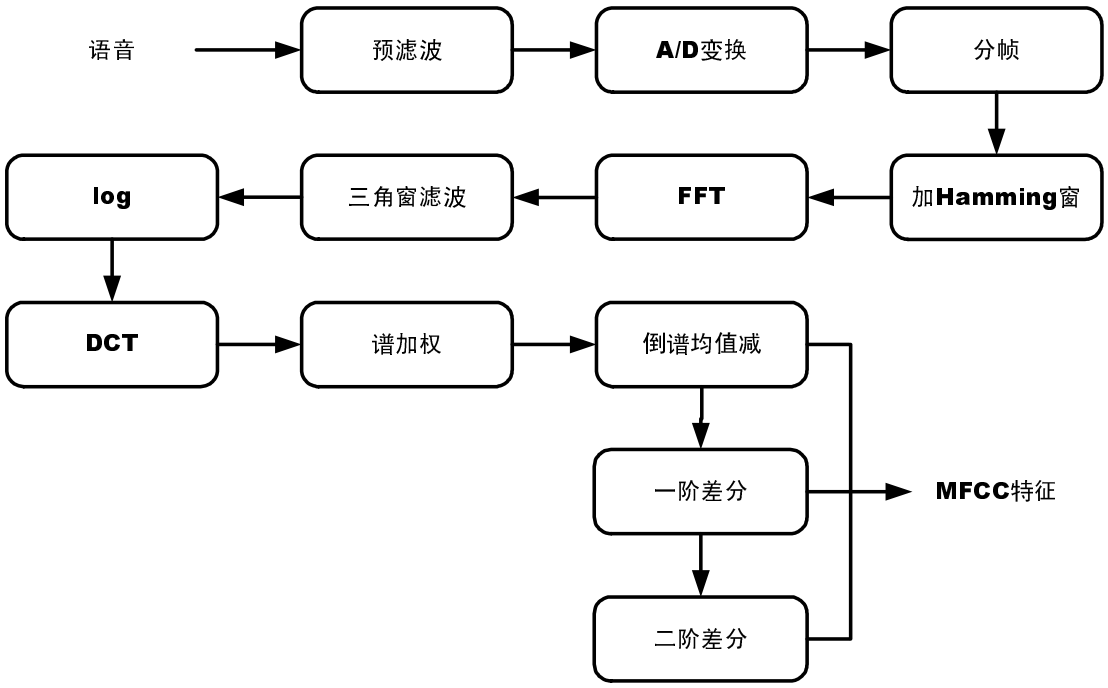


图 2.1　MFCC特征提取流程

图 2.1为本系统中MFCC特征的提取流程。

在本章描述的系统中，选用了27维语音特征参数，其中包括12维MFCC、12维ΔMFCC（即MFCC的一阶差分）、短时归一化对数能量$\log E$及其一阶差分$\Delta\log E$ 和二阶差分$\Delta^2\log E$。

2.2　声学建模

为了确保系统的高识别性能，本系统采用了连续隐含马尔可夫模型（Continuous Hidden Markov Model，CHMM）[28]作为系统声学模型的基本框架。本节将就CHMM的基本概念、模型训练算法以及汉语普通话中声学模型基本单元的选择作简要介绍。

2.2.1　隐含马尔可夫模型（HMM）的基本概念

隐含马尔可夫模型是一种随机过程模型，该模型假设外界可观察到的观察

矢量序列$O=o_1,o_2,\cdots,o_T$是由模型隐含层中的一串状态序列产生的。在语音识别系统中，模型的观察矢量为语音特征矢量，而模型中的状态序列则对应了语音的实际内容，如图 2.2所示。

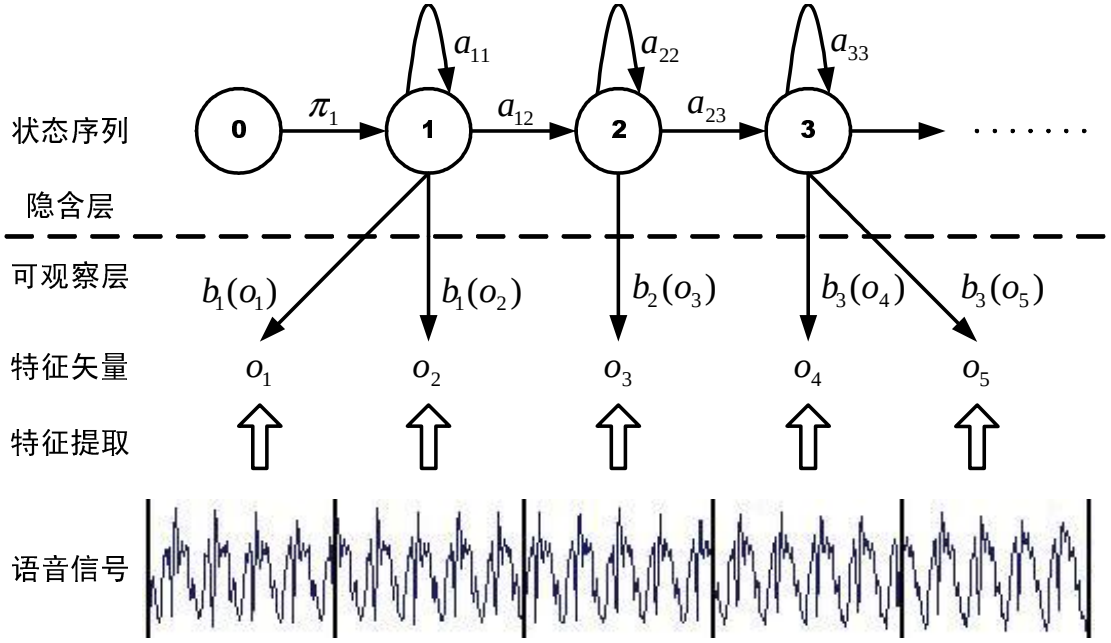


图 2.2　语音识别系统中HMM的基本框架示意图

一套HMM模型由三组参数唯一确定，状态初始概率分布矢量$\boldsymbol{\pi}=\{\pi_i\}_{1\le i\le N}$，转移概率矩阵$\mathbf{A}=\{a_{ij}\}_{1\le i,j\le N}$和输出概率密度分布函数矩阵$\mathbf{B}=\{b_j(o_t)\}_{1\le i\le N,1\le i\le T}$。其中$T$为观察矢量的总帧数，$N$为状态的总个数。

2.2.2　连续HMM框架下的训练算法

在HMM的框架下，模型的训练问题可以表达如下：给定一个HMM模型$\Phi=(\boldsymbol{\pi},\mathbf{A},\mathbf{B})$和一组观察矢量集合$O=o_1,o_2,\cdots,o_T$，如何调整模型参数，使得新的模型$\hat{\Phi}$在给定的观察矢量集合$O$下的似然度$p(O|\hat{\Phi})$最大。

这一问题是一个典型的不完全数据学习问题，由于模型中隐含层的存在，我们无法确知每一个观察矢量是由哪一个状态产生的。EM（Expectation-Maximization，期望最大化）迭代算法[29]非常适合解决此类问题。每一次迭代都分为两步：

1. **期望**（Expectation）：根据给定的初始模型Φ计算出观察矢量对应隐含层中所有可能的状态序列输出的后验概率；
2. **最大化**（Maximization）：根据第一步估计的结果，利用最大似然准则估计新的HMM模型参数$\hat{\Phi}$。

所得到的新的模型参数用来进行下一次迭代，以得到更优的HMM模型参数。

2.2.3　声学模型基本单元的选择

我们的孤立词识别系统中主要使用了上下文无关子词模型和上下文相关子词模型[30] [8] [31] [32] [28]。

根据基本子词建立的声学模型，又称作上下文无关模型。我们选取单音子（Monophone）模型作为我们的上下文无关模型；根据音素的上下文相关性建立的声学模型，又称作上下文无关模型。我们选取Biphone模型作为我们的上下文相关模型。

2.3　识别算法

2.3.1　Viterbi解码算法

HMM模型中的识别问题可以描述如下：给定HMM模型Φ和观测矢量序列$O=o_1,o_2,\cdots,o_T$，产生这个观察矢量序列的最佳状态序列$\mathbf{S}=s^1,\cdots,s^T$应该是所有状态序列中使得下式取得最大值的那一条状态序列：

$$\hat{\mathbf{S}}=\arg\max_{S} P(\mathbf{S},O|\Phi) \tag{2-1}$$

实际上，此问题等效于动态规划中的最优路径问题，可采用Viterbi算法来解决，从而找到隐含在HMM模型中的最佳状态序列。

2.3.2　识别网络

根据Viterbi算法的基本原理，在Viterbi解码的过程中，需要确认识别网络由哪些状态组成，而每一个状态有哪些可能的前续状态。我们的基本算法中，选择使用2个状态描述声母，4个状态描述韵母。此外还分别使用了1个状态描述语音

前后的静音模型（Silence）和字间的暂停（Pause）模型。在图2.3和图2.4中，我们以词条“中国”“北京”为例，说明了基于子词模型的线性状态网络的构建过程。

2.3.3　两阶段孤立词识别系统

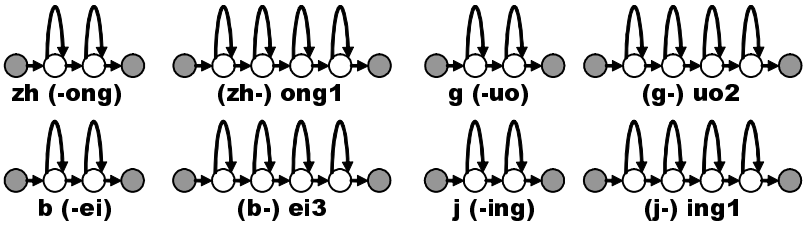
现有的基线系统全面地考虑了片上系统的特点，在运算复杂度和内存限制下尽量实现更好的性能，下面主要介绍两阶段识别的框架[33] [34]。

获得语音信号的特征矢量和前后端点信息之后，就可以进入识别阶段。读入所有的模型参数和整个识别网络，计算每帧特征对所有状态的输出概率，最终得到一个帧数乘以状态数的输出概率矩阵。

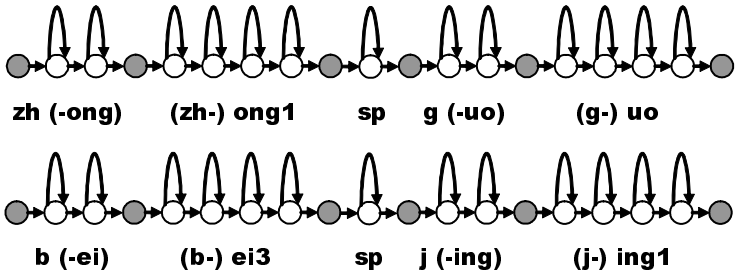
在识别中需要保证系统的识别性能，这就需要使用相对比较复杂的声学模型，如果使用我们的左相关的Biphone模型，就需要保留一个358乘以时间长度的内存空间，这个空间对片上系统来说是一个不小的负担，因此我们的孤立词识别系统采取了一种两级搜索的解码策略，在这种方案下嵌入式语音识别引擎的结构如图2.5所示。

中国 zhong1 guo2
北京 bei3 jing1
……

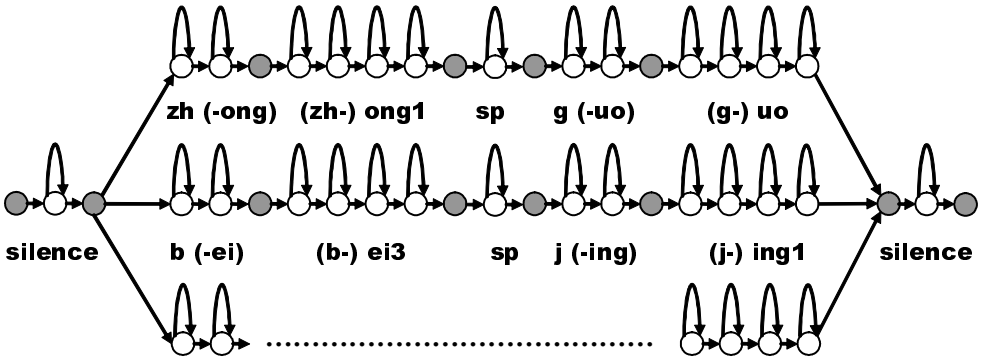
图 2.3　识别词表



a．各子词模型的状态连接关系



b．由子词模型拼接的各词条的线性网络



c．整个词表的识别网络

图 2.4　识别网络的生成过程

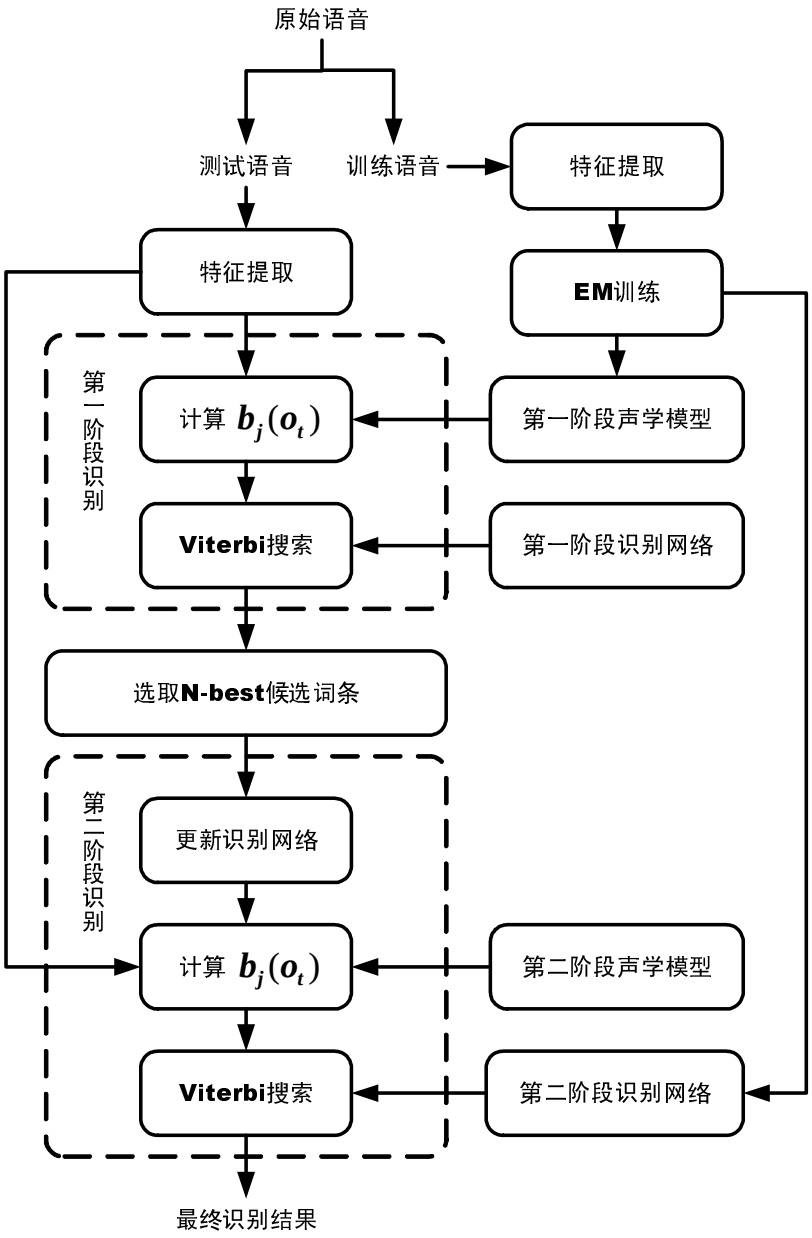


图 2.5　两阶段孤立词语音识别系统

2.3.4　基线系统的性能

2.3.4.1　两阶段孤立词识别系统使用的数据库

孤立词识别系统使用的数据库包括：

1. **声学模型的训练数据库**

   声学模型训练数据库包括：

   - 国家863计划智能计算机主题办公室提供的数据库

     连续语音，男、女性各83人，其中41人每人大约520句，42人每人大约650句，语料为人民日报内容，863专用话筒16KHz采样，录音环境为实验室安静环境，男性语音总长度为54小时45分钟，女性语音总长度为57小时5分钟。

   - 微软亚洲研究院提供和本实验室采集的数据库

     连续语音，男、女性各100人，其中每人大约200句，高性能话筒16KHz采样，录音环境为实验室安静环境，男性语音总长度为31小时3分钟，女性语音总长度为38小时54分钟。

2. **孤立词识别库**

   600孤立词，包括200地名、200人名、200股票名称，高性能麦克风16KHz采样，男性10人，女性5人，录音环境为安静实验室环境。

2.3.4.2　两阶段孤立词识别系统的参数

两阶段孤立词识别系统的使用的特征参数为2.1节中描述的27位特征矢量。系统采用2.2节中描述的EM算法训练连续HMM声学模型。第一第二阶段采用的声学模型分别为Monophone和左相关的Biphone，每个状态的高斯混合变量数分别为1和3，总的状态数分别为208和358。出于运算复杂度的考虑，两阶段的状态输出概率密度分布的协方差矩阵均被设定为对角阵。声学模型基本单元的结构和参数以及识别网络在本章第2节中都有详细描述。系统的识别词表规模为600词。

2.3.4.3　两阶段孤立词识别系统的性能

在我们600个孤立词识别库上对基线系统作了相应的测试，得到了基线系统的识别性能：如果采用男性声学模型测试男性语音、女性声学模型测试女性语音和男女性混合的声学模型测试整个测试集。对于一条测试词条，如果基线系统第一阶段产生的多候选识别结果中的任意一条识别结果与测试词条的实际内容一致，记作该词条识别正确；反之，记作该词条识别错误。第一阶段多候选识

别性能如表 2.1所示。

表 2.1　基线系统第一阶段识别率

候选数目	男性识别率 (%)	女性识别率 (%)	混合识别率 (%)
1	91.98	90.66	90.69
2	96.52	94.87	94.10
3	97.70	97.08	96.20
4	98.33	97.95	97.26
5	98.76	98.41	97.92
6	99.10	98.79	98.56
7	99.15	99.05	98.83
8	99.28	99.18	98.90

从表 2.1中可以发现，虽然模型的一选识别率较低，但是八选识别率基本可达到99.0%。而且，随着候选数的增加，第一阶段的识别率呈现了一个缓慢的上升趋势，因此在我们的基线系统中，选择第一阶段的候选数为8进入第二阶段的精细识别。

对于第二阶段识别过程的测试，我们分别考察了男性模型测试男性语音的情况、女性模型测试女性语音的情况、混合模型测试的情况，识别性能如表 2.2所示。

表 2.2　基线系统第二阶段识别率

男性识别率 (%)	女性识别率 (%)	混合识别率 (%)
97.90	97.84	97.26

2.4　本章小结

在本章中，首先介绍了论文研究的语音识别引擎的基本识别算法，描述了特征提取的流程，介绍了声学模型基本单元和识别网络的构造。

针对嵌入式语音识别引擎的特点，本章还提出了基线系统的两阶段段点判决方法和两级搜索的识别框架。

在两阶段识别框架下，第一级采用了相对简单的Monophone模型，各状态的输出概率密度分布为协方差矩阵为对角阵的单高斯分布，第二级采用了左相关

的Biphone模型，各状态的输出概率密度分布为协方差矩阵为对角阵、Mixture数为3的混合高斯分布。

最终，在识别词表为600词的情况下，一阶段识别率在8候选的情况下达到了98.9%；在第二阶段识别中，最终的识别率达到了97%以上。

第 3 章　孤立词语音识别系统中的干扰噪声拒识

语音识别系统工作的过程中由于使用者输入时的犹豫、发音习惯、环境噪声干扰等原因，会将一些噪声信号（例如清嗓子、咳嗽、叹气、咂嘴、开门声、拍手声、走动声等无关语音和干扰噪声）输入到原本只是为单纯语音输入而设计的识别系统。这不但会使系统输出本来应不该输出的错误识别结果，也为用户使用语音识别系统造成了极大的不便。

这样就需要设计一种分类方法将语音与干扰噪音分离开，识别系统将对分离出的语音送入识别网络，输出用户期待的正确识别结果；并拒绝识别分离出的干扰噪音，同时向用户输出有用的提示信息，等待下一次语音输入。

本章介绍并分析了干扰噪声的类型与特点，然后提出了对干扰噪声建模的方法，并提出了我们自己的干扰噪声拒识的系统框架，并对系统中的各个部分作详细说明，最后给出系统参数、训练测试数据库以及实验性能及分析。

3.1　干扰噪声的类型

我们在实验室环境下采集了一个噪声库，噪声库中收集了在语音识别系统中常见的10种噪声：清嗓子的声音、嘴吹气声、鼻子送气声、鼻腔长音、咂嘴声、叹气声、咳嗽、拍手声、拍桌子声音、关节敲桌子声音。采集噪声的能量大致和实际应用环境下出现信号的强度相当。

这些噪声如果按照发音的机理大致可以分为两类：一类和语音中的元音发音非常相似，是由于人的鼻腔、口腔的爆破声音，如清嗓子的声音、鼻腔长音、叹气声、咳嗽等；另一类就是和语音信号明显不同，包括平缓的气流、短促的敲击拍打声音，如嘴和鼻子平缓送气发出的声音、咂嘴声音、拍手、拍桌子等声音。

3.2　干扰噪声拒识的处理方法

据突发噪声不同的特点，我们从两个方面对噪声进行处理。

3.2.1　基于时频特征检测的方法

采用时频信号处理技术分析噪声和语音不同的特征，直接将噪声信号去除。由于第二类噪声的时频域特征明显不同于语音，通常不具有语音的韵律特征，比如基音周期和共振峰特征等。因此可以采用信号处理方法进行检测。

3.2.2　基于模式识别的方法

对于发音机理和语音信号相类似的噪声就不容易通过简单的信号处理的方法处理，需要根据噪声信号的特征设计匹配噪声的吸收模型，从模式识别的角度处理这类噪声，如果某帧信号的观察矢量更匹配噪声模型就认为这帧信号是噪声。

由于基于时频特征检测的方法对声带发声的干扰噪声的分辨能力较差，本论文决定采用基于模式识别的方法来对语音与干扰噪声分类。

3.3　干扰噪声建模

采用模式识别的方法对干扰噪声快速而准确地拒识，首要的问题就是如何合理地对干扰噪声建模。常用的基于模式识别的对噪声进行建模的方法主要有：

3.3.1　基本模式分类器

基于模式分类器的语音干扰噪声分类方法先提取语音和干扰噪声的特征矢量，然后将语音和干扰噪音的特征数据单纯地看作两类需要分类的数据。这样语音干扰噪音分类问题就转化成为数据分类的问题，而不考虑这些数据的来源。

常用的模式分类方法有很多：使用线性区分函数的线性分类方法[35]，多层神经网络分类方法[36]，以及近年引起科研工作者广泛关注的支持向量机（Support Vector Machine）分类方法[37]等。

由于上述模式分类的方法大多对一帧音频数据提取对应的特征，然后在帧的级别上区分语音和干扰噪声。这些数据类型无关的模式分类算法只考虑了语音和干扰噪声的短时信息，而没有考虑到实际的语音与干扰噪声长时信息，所以建立的模型不能够很好地反映语音与干扰噪声的实际特点。在实验结果上反映为语音与干扰噪声的帧特征混淆度高，区分度不够。

基于上述原因，本论文的干扰噪音模块没有采用数据类型无关的模式分类方法对语音和干扰噪声分类。

3.3.2　高斯混合模型（Gaussian Mixture Model，GMM）

基于GMM的建模方法先提取帧级别的语音和干扰噪声的特征矢量，然后采用基于最大似然度准则的训练方法得到干扰噪声的高斯混合模型[38]。当噪声根据类型不同分为N类时，可以使用这N组训练数据获得N个不同噪声的GMM模型。一般地，由于噪声的多变性，以及干扰噪声数据量相对语音数据量的不足，我们会使用训练集中的所有干扰噪声训练出一个干扰噪声GMM模型，即$N=1$。

GMM同样需要训练与干扰噪声GMM模型对应的语音GMM模型。这个语音的GMM模型是用训练集中的所有语音训练得到的。由于实际的应用环境中，语音词条和干扰噪声的长度普遍在2秒以下，所以我们可以采用基于整段判别的策略对语音和干扰噪声进行分类。执行步骤如下：

对于一段提取的音频特征$O_1^T=o_1,o_2,\cdots,o_T$，使用训练得到的干扰噪声的GMM模型计算一帧音频特征是由第j类干扰噪声产生的对数似然度：

$$\begin{aligned} ll(o_t|\Phi_j)=\log p_j(o_t|\Phi_j) &= \sum_{m=1}^{M} c_{jm}b_{jm}(o_t) \\ &= \log\sum_{m=1}^{M} c_{jm}N(o_t;\mu_{jm},\Sigma_{jm}) \end{aligned} \tag{3-1}$$

其中m为高斯混合分量的标号，M为高斯混合分量的数量。j代表噪声的种类。

由于语音和干扰噪声的短时平稳特性，可以认为特征矢量是帧间独立的。这样就可以计算一段输出特征矢量是由第j类干扰噪声产生的对数似然度：：

$$\begin{aligned} ll(O_1^T|\Phi_j) &= \log p_j(O_1^T|\Phi_j) \\ &= \log\left[p_j(o_T|o_1^{T-1},\Phi_j)p_j(o_{T-1}|o_1^{T-2},\Phi_j)\cdots p_j(o_1)\right] \\ &= \log\prod_{t=1}^{T} p_j(o_t|\Phi_j) \\ &= \sum_{t=1}^{T}\log p_j(o_t|\Phi_j) \end{aligned}$$

$$= \sum_{t=1}^{T} ll(o_t|\Phi_j) \tag{3-2}$$

用上述同样的办法计算出这段输出矢量是由语音产生的对数似然度$ll(O_1^T|\Phi_{speech})$，从而确定该段音频数据的类型：

$$\begin{cases} \text{如果} \quad ll(O_1^T|\Phi_{speech}) \geq \max\limits_j ll(O_1^T|\Phi_j) & \text{该段音频为正常语音} \\ \text{如果} \quad ll(O_1^T|\Phi_{speech}) < \max\limits_j ll(O_1^T|\Phi_j) & \text{该段音频为干扰噪声} \end{cases}$$

在使用GMM模型判断一段音频特征矢量的类型时，没有计算输出概率，而是计算了输出的对数似然度。主要因为计算一段特征对数似然度相对于输出概率有以下两个优点：

- 如公式 3-2所示，我们将计算输出概率的乘法运算转化为了计算对数似然度的加法运算，大大降低了运算量，并且有利于硬件实现
- 由于输出概率密度分布函数的动态范围很大，求对数运算后可压缩了输出概率密度分布的动态范围，有利于硬件平台的定点化

3.3.3　隐含马尔科夫模型（Hidden Markov Model，HMM）

虽然上面提到的基于GMM模型的语音和干扰噪声的建模方法可以取得相对于基于帧级别特征的模式分类方法更好的分类效果，但是这种方法也是有一定局限性的。

因为对于一段观测矢量$O_1^T = o_1, o_2, \cdots, o_T$而言，每一帧$o_t$的时域特征以及谱特征也都是不尽相同的。但是GMM分类器的训练阶段是使用了训练集中的相同类型的音频的所有特征帧来训练出一个GMM模型。大多数情况下，由于噪声类型的多样性和不可预测性，以及数据量的不足，我们甚至会选择使用训练集中的所有噪声训练出一个GMM模型。但是实际上这个泛化的GMM模型的概率密度分布函数是无法准确描述一种噪声的不同帧的特征在不同时刻的实际分布的。

我们认为虽然一段噪声内部所有帧的特征分布并不是完全一致的，但是如果将这一段噪声切分成若干小段，每个小段内部所有帧的特征分布还是基本一致的。并且在实际的孤立词识别系统中出现的干扰噪声（开关门声、敲桌子声、咳嗽声等）有很强的时序性。因此我们选择如图 3.1自左向右的3状态HMM模型

对一段干扰噪声进行建模。相比于GMM模型，这里的HMM模型使用了一个有限状态机来在刻画一段噪声内部的状态的跳转关系。相比于GMM模型（实际相当于单状态的HMM模型），HMM模型不但增加了状态的数目，而且增加了状态之间的转移概率，这样就可以克服GMM模型对噪声内部结构刻画不够细致的缺陷，从而提高系统的性能。

此外，在HMM的框架下，我们无需象在GMM的框架下那样使用所有语音帧训练一个GMM模型。我们可以直接使用第二章中提到的子词模型（一种HMM模型）来描述系统中出现的词条。显然，子词模型相比于GMM模型能够更准确地描述一个语音词条的内部结构。

综上所述，我们的干扰噪声拒识模块中最终采用了基于HMM的方法对干扰噪音进行建模。具体的识别网络结构将在下一节中介绍。

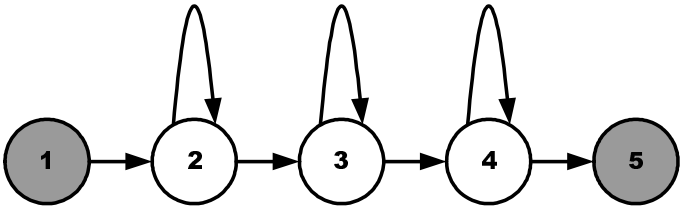


图 3.1　描述噪声的HMM模型结构示意图

3.4　包含干扰噪声拒识的识别网络

本论文中，我们提出了一种新的观点：即在孤立词识别系统中出现的干扰噪声可以看作一种特殊的“词条”。因此我们不但可以使用自左向右的HMM模型来对干扰噪声建模，而且可以采用第二章相类似的识别网络来对一段干扰噪声作Viterbi译码。干扰噪声Viterbi译码使用的识别网络如下图 3.2所示：

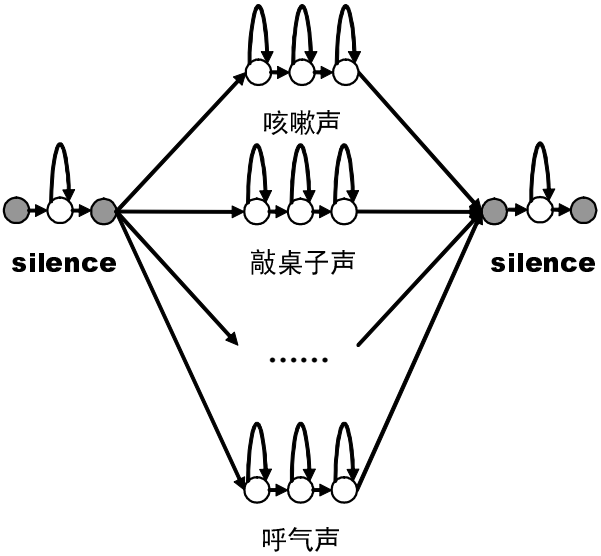


图 3.2　干扰噪声的识别网络

从图中可以看出，干扰噪声的识别网络采用了类似前面第二章提到的基于子词模型的线性状态网络。其中的每一个噪声“词条”模型的参数由训练集中的对应类型干扰噪声训练得到。在网络的前后两端的silence模型用于吸收干扰噪音前后的静音。

我们最终识别网络的结构采用了如图 3.3所示的一个包含了干扰噪声和正常语音两个识别网络的并行网络。一段音频的特征并行地通过两个识别网络，这两个识别网络可以分别计算由噪声训练得到的HMM模型以及语音训练得到的子词模型匹配这段观察矢量得到的对数似然度分数。

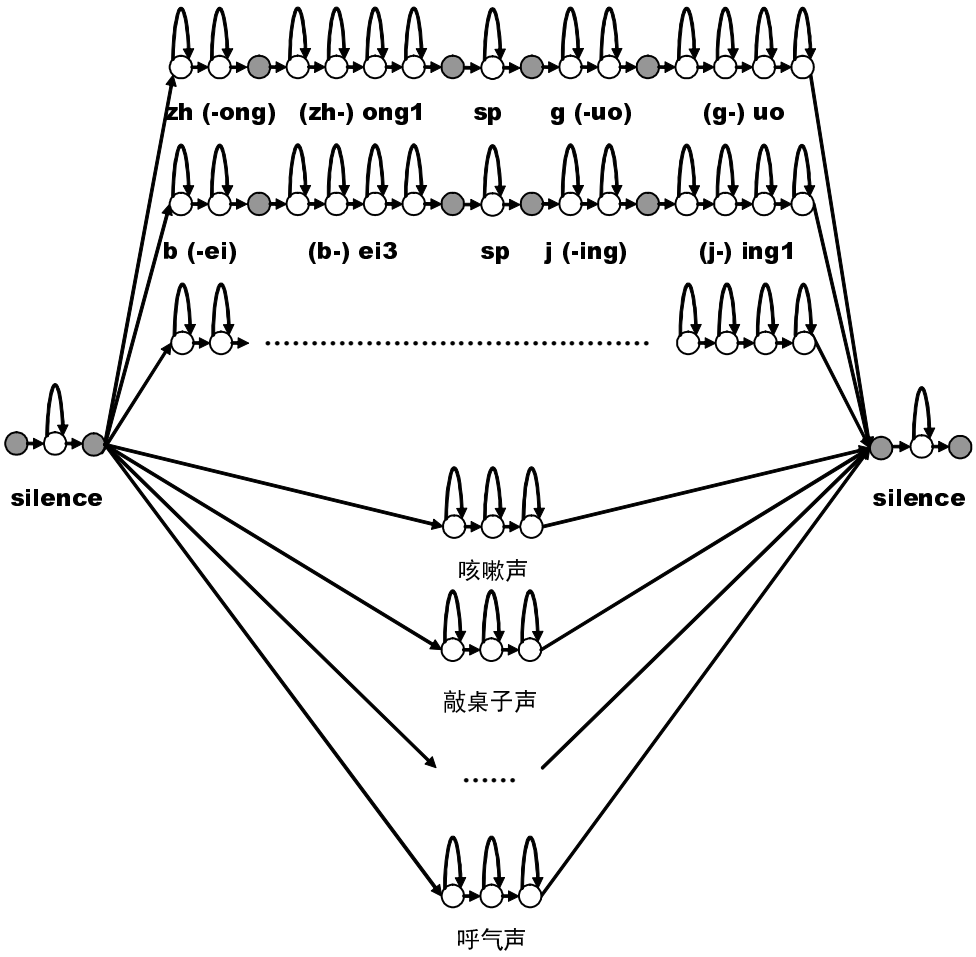


图 3.3　合并后的识别网络

从图 3.3中可以看出，silence模型在识别网络的开始和结束位置吸收静音。如果词条总数为M，干扰噪声种类为N，那么系统输入一段音频后，识别网络将会输出$M+N$个对数似然度分数。如何根据识别网络的输出对干扰噪声和正常语音分类，涉及到置信度分数处理的问题。本论文将在下一节详细讨论置信度分析与干扰噪音的拒识。

3.5　利用置信度分数拒识干扰噪声

3.5.1　置信度的定义

语音识别中的置信度（Confidence Measure，CM）代表识别结果的可信程度，反映了识别结果词条的声学模型和输入信号的观察矢量之间的匹配程度。对于有拒识功能的识别系统而言，如果某次识别结果具有足够的可信程度那么就可以接受这样的识别结果，否则系统就要对输入的语音信号拒识。

我们可以从统计假设检验的角度来解释语音识别中的置信测度。假设某次识别过程中的输入信号的观察矢量为O，则观察矢量O相对于词条W的置信度$CM(O|W)$表示语音O由词条W产生的可信程度。系统输出的识别结果词条W实际上是对于输入语音的一个假设H，这个假设H可能有“真”和“假”两种状态：

- **原假设**H_0：表示语音的观察矢量O由词条W产生，识别结果正确；
- **备选假设**H_1：表示语音的观察矢量O由词条W以外的其它模型产生，识别结果错误。

对此识别结果验证（Verification）的主要手段就是进行统计假设检验，即利用搜集到的数据对事先提出的假设按照某种方法进行检验，计算词条W 产生语音观察矢量O的置信测度$CM(O|W)$，从而决定是否接受假设H_0。

这样，检验统计量的取值空间被分为两部分，接受域（Acceptance Region）和拒绝域（Reject Region）。当检验统计量落在接受域中时接受假设假设H_0，落在在拒绝域中时时拒绝假设假设H_0。

根据假设本身的性质（真/假）以及检验的判断结果（接受假设/拒绝假设），假设检验中出现的情况包括以下四种情况：正确接受（H_0为真时被接受）、错误拒绝（H_0为真时被拒绝）；错误接受（H_0假时被接受）和正确拒绝（H_0为假时被拒绝）。因此假设检验通常会产生两类错误判断：

- **第Ⅰ类错误**（Type I error）**错误拒绝**
  发生的概率为$Pr(reject(H_0)|true(H_0))$，此类错误又称漏报（False Rejection，FR）。
- **第Ⅱ类错误**（Type II error）**错误接受**
  发生的概率为$Pr(accept(H_0)|false(H_0))$，此类错误又称虚警（False Alarm，FA）。

这两类错误发生的概率统称为条件错误率（Conditional Error Rate），两类错误之和称为总体证实错误（Total Verification Error，TVE）两类错误概率相等时的错误率称为相等错误率（Equal Error Rate，EER）。通常的语音识别错误率指的是在第 I 类错误为零（即全部接受）的情况下发生的第 II 类错误率。从这个意义上讲，可以说传统语音识别是引入了置信度的语音识别的一种特例。

对于总数为$N(H_0)$的识别结果的样本，可以分成如表 3.1所示的四类，其中的$N(Decision, HypoStatus)$表示在假设H_0的状态为$HypoStatus$（真/假）的情况下进行$Decision$（接受假设/拒绝假设）判断（接受或拒绝）时得到的样本数目。

表 3.1　假设检验的混淆矩阵

	接受H_0	拒绝H_0
H_0为真	$N(accept(H_0), true(H_0))$	$N(reject(H_0), true(H_0))$
H_0为假	$N(accept(H_0), false(H_0))$	$N(reject(H_0), false(H_0))$

第 I 类错误和第 II 类错误可以分别估计如下：

$$Pr(reject(H_0)|true(H_0)) = \frac{N(reject(H_0), true(H_0))}{N(true(H_0))} \tag{3-3}$$

$$Pr(accept(H_0)|false(H_0)) = \frac{N(accept(H_0), false(H_0))}{N(false(H_0))} \tag{3-4}$$

在确定统计假设检验的法则时，应该尽可能使两种错误发生的概率都较小。但一般来讲，当样本容量固定时，若减小一类错误的概率，则另一类错误出现的概率往往增大。因此，实际应用中往往需要找到两种错误率之间的一个平衡，两类错误之间的关系可以作为置信测度本身性能的一个评价方案。

3.5.2　拒识模型的评价方法

拒识模型的评价方法[39] [40]比较多，上面提到的第 I 类错误、第 II 类错误、总体证实错误、相等错误率等都是衡量置信测度的重要量度。由于置信测度是为了拒绝一部分识别结果来提高系统的识别性能，因此拒绝率（Reject Rate，RR）和拒绝后的准确率（Accuracy after Rejection，AR）也是置信测度的一个重要量度。拒绝率的概念包括对正确输入的拒绝率、对错误输入的拒绝率和总体拒绝率。对正确输入的拒绝率指拒绝正确输入的次数$N(reject(H_0), true(H_0))$占

总输入次数$N(H_0)$ 的比例；对错误输入的拒绝率指的是拒绝错误输入的次数$N(reject(H_0), false(H_0))$占总输入次数$N(H_0)$的比例。总体拒绝率$RR$是两者之和：

$$RR = \frac{N(reject(H_0), true(H_0)) + N(reject(H_0), false(H_0))}{N(H_0)} \tag{3-5}$$

拒绝后的准确率AR为：

$$AR = \frac{N(accept(H_0), true(H_0))}{N(accept(H_0), true(H_0)) + N(accept(H_0), false(H_0))} \tag{3-6}$$

拒绝率表示了系统对用户的友好程度，拒绝率越大识别率就越高，但是相应的，系统就拒绝了用户多次正确的输入，而且需要用户进行多次的重复，为使用带来很多不便。反之，拒绝率太小，就不足以拒绝掉集外词发音、噪声等无关信号，系统地识别率就受到影响，使用的稳健性也比较差。

从假设检验的角度可以通过更加直观的方法来评价置信测度的性能。主要使用两个评价方式：ROC曲线、DET曲线。

ROC曲线（Receiver Operating Characteristics）指以漏报率（Detection Rate，DR）即$Pr(accept(H_0), false(H_0)))$为纵轴，以虚警率（False Alarm，FA）即第II类错误为横轴的关系曲线。

DET曲线（Detection Error Tradeoff）是分别以两类错误为横轴和纵轴而作出的变化曲线，典型的曲线例如图 3.4所示，其中粗实线是随机接收时的DET曲线，表明在随机接受时，错误拒绝和错误接受的概率是完全相同的；虚线是理想的假设检验性能，即不管其中一种错误率为何值，系统都使得另一种错误率为零；两者之间的细实线实际情况下的假设检验曲线。显然DET曲线越接近零点表明置信测度的性能越好，这就可以确定置信测度另一个重要的指标等错误率（Equal Error Rate，EER），即第一类错误和第二类错误相等时的值，这个值越低，表明置信测度越稳健。

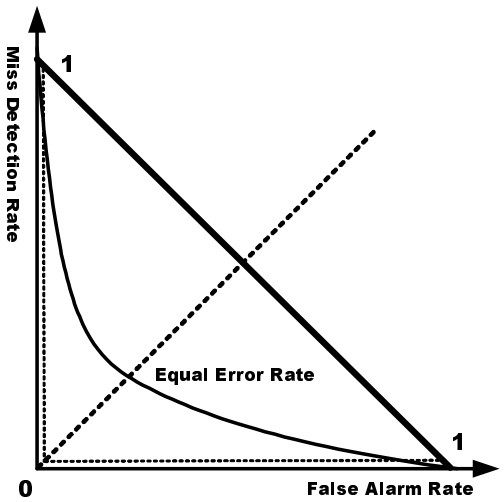


图 3.4　置信测度DET曲线示意图

在下面实验中，我们一般使用DET曲线和等错误率EER来评价置信测度的性能。

3.5.3　本文采用的置信度

本节将详细介绍无关噪声拒识模块中使用的置信测度。我们设计了两种置信测度用于分类在孤立词识别系统中出现的正常语音和干扰噪声：

3.5.3.1　语音网络输出的对数似然度$ll(O|W)$

在一段观测矢量通过语音的基于子词模型的线性状态网络后，Viterbi译码算法会生成与网络中M个词条相对应的M个对数似然度的匹配分数$ll(O|W_j)$，j为词条的序号。这个网络生成的M个对数似然度中最大的一个反映了观测矢量O由是正常语音的的可信程度，所以，本系统将识别结果的第一选的对数似然度分数作为一种置信度$CM_1(O|W)$：

$$CM_1(O) \quad = \quad \max_j ll(O|W_j) \qquad W_j \in \Omega_1 \tag{3-7}$$

Ω_1代表所有语音词条组成的集合。

3.5.3.2　噪声网络输出的对数似然度$ll(O|W)$

同理，在一段观测矢量通过噪声网络后，得到的对数似然置信度为：

$$CM_2(O) = \max_j ll(O|W_j) \qquad W_j \in \Omega_2 \tag{3-8}$$

Ω_2代表所有噪声词条组成的集合。

3.5.3.3　语音网络输出的后验概率$Pr(W|O)$

在识别系统中，输入特征是观测矢量O，对于基于HMM模型的语音识别器而言，其目的是找到能够令后验概率$Pr(W|O)$最大的词序列W^*，即

$$W^* = \arg\max_{W\in\Omega} Pr(W|O) \tag{3-9}$$

其中，Ω是所有可能词序列的集合。根据贝叶斯定理：

$$Pr(W|O) = \frac{Pr(O|W)Pr(W)}{Pr(O))} \tag{3-10}$$

在识别过程中，由于O给定，$Pr(O)$对所有的W都相同，实际计算过程中通常将$Pr(O)$省略去。这样，识别过程就相当于求解：

$$W^* = \arg\max_{W\in\Omega} Pr(O|W)Pr(W) \tag{3-11}$$

对于语音网络，我们认为每词条出现的概率是相同的：

$$Pr(W) = \frac{1}{M}, \qquad W \in \Omega_{speech} \tag{3-12}$$

对于噪声网络，我们认为每种噪声出现的概率也是相同的：

$$Pr(W) = \frac{1}{N}, \qquad W \in \Omega_{noise} \tag{3-13}$$

所以我们前面提出的对数似然度的置信度$ll(O|W)$实际上与$Pr(O|W)Pr(W)$是等价的。

对词表中不同词条的似然度加以比较，总可以得到满足要求的一个。但是由于丢失了$Pr(O)$信息，得出的最大值$\max Pr(O|W)Pr(W)$仅仅是一种相对测度。

在$Pr(O)$相等的情况下，各词条的$Pr(O|W)Pr(W)$是可比的；但在$Pr(O)$不等的情况下，它们就不可比了。

如果简单地用$p(O|W)Pr(W)$作为识别结果的可信度准则，就有可能出现$Pr(W|O)$很小，但仍被作为可信的识别结果输出的情况，从而造成误识。因此我们的系统中增加了绝对的置信测度，即基于$Pr(W|O)$的置信度。

$$\begin{aligned}
CM_3(O) &= \log Pr(W^*|O) \\
&= \log \frac{p(O|W^*)Pr(W^*)}{\sum_{W\in\Omega_1} p(O|W)Pr(W)} \\
&= \log \frac{p(O|W^*)}{\sum_{W\in\Omega_1} p(O|W)} \qquad (Pr(W) = Pr(W^*) = \frac{1}{M}) \\
&= \log p(O|W^*) - \log \sum_{W\in\Omega_1} p(O|W) \\
&= \max_j ll(O|W_j) - \log \sum_{k=1}^{M} e^{ll(O|W_j)} \\
&\approx \max_j ll(O|W_j) - \log \sum_{k=1}^{N_{best}} e^{ll(O|W_k)}
\end{aligned} \tag{3-14}$$

其中Ω_1代表所有语音词条组成的集合，集合中词条数目为M，N_{best}代表识别网络输出的对数似然度$ll(O|W_j)$分数经过排序后的前N_{best}个候选结果。实验结果表示，在N_{best}候选之外的识别结果的对数似然度分数之和在全部识别结果的对数似然度分数之和中只占很小的比例。所以我们采取排序后的前N_{best}个识别结果的似然度分数之和代替全部识别结果的对数似然度分数之和。这样就大大节省了运算量，在语音词条数目比较多（我们的系统$M = 600$）的情况下，可以适应实时性要求高的片上系统的计算要求。在系统中，N_{best}被设定为8。

3.5.3.4　噪音网络输出的后验概率$Pr(W|O)$

同理，在一段观测矢量通过噪声网络后，得到的基于$Pr(W|O)$的置信度为：

$$\begin{aligned}
CM_4(O) &= \log p(O|W^*) - \log \sum_{W\in\Omega_2} p(O|W) \\
&= \max_j ll(O|W_j) - \log \sum_{k=1}^{N} e^{ll(O|W_j)}
\end{aligned}$$

$$\approx \max_j ll(O|W_j) - \log \sum_{k=1}^{N_{best}} e^{ll(O|W_k)} \tag{3-15}$$

Ω_2代表所有噪声词条组成的集合，噪声种类为N类。

我们已经设计好了D种置信测度来评价识别结果的可信程度。一段观测矢量O分别通过语音网络和噪声网络之后，可以得到该段观测矢量D维的相对于语音和无关噪声的置信度分数。我们可以采用以下的置信度处理方法来来拒识无关噪声：

将正常语音和干扰噪声的两个D维的置信度分数矢量$CM(O) = [CM_1(O) \quad CM_2(O) \quad CM_3(O) \quad CM_4(O)]^T$ 投影为一个1维的置信度分数$\vec{CM}(O)$：

$$\vec{CM}(O) = w^T CM(O) \tag{3-16}$$

其中下标1和2分别对应正常语音和干扰噪声。线性投影向量w可以根据Fisher线性区分分类器的准则函数训练得到：

$$J_F(w) = \frac{(\vec{m}_1 - \vec{m}_2)^2}{\vec{S}_1^2 + \vec{S}_1^2} \tag{3-17}$$

其中$\vec{m}_i$为各类置信度分数向量投影后的均值：

$$\begin{aligned} \vec{m}_i &= \frac{1}{N_i} \sum_{\vec{CM}(O) \in \vec{\Omega}_i} \vec{CM}(O) \\ &= \frac{1}{N_i} \sum_{CM(O) \in \Omega_i} w^T CM(O) \\ &= w^T \left(\frac{1}{N_i} \sum_{CM(O) \in \Omega_i} CM(O) \right) \\ &= w^T m_i \end{aligned} \tag{3-18}$$

其中m_i为各类置信度分数的均值向量：

$$m_i = \frac{1}{N_i} \sum_{CM(O) \in \Omega_i} CM(O) \qquad i = 1,2 \tag{3-19}$$

这里的$\vec{S}_i$表示各类置信度分数向量投影后的类内离散度矩阵：

$$\begin{aligned}\vec{S}_i^2 &= \sum_{\vec{CM}(O)\in\vec{\Omega}_i}(\vec{CM}(O)-\vec{m}_i)^2 \\ &= \sum_{CM(O)\in\Omega_i}(w^T CM(O)-w^T m_i)^2 \\ &= w^T\left[\sum_{CM(O)\in\Omega_i}(CM(O)-m_i)(CM(O)-m_i)^T\right]w \\ &= w^T S_i w \end{aligned} \tag{3-20}$$

其中S_i为各类置信度分数的均值向量：

$$S_i^2 = \sum_{CM(O)\in\Omega_i}(CM(O)-m_i)(CM(O)-m_i)^T \qquad i=1,2 \tag{3-21}$$

我们再定义总类内离散度矩阵S_ω和置信度分数类间离散度矩阵S_b为：

$$\begin{aligned} S_\omega &= S_1+S_2 \\ S_b &= (m_1-m_2)(m_1-m_2)^T \end{aligned} \tag{3-22}$$

这样$J_F(w)$可以表示为：

$$\begin{aligned} J_F(w) &= \frac{(\vec{m}_1-\vec{m}_2)^2}{\vec{S}_1^2+\vec{S}_1^2} \\ &= \frac{(w^T m_1-w^T m_2)^2}{w^T S_1 w+w^T S_2 w} \\ &= \frac{w^T(m_1-m_2)^2 w}{w^T(S_1+S_2)w} \\ &= \frac{w^T S_b w}{w^T S_\omega w} \end{aligned} \tag{3-23}$$

我们希望找到这样的投影向量w，使得投影后的类内离散度尽量小，类间离散度尽量大。即求使得$J_F(w)$取得极大值时的w^*。可以使用Lagrange乘子法求得w^*（具体求解过程从略）：

$$w^* = S_w^{-1}(m_1-m_2) \tag{3-24}$$

利用训练集数据获得最佳投影向量，并将置信度分数矢量投影到一维空间

后，我们还需要确定一个阈值y_0，作为分类决策的依据。

$$\begin{cases} \text{如果} \quad \vec{CM}(O) \geq y_0 \quad \text{该段音频为正常语音} \\ \text{如果} \quad \vec{CM}(O) < y_0 \quad \text{该段音频为干扰噪声} \end{cases}$$

实际上这个阈值就是我们用于控制系统工作点的参数。增大y_0的值，将会有更多的干扰噪声被判定为正常语音，此时系统的漏报率降低，虚警率升高；减小y_0的值，将会有更多的正常语音被判定为干扰噪声，此时系统的虚警率降低，漏报率升高。我们可以根据实际系统运行中对漏报率与虚警率的要求，选取y_0的值，使系统工作在令人满意的工作点上。

3.6　包含干扰噪声拒识模块的孤立词识别系统框架

本论文的干扰噪声拒识模块添加到原有的两阶段孤立词语音识别平台上后，系统的结构如图 3.5所示。

考虑到有限的片上系统的计算能力，和高实时性的要求，我们将干扰噪声拒识模块添加到第一阶段识别中。如果根据第一阶段输出的置信度判别该观测矢量为无关噪声，则拒识该输入，并不再继续第二阶段识别；如果判别该观测矢量为正常语音，则接受该输入，并利用第一阶段输出的N-best候选识别结果在第二阶段进行精细识别并输出最终的语音识别结果。

从系统框图中也可以看出，由于在第一阶段识别中加入的噪声词条网络与原有的语音词条网络是并行的、互不干扰的。因此正常的语音输入如果在第一阶段被干扰噪声拒识模块接受，最终的识别结果将与原有的两阶段孤立词识别系统得到的识别结果完全相同。所以在下一节的系统性能与实验结果描述中，我们不再赘述新的具有干扰噪声拒识功能的孤立词识别系统的语音识别性能，而只介绍干扰噪声拒识的相关性能，例如虚警率与漏报率。

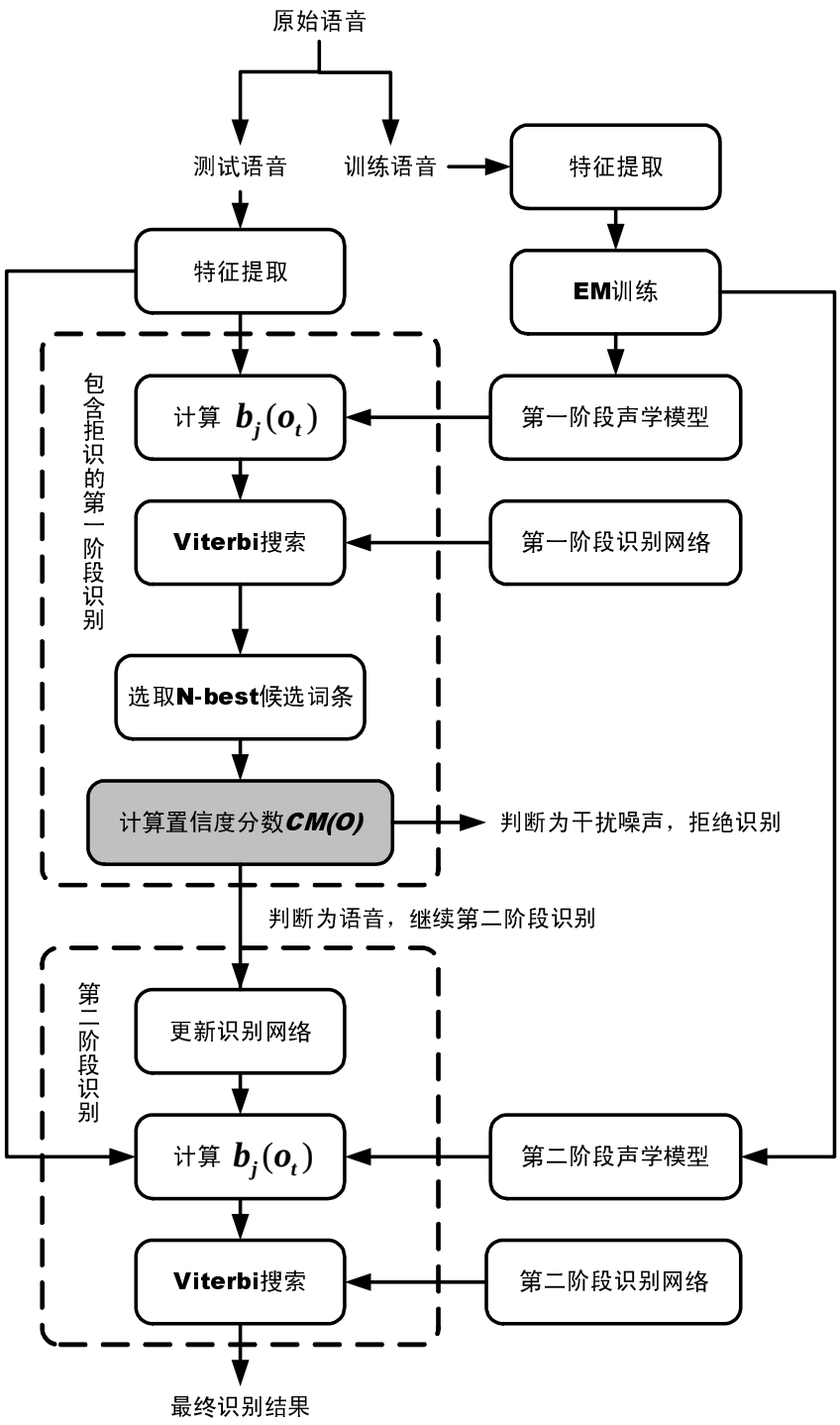


图 3.5　添加了干扰噪声拒识的孤立词语音识别系统

3.7 干扰噪声拒识系统的性能

3.7.1 干扰噪声拒识系统使用的数据库

包含干扰噪声拒识模块的孤立词识别系统用于训练和测试的语音库与第二章中描述的语音数据库完全相同。所不同的是，我们这里还增加了用于训练噪声声学模型和测试干扰噪声拒识性能的干扰噪声库：

10种常见突发噪声，包括清嗓子的声音、嘴吹气声、鼻子呼气声、鼻腔长音、咂嘴声、叹气声、咳嗽、拍手声、拍桌子声音、关节敲桌子声音，采样频率为8KHz，16bit量化，录音环境为安静实验室环境，男性60人，女性24人，每人每种噪声重复5遍。音频总长度为20分钟，共计2200个音频文件。

3.7.2 系统参数以及实验结果分析

包含干扰噪声拒识模块的孤立词识别系统的特征参数为2.1节中描述的27位特征矢量，与语音词条有关的声学模型的基本单元和参数以及两阶段的识别网络结构和参数与第二章中描述的完全一致。这里，我们主要介绍与干扰噪声相关的声学模型参数和识别网络参数。

噪声的声学模型为自左向右的3状态的HMM模型，每个状态的高斯混合变量数为3。噪声种类为10，因此总的状态数为30。

我们随机选取了整个噪声数据库的2/3共计1467个音频文件以及整个孤立词识别数据库的2/3共计1867个语音文件做为训练集；并使用整个噪声数据库余下的1/3共733个音频文件以及孤立词识别数据库的1/3共计933个语音文件作为测试集，

噪声的声学模型与都是根据训练集中的噪声数据训练得到的。置信度分数的投影向量是由训练集中的噪声数据和语音数据通过识别网络后得到的置信度分数训练得到的。

我们随机地在训练集中全部通过噪声和语音网络生成的置信度分数中随机选取了1/10绘制了置信度分数矢量$CM(O)$的实际分布。如图 3.6 和图 3.7 所示。

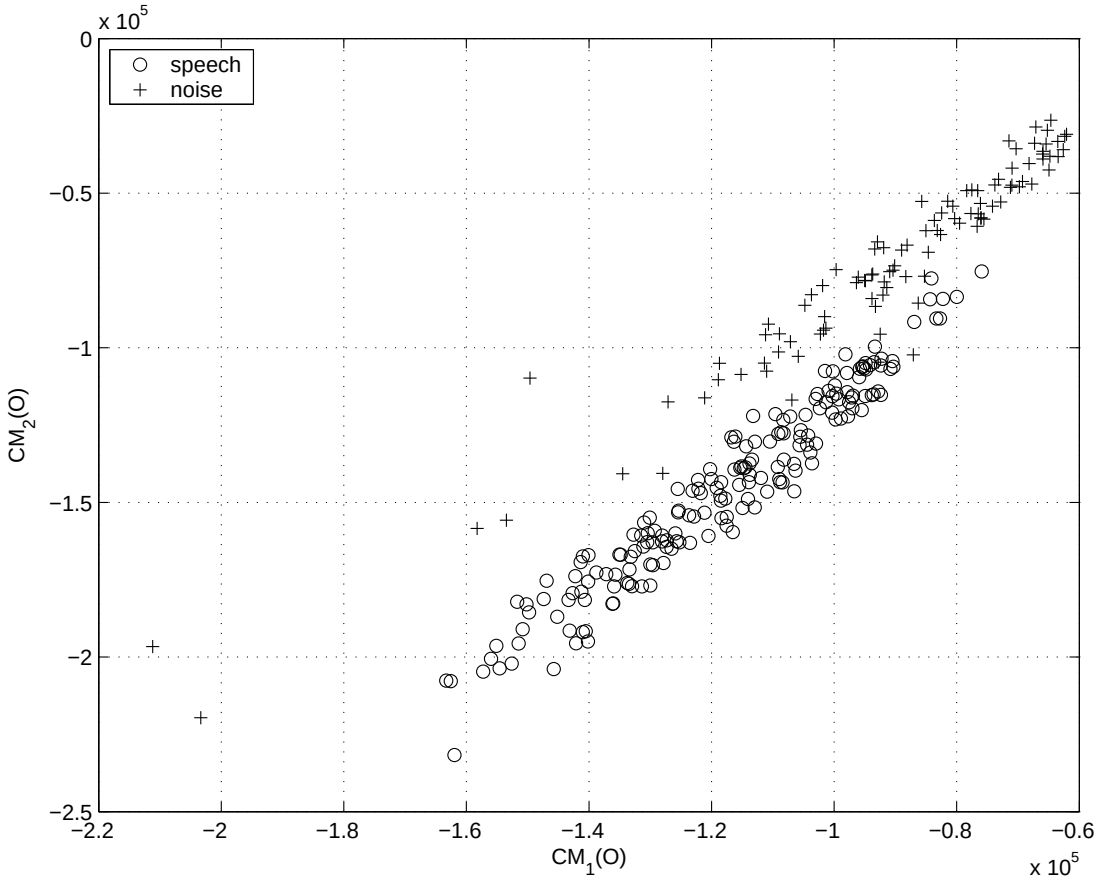


图 3.6　干扰噪声与正常语音基于对数似然度的置信度分数分布

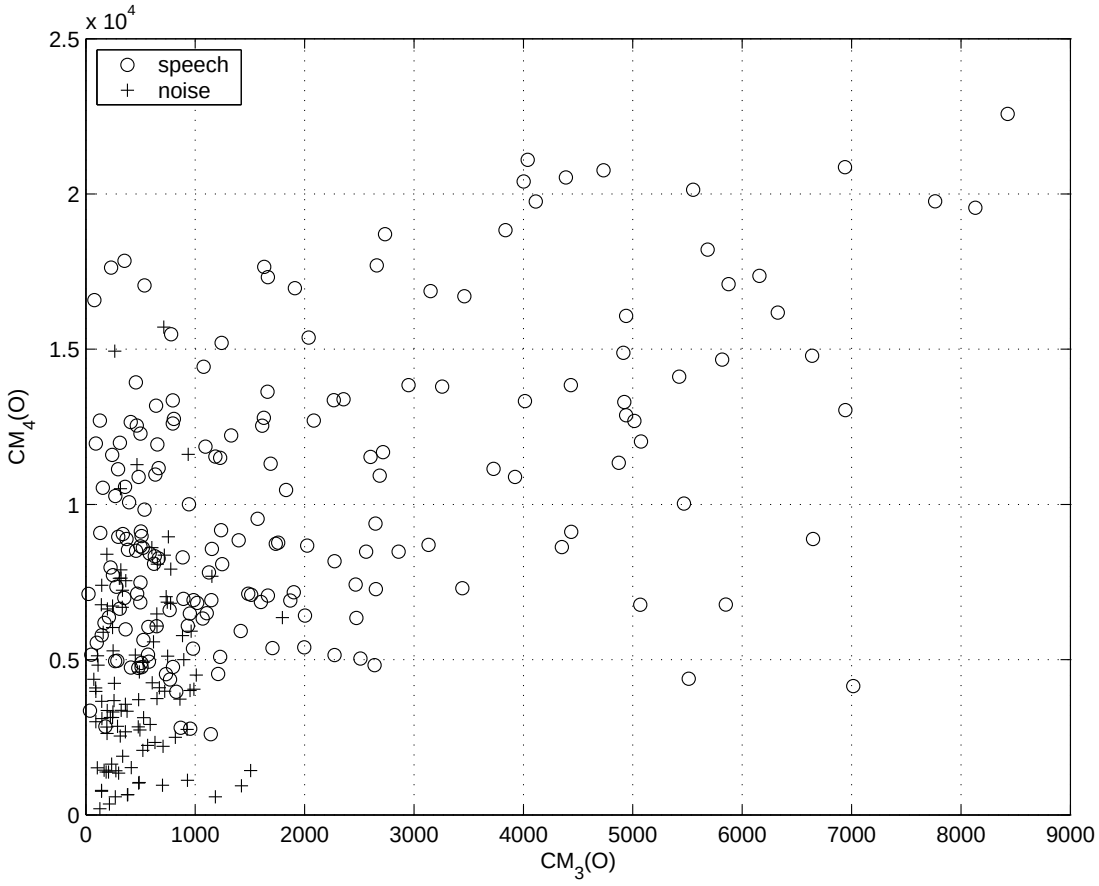


图 3.7　干扰噪声与正常语音基于后验概率的置信度分数分布

从图中可以看出基于对数似然度的2维置信度有着很好的噪声语音区分度；并且基于后验概率的置信度分数的区分度也很高，是基于对数似然度的置信度的有益补充。

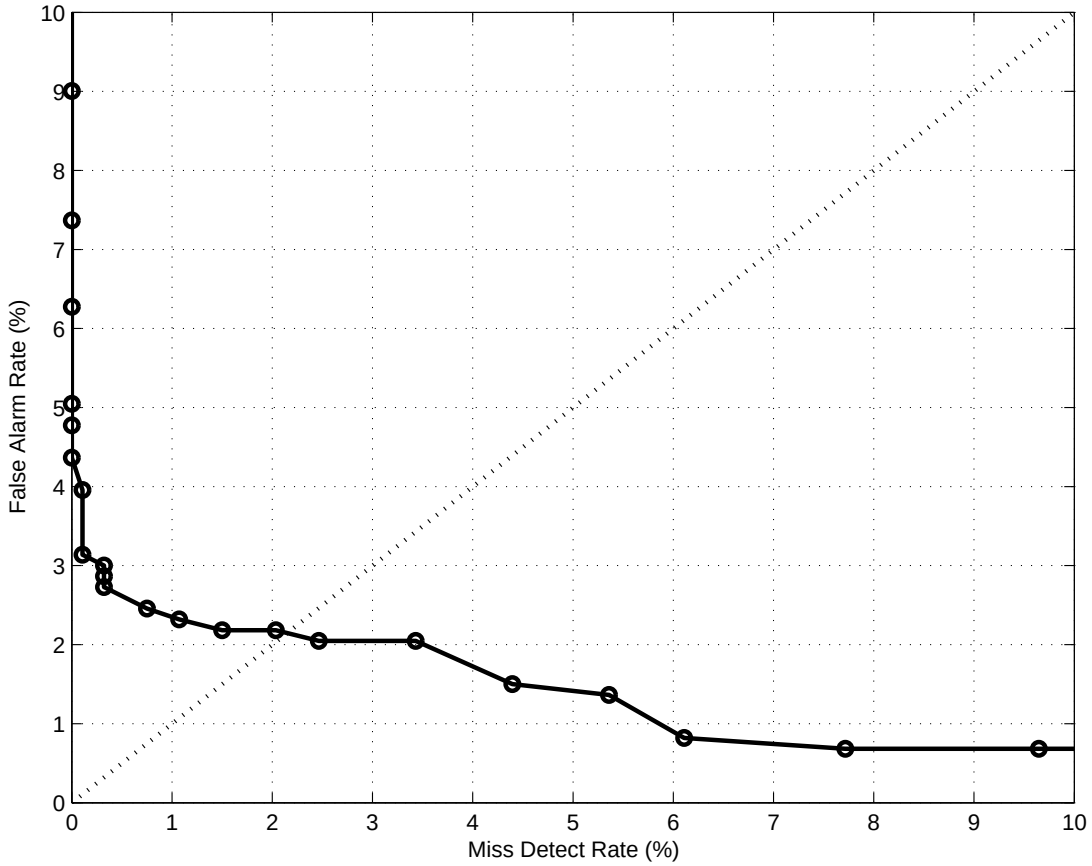


图 3.8　干扰噪声拒识DET曲线

对于我们系统而言，前面提到的假设H_0为假设该段音频输入为语音。则系统的漏报率（Miss Detection Rate）定义如下：

$$MDR = Pr(reject(H_0)|true(H_0)) \tag{3-25}$$

系统的虚警率（False Alarm Rate）定义如下：

$$FAR = Pr(accept(H_0)|false(H_0)) \tag{3-26}$$

使用拒识模块在测试集上获得的测试结果绘制的ROC曲线如图 3.8所示。

从图中可以看出本系统在测试集上的等错点（EER）被控制在2.5%以下，实际情况中，我们可以根据需要选取合适的工作点。例如当系统对漏报率有着比虚警率更高的要求时，我们可以选择从等错点出发，将工作点沿着图

3.8中的DET曲线向左上方移动。例如，系统可以工作于漏报率为0.75%，虚警率为2.46%之下。

最终，整个干扰噪声拒识模块也整合到了片上孤立词识别系统中。

3.8　本章小结

在本章中，我们根据干扰噪声的类型和特点以及当前常用的噪声拒识方法，最终设计并实现了我们自己的干扰噪声拒识模块。

我们的干扰拒识模块采用了HMM模型对噪声进行声学建模，并且采用了与正常语音类似的针对干扰噪声的线性识别网络，最后对网络输出的识别结果进行了置信度分析来确认系统输入是否为干扰噪声。

我们的干扰噪声拒识模块的在测试集上的等错点被控制在2.5%以下。最后，我们将干扰噪声拒识模块整合到了片上孤立词识别系统中，增强了原有孤立词识别系统的可用性。

第 4 章　干扰噪声拒识模块的片上实现

在本章中，我们主要介绍了我们的干扰噪声算法的定点化、汇编优化以及片上实现。

4.1　系统硬件平台介绍

我们选择了基于TMS320VC5509的DSP硬件系统作为我们的硬件平台。

4.1.1　DSP芯片介绍

DSP芯片，也称数字信号处理器，是一种具有特殊结构的微处理器。DSP芯片的内部采用程序和数据分开的哈佛结构，具有专门的硬件乘法器，广泛采用流水线操作，提供特殊的DSP 指令，可以用来快速地实现各种数字信号处理算法。根据数字信号处理的要求。

世界上第一个单片DSP芯片是1978年AMI公司宣布的S2811，此后DSP芯片得到了突飞猛进的发展，DSP芯片的应用越来越广泛。从运算速度来看，MAC（一次乘法和一次加法）时间已经从80年代初的400ns（如TMS32010）降低到40ns（如TMS32C40），处理能力提高了10多倍。DSP芯片内部关键的乘法器部件从1980年的占模区的40左右下降到5以下，片内RAM增加一个数量级以上。从制造工艺来看，1980年采用4μ的N沟道MOS工艺，而现在则普遍采用亚微米CMOS工艺。DSP芯片的引脚数量从1980年的最多64个增加到现在的200个以上，引脚数量的增加，意味着结构灵活性的增加。此外，DSP芯片的发展，使DSP系统的成本、体积、重量和功耗都有很大程度的下降。

在这么多的DSP芯片种类中，应用比较广泛的是美国德克萨斯仪器公司（Texas Instruments，TI）的一系列产品。TI公司在1982年成功推出启迪一代DSP芯片TMS32010及其系列产品TMS32011、TMS32C10/C14/C15/C16/C17等，之后相继推出了第二代DSP芯片TMS32020、TMS320C25/C26/C28，第三代DSP芯片TMS32C30/C31/C32，第四代DSP芯片TMS32C40/C44，第五

代DSP芯片TMS32C50/C51/C52/C53以及集多个DSP于一体的高性能DSP芯片TMS32C80/C82等。

TMS320C5000系列是TI公司推出的低功耗定点DSP芯片系列，待机功耗低至0.12mW，性能高达900MIPS，C5000主要应用包括：语音和图像的处理，数字音乐播放器、GPS 接收器、便携式医疗设备、MIPS 密集型语音和数据处理等个人和便携式产品，以及极其经济高效的单通道和多通道应用。本文系统选用C5000系列中的TMS320VC5509 DSP[41]作为核心运算芯片。该DSP属于C5000系列中的C55x系列DSP。通过强大的并行计算和专注于整体功耗的减少实现高性能和低功耗依然是其最大的特点。与很多C55x系列DSP相同，VC5509采用双乘法累加器，每一个可以实现17 ×17比特的乘法。为了提高代码的密度和程序空间的利用率，C5509使用可变长度的指令代码，最大长度为32比特。DSP核在1.6V的电压下，工作可达144MHz，每秒钟可进行288M乘累加运算。片内含有258K字节的RAM空间，64K字节的ROM空间。

4.1.2　基于TI TMS320VC5509 DSP的硬件系统介绍

语音信号处理系统采用基于TMS320VC5509的DSP硬件系统，系统结构如图 4.1所示。在系统中，除了DSP以外，涉及到语音信号处理算法实现的外围器件主要有：

- CODEC 采用TI公司生产的CODEC AIC23，主要实现语音信号的模数数模转换。按照算法需求，CODEC通过I2C总线设置为8KHz采样，16比特量化。线性PCM的数字语音信号通过DSP的多通道串行缓冲端口（Multichannel buffered serial ports MCBSP）与DSP进行数据的传输。
- FLASH FLASH用于长期存储程序代码，在器件上电复位时将程序载入DSP内部RAM。同时，可以用于保存长期需要的数据，例如识别用的声学模型，词表网格以及部分的语音。
- SDRAM 用于临时数据的存放，当内存空间不够的时候，可以使用SDRAM存放部分数据，一般大部分的语音数据（比如算法处理的最终结果）都存放在SDRAM中。
- CPLD 在系统中主要用于外部器件的扩展，可以实现DSP对外设的灵活控制，比如开关和发光管等等。

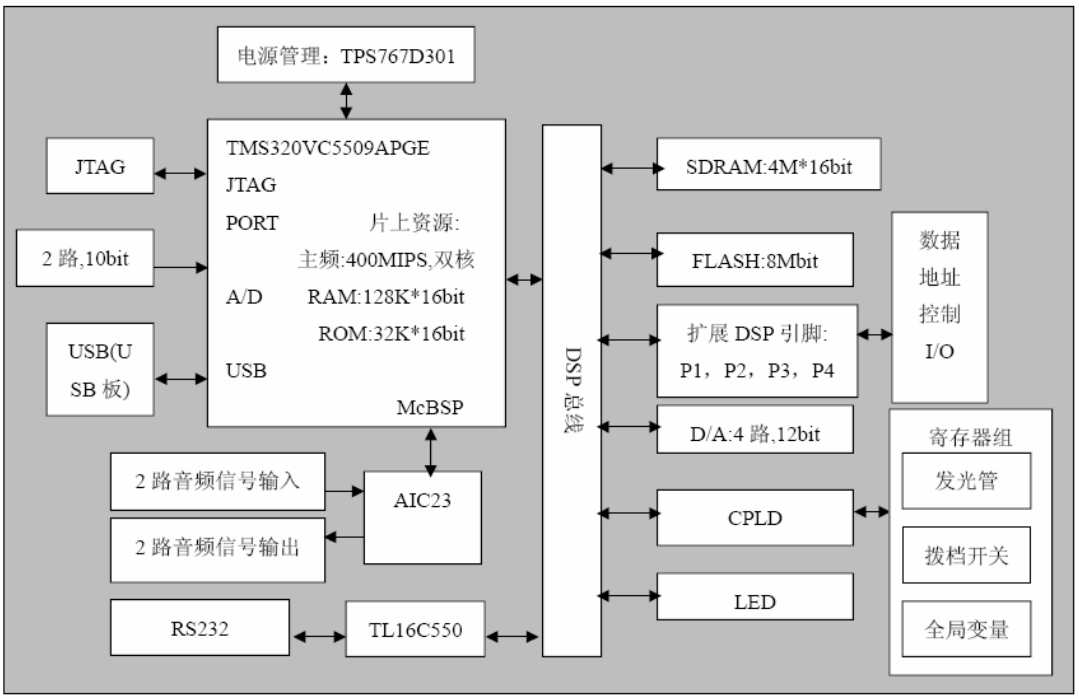


图 4.1　孤立词语音识别系统硬件系统器件结构

4.2　干扰噪声拒识算法的移植

4.2.1　定点化工作

由于本系统的核心处理单元为16位定点DSP，因此在片上实现之前需要将所有的模型参数和计算程序定点化。我们采用了基于统计的方法对干扰噪声拒识算法进行定点化处理。基本步骤如下：

在对参数x做16bit定点化时，首先统计出参数x的动态范围$[x_{min},x_{max}]$，然后根据这个动态范围选取动态范围扩展参数Q_x，使得$int(x_{min}\times 2^{Q_x},x_{max}\times 2^{Q_x})$，则参数$x$定点化后用16bit数$\hat{x}$表示为$int(x\times 2^{Q_x})$，即将x左移$Q_x$位后取整。

在我们嵌入式语音识别系统的干扰噪声拒识模块中的噪声声学模型以及置信度分数的定点化，采用的就是这种基于统计的方法，参数的动态范围是从训练数据库统计得到的。为了防止测试数据可能出现的动态范围溢出，我们在识别程序中对参数的最大值做了限制。

4.2.2　代码的汇编优化

在高级语言编译器出现以前，由于软件部分都是由汇编来完成，并且写出的代码性能都比较高，所以代码的优化在开发过程中已经完成，不需要把优化单独地作为开发的一个步骤。现在随着高级语言应用到DSP系统的开发中，在软件功能实现的基础上，软件执行效率的优化显得愈加重要。

对我们的嵌入式语音识别系统的开发而言，代码的优化主要包括两个方面：

- 根据TMS320VC5509DSP编译器的特点，进行C代码的优化，使编译器能生成更高效的汇编代码；
- 对编译器生成的汇编代码进行优化，提高执行速度。在实际的开发过程中，我们对识别过程中消耗时间较长的关键代码（如状态输出概率的计算部分）进行了汇编代码的优化。

4.3　干扰噪声拒识模块在硬件平台上的性能

完成定点化和汇编优化工作，并将干扰噪声拒识算法移植到硬件平台上之后。我们测试了片上干扰噪声拒识的性能。结果见下表：

表 4.1　片上干扰噪声拒识模块的资源消耗

	代码汇编优化前	代码汇编优化后
最短识别时间（倍实时）	1.97	0.59
最长识别时间（倍实时）	7.24	0.90
干扰噪声模块占用时间（倍实时）	0.44	0.03
干扰噪声模块程序RAM占用量（字节）	4k	4k
干扰噪声模块数据RAM占用量（字节）	2k	2k

表中的最短识别时间对应输入为干扰噪声情况，系统在第一阶段识别结束后根据输出的置信度分数直接拒识噪声，不做第二阶段识别；最长识别时间对应输入为正常语音情况，统在第一阶段识别结束后根据输出的置信度分数接受正常语音，并继续第二阶段识别，并输出最终识别结果。在代码汇编优化后，我们干扰噪声模块占用时间仅为0.03倍实时，相对于基线两阶段孤立词识别的实时0.90只增加了3%的时间开销；同时我们干扰噪声模块只增加了4kB的程序RAM占用量和2kB的数据RAM占用量。

在片上系统的端点检测算法比较稳健的情况下，使用麦克风输入时的干扰噪声拒识模块的等错点仍然能保持在5%以下。

4.4　本章小结

在本章中，我们主要介绍了我们的干扰噪声拒识模块的定点化、汇编优化以及片上实现。

与两阶段孤立词识别系统相比，我们的干扰噪声拒识模块的程序和数据占用量以及实时率都是比较低的，符合片上实现的要求。我们干扰噪声拒识模块的拒识性能在端点检测准确的情况下也是符合实际应用要求的。

第 5 章　连续音频流中的非语音音频移除

本论文的另一个主要工作就是设计并实现一个高性能连续音频流的语音/非语音分类前端。这个前端可以将无关音频段从连续广播音频流中移除，并提取出语音段作为后续系统的输入。

本章介绍并分析常用的语音/非语音分类方法的优点和不足，然后提出我们自己的语音/非语音分类器的系统框架，并对系统中的各个部分作详细说明，最后给出系统参数、训练测试数据库以及实验性能及分析。

5.1　常用的语音/非语音分类方法

当前研究人员进行了广泛而深入的语音/非语音分类研究。这些分类方法可以归类为以下三种：

5.1.1　固定时长分类法

固定时长分类（Fixed Time Classification）的方法将一段连续的音频流按照固定的长度切分成若干等长的音频段（例如2.4秒），然后使用基于能量的特征和谱特征，例如过零率（Zero Crossing Rate）特征和MFCC特征来对固定时长的音频段进行分类[42] [43] [44]。

但是这种定长切分再分类的方法容易将语音和非语音音频划入同一个音频段，这将增加语音/非语音分类的难度。减小音频段的长度有利于缓解这个段内语音/非语音混淆问题，但是这样做会增加新的分类难题：音频段长度愈短，语音/非语音分类就愈困难。

实验结果表明，对于一段段内音频数据均为相同类型（Homogeneous）的音频数据段，它的长度愈长，语音/非语音分类器就愈容易确定该音频段的类型。

5.1.2　基于分析窗的分类法

基于分析窗判决的分类法（Analysis Window-Based Classification）逐帧判断音频的类型。在判断某一帧是否为语音时，基于分析窗判决的分类法借助于以

该帧为中心的一个分析窗（例如分析窗长度为80帧），认为分析窗的音频类型就是该帧音频的类型[23]。在确定音频流中所有帧的类型后，还可以使用一些后处理的方法使判断结果更加合理。例如将若干非语音帧中出现的一帧语音的类型改变为非语音。

但是这种基于分析窗判决的方法实际上是固定时长分类方法的升级版本：在判断音频类型时，分析窗取代了固定时长的音频段。在基于分析窗判决的方法中使用的分析窗仍然有可能包含不同类型的音频，所以仍然面临着语音/非语音混淆问题。

5.1.3　基于模型的分类法

基于模型的分类法可以使用语音/非语音的HMM模型对连续音频流作Viterbi译码。Viterbi译码的结果就是这段音频流的分类结果 [45]。

但是这种基于模型分类法存在着测试时出现的音频数据与训练时使用的音频数据不匹配，即模型失配的问题。为了解决训练识别数据不一致的问题，基于模型的分类器也可以增加在线无监督自适应模块，根据新出现的音频数据更新HMM模型中的参数值，以取得更好的分类效果。但是由于广播音频流中包含了频繁的数据类型转换，这样在线无监督自适应算法就很难保证稳健地更新模型的要求。

此外，基于模型的分类法还需要提供充足的合适音频数据用于训练初始的HMM模型，这对于某些任务是不可能的。

5.2　广播音频流中的非语音移除系统介绍

我们的针对广播音频数据中的非语音移除系统最终的目的是最小化语音/非语音分类器对所有语音帧的区分错误率FER_S，对所有非语音帧的区分错误率FER_N以及对所有音频帧级别的区分错误率FER_A。FER_S、FER_N以及FER_A定义如下：

$$FER_S = \frac{N_{S\to N}}{N_S} \times 100\%$$

$$
\begin{aligned}
FER_N &= \frac{N_{N\to S}}{N_N}\times 100\% \\
FER_A &= \frac{N_W}{N_A}\times 100\% \\
&= \frac{FER_S N_S + FER_N N_N}{N_S + N_N}\times 100\%
\end{aligned}
\tag{5-1}
$$

其中$N_{S\to N}$为语音帧被判定为非语音帧的帧数，N_S为语音的总帧数，$N_{N\to S}$为非语音帧被判定为语音帧的帧数，N_N为非语音的总帧数，N_W为被错误判断类型的帧的总帧数，N_A为该段音频流包含的总帧数。

我们希望能够在判断音频的类型时，能够得到尽可能长的相同性质的音频段。也就是说，我们希望能够准确地找到不同类型音频的转换点，然后对转换点之间的相同性质的音频判断类型。

与传统的语音/非语音分类器不同，我们设计了一个多步语音/非语音区分器，能够快速并有效地将语音音频段从连续广播音频流中提取出来，并移除无关的非语音音频。系统流程图如图 5.1所示：

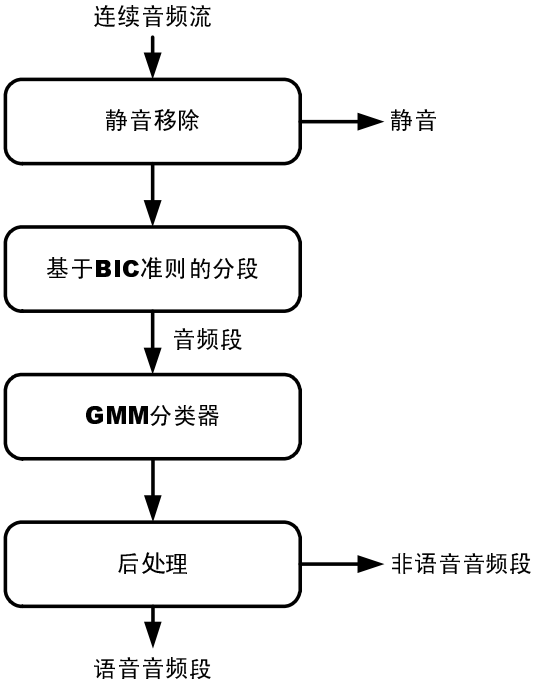


图 5.1　运行在广播音频流上的多步语音/非语音区分器的系统流程图

我们的多步语音/非语音区分系统主要采用了先分段后判断每段音频类型的

策略。下面分步介绍本系统各个组成部分。

5.3　静音移除

根据实验结果，我们发现分段模块输出的一部分语音/非语音转换点位于某些静音段的中心。这样语音音频段就会不可避免地包含一些静音帧。这些静音帧将会以异质（Heterogeneous）音频段的形式混入到原先是同质的语音音频段中。语音和静音的混合将会增加判断该段音频段是否为语音段的难度，因为后面我们会提到，系统的语音/非语音分类器中的语音GMM模型是通过不含静音的语音训练得到的。

考虑到静音在语音/非语音分类中的影响，我们认为有必要在自动分段之前将静音移除。由于广播音频的信噪比比较高，我们使用了经典的基于能量和过零率的端点检测（Voice Active Detection）算法对连续广播音频流作预处理，将静音部分移除。这里要指出的是，不只是语音部分，带噪语音、高能量噪声以及音乐部分在静音移除预处理之后仍然不会被移除。

5.4　基于BIC准则的分段

BIC准则（Bayesian Information Criterion）是一种基于最大似然度的准则[46]，因此可以应用于混合音频数据的分段。

从一段连续的音频数据中提取一段特征序列$O = o_1, o_2, \cdots, o_N, O \in \Re^d$，在这段特征序列$O$中，可能存在许多种音频类型的转换点，包括语音/非语音转换点，说话人性别转换点，说话人转换点，信道转换点等。判断该段特征序列中是否存在一个转换点，实际上是一个假设检验问题：

- **假设**H_0：表示观察矢量O对应的原始音频中不存在转换点。
- **假设**H_1：表示观察矢量O对应的原始音频中至少存在一个转换点。

假设H_0相当于假设该段音频中不存在转换点，我们可以认为观测矢量$O = o_1, o_2, \cdots, o_N$是由独立同分布的高斯过程$N_d(\mu_0, \Sigma_0)$生成的。假设$H_1$相当于假设该段音频中在$i$时刻存在一个转换点，那么我们可以认为从$i$时刻之前的观测矢量$O_1^i = o_1, \cdots, o_i$ 是由独立同分布的高斯过程$N_d(\mu_1, \Sigma_1)$生成，从t时刻之后的观

测矢量$O_{i+1}^{N}=o_i,\cdots,o_N$是由独立同分布的高斯过程$N_d(\mu_2,\Sigma_2)$生成。这样，就可以计算对于时刻$i$为音频类型转换点的似然比$Lr(i)$：

$$
\begin{aligned}
Lr(i) &= \log\frac{L(O_1^N,\mu;\Sigma)}{L(O_1^i,\mu_1;\Sigma_1)L(O_{i+1}^N,\mu_2;\Sigma_2)} \\
&= \log p(O_1^N|M)-\log p(O_1^i|M_1)-\log p(O_{i+1}^N|M_2) \\
&= \log\prod_{j=1}^{N}p(o_j|M)-\log\prod_{k=1}^{i}p(o_k|M_1)-\log\prod_{l=i+1}^{N}p(o_l|M_2) \\
&= \sum_{j=1}^{N}\log\left[\frac{1}{\sqrt{2\pi|\Sigma|^N}}e^{-\frac{(o_j-\mu)^T\Sigma^{-1}(o_j-\mu)}{2}}\right]-\sum_{k=1}^{i}\log\left[\frac{1}{\sqrt{2\pi|\Sigma_1|^i}}e^{-\frac{(o_k-\mu_1)^T\Sigma_1^{-1}(o_k-\mu_1)}{2}}\right] \\
&\quad -\sum_{l=i+1}^{N}\log\left[\frac{1}{\sqrt{2\pi|\Sigma_2|^{N-i}}}e^{-\frac{(o_l-\mu_2)^T\Sigma_2^{-1}(o_l-\mu_2)}{2}}\right] \\
&= -\sum_{j=1}^{N}\left[\log|\Sigma|^{\frac{N}{2}}-\frac{(o_j-\mu)^T\Sigma^{-1}(o_j-\mu)}{2}\right] \\
&\quad +\sum_{k=1}^{i}\left[\log|\Sigma_1|^{\frac{i}{2}}-\frac{(o_k-\mu_1)^T\Sigma_1^{-1}(o_k-\mu_1)}{2}\right] \\
&\quad +\sum_{l=i+1}^{N}\left[\log|\Sigma_2|^{\frac{N-i}{2}}-\frac{(o_l-\mu_2)^T\Sigma_2^{-1}(o_l-\mu_2)}{2}\right] \qquad (5\text{-}2) \\
&= \left[-\frac{N}{2}\log|\Sigma|+\frac{i}{2}\log|\Sigma_1|+\frac{N-i}{2}\log|\Sigma_2|\right]+\frac{1}{2}\sum_{j=1}^{N}\left[(o_j-\mu)^T\Sigma^{-1}(o_j-\mu)\right] \\
&\quad -\frac{1}{2}\sum_{k=1}^{i}\left[(o_k-\mu_1)^T\Sigma_1^{-1}(o_k-\mu_1)\right]-\frac{1}{2}\sum_{l=i+1}^{N}\left[(o_l-\mu_2)^T\Sigma_2^{-1}(o_l-\mu_2)\right] \qquad (5\text{-}3)
\end{aligned}
$$

其中

$$
\begin{aligned}
\mu &= \hat{\mu}=\frac{1}{N}\sum_{j=1}^{N}o_j \\
\mu_1 &= \hat{\mu}_1=\frac{1}{i}\sum_{k=1}^{i}o_k \\
\mu_2 &= \hat{\mu}_2=\frac{1}{N-i}\sum_{l=i+1}^{N}o_l \qquad (5\text{-}4)
\end{aligned}
$$

$$\begin{aligned}\Sigma &= \hat{\Sigma} = \frac{1}{N}\sum_{j=1}^{N}(o_j-\mu)(o_j-\mu)^T \\ \Sigma_1 &= \hat{\Sigma}_1 = \frac{1}{i}\sum_{k=1}^{i}(o_k-\mu_1)(o_k-\mu_1)^T \\ \Sigma_1 &= \hat{\Sigma}_2 = \frac{1}{N-i}\sum_{l=i+1}^{N}(o_l-\mu_2)(o_l-\mu_2)^T \end{aligned} \tag{5-5}$$

如果我们只保留$Lr(i)$最终推导结果的起前半部分取相反数，就可以得到最大似然比例统计量（maximum likelihood ratio statistics）$R(i)$：

$$R(i) = \frac{N}{2}\log|\Sigma| - \frac{i}{2}\log|\Sigma_1| - \frac{N-i}{2}\log|\Sigma_2| \tag{5-6}$$

因此i的最大似然估计，即最有可能为音频类型转换点的时刻为

$$i = \arg\max_i R(i) \qquad i \in [1,N] \tag{5-7}$$

由于可以证明$R(i) > 0$，这说明最大似然比例统计量总是倾向于认为假设H_1为真，即任意一段音频中总会存在一个类型转换点。实际上，BIC准则是一个的模型选择算法，可以对根据模型复杂度对模型施加惩罚。

$$\Delta BIC(i) = R(i) - \lambda P \tag{5-8}$$

其中，λ 为经验参数，P为惩罚因子：

$$P = \frac{1}{2}\left[d + \frac{d(d+1)}{2}\right]\log N \tag{5-9}$$

这里的d为特征维数，也代表了独立同分布高斯过程中均值向量的参数个数；$d(d+1)/2$代表独立同分布高斯过程协方差矩阵（为对称矩阵）中参数的个数。

如果i满足：

$$i = \arg\max_i \Delta BIC(i) \qquad i \in [1,N] \tag{5-10}$$

$$\text{并且} \qquad \max_i \Delta BIC(i) > 0 \tag{5-11}$$

BIC准则认为i为该段音频中的类型转折点。

下面介绍BIC准则是如何在一段连续的音频流中寻找音频类型转换点并对音频流分段的[47]。

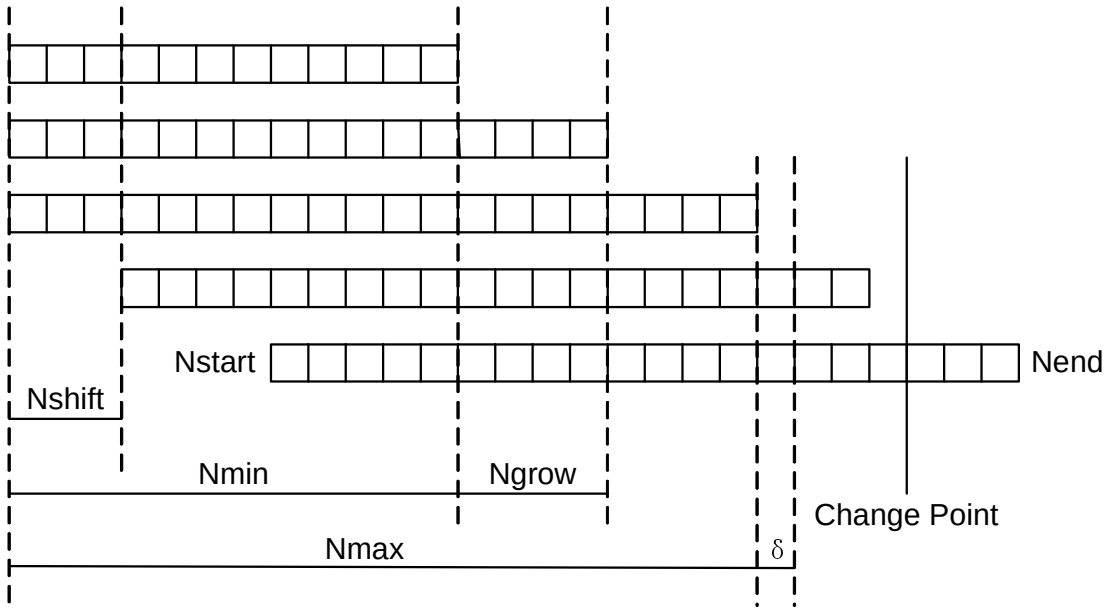


图 5.2　局部最优的基于BIC的分段示意图

如图 5.2所示，BIC分段算法步骤如下：

- **第1步**：初始化分析窗$[N_{start}, N_{end}]$：$N_{start} = 0, N_{end} = N_{min}$。
- **第2步**：如果到达了音频流的末尾：$N_{end} > N$，退出；否则，前进到第3步。
- **第3步**：使用BIC准则检测在分析窗内是否存在音频类型转换点。我们设定分辨率为δ，这就意味着该分析窗内可能的转换点的数目为$(N_{start} - N_{end})/\delta - 1$。如果未能在分析窗中检出转换点，这时应保持分析窗起始帧固定，增加分析窗的长度：$N_{end} += N_{grow}$，并前进到第4步；如果在分析窗中找到了符合BIC准则的转换点i_{max}，那么记录当前的类型转换点i_{max}，然后初始化下一次搜索使用的分析窗：$N_{start} = i_{max}, N_{end} = N_{start} + N_{min}$，并跳转到第2步。
- **第4步**：如果分析窗长度超过了N_{max}，保持分析窗长度不变向前滑动：$N_{start} += N_{shift}, N_{end} = N_{start} + N_{max}$，并跳转到第2步。

这样，在连续音频流通过基于BIC准则的分类器后，我们就得到了若干段内性质相同的音频段。

5.5　语音/非语音GMM分类器

本论文使用的语音/非语音GMM分类器与第3章中描述的GMM分类器结构基本一致，

所不同的是，我们只训练了语音和非语音的GMM模型，而没有训练更多的GMM模型，原因如下：

- 对于带噪语音或者带有背景音乐的语音，很难确定信噪比或者信号音乐比，所以训练相应的带噪语音GMM模型或者带有背景音乐的语音GMM模型也很难。
- 由于噪声类型的多样化，训练一个能够准确描述噪声分布的GMM模型并不是一件容易的事情。
- 在噪声与带有背景音乐的数据量不足的情况下训练得到的背景音乐的语音GMM模型和噪声GMM模型很难准确反映噪声或带有背景音乐的语音的真实分布。

根据以上原因，我们在语音/非语音GMM分类器中我们训练了2个GMM模型：

- **语音**（Speech，S）**模型**：使用广播音频中的纯净语音与带噪语音训练得到。
- **非语音**（Non-Speech，N）**模型**：使用广播音频中的音乐与较高音量的噪声训练得到。

因为基于BIC准则的分段模块的预期输出为同质的音频段，这对后续GMM模块能够输出可靠的分类结果是非常有利的.

5.6　分类结果后处理

在分段的音频通过语音/非语音的自动分类器之后，我们可以获得这些分段的音频的类型。从实验结果中我们发现，由基于BIC准则的分段模块输出的转换

点会与我们标定的转换点有一定的偏差，如图 5.3所示。不准确的语音－非语音转换点会造成后续的GMM分类器产生帧级别的分类错误。

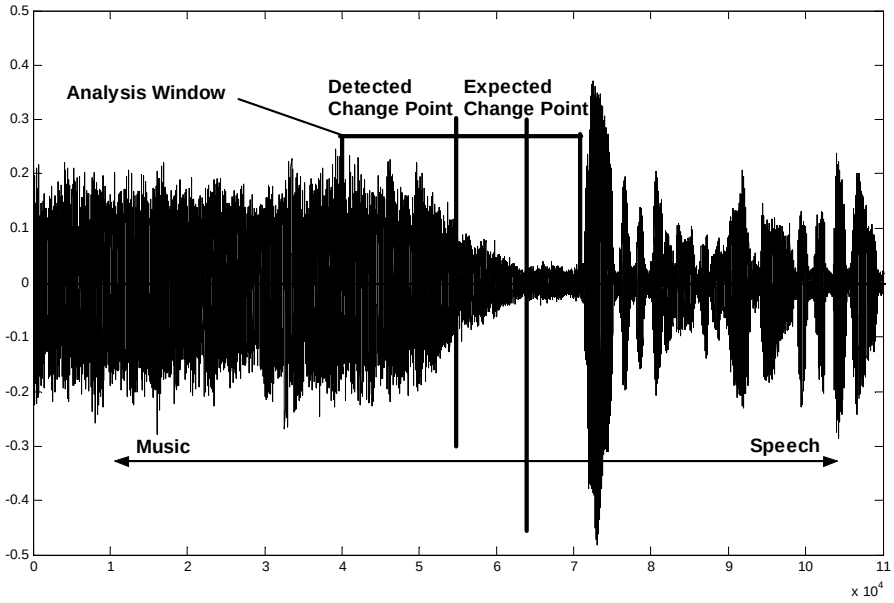


图 5.3　BIC分段程序获得的语音/非语音类型转换点与后处理模块修正的语音/非语音类型转换点示意图

为了降低帧级别的分类错误，我们在GMM分类器后面又增加了一个后处理模块。这个后处理模块能够根据事先设定的一系列规则对音频类型转换点作调整。后处理算法操作步骤如下：

- **选取转换点：** 因为基于BIC准则分段程序可能会产生过分段（Over Segment）问题，即将一整段同质音频切分为若干音频段。所以GMM分类器的输出可能会出现一个转换点两侧音频类型相同的情况。因为调整两侧音频类型相同的转换点不会影响最终帧级别的分类结果，所以这里我们将要进行调整的转换点为两侧音频类型不同的转换点。
- **设定分析窗：** 以转换点i为中心生成一个宽度为M帧的分析窗，并且这个分析窗中包含了N个音频小段，每个音频小段长度为M/N帧。
- **调整转换点：** 找到分析窗内能量包络的最低点作为调整后的转换点的位置，即找到分析窗中N个音频小段中能量最小的一个，将该音频小段的中

心i'作为调整后的音频类型转换点。

5.7　多步语音/非语音分类器的性能

5.7.1　实验使用的数据库

我们的语音/非语音分类器测试使用的数据库为实验室采集的7小时中央电视台新闻联播数据库，组成成分包括：

- **纯净语音**（63.10%）：在播音室由播音员朗读的国内/国际新闻或者现场由记者朗读的特别报道。
- **带噪语音**（15.31%）：现场录制的带有背景噪声的由记者与被采访人的对话。
- **混有音乐的语音**（3.93%）：带有背景音乐的语音段，例如广告。
- **音乐**（7.04%）：新闻与报道之间的音乐段。
- **其他**（10.41%）：语音段间与段内的静音，背景噪声，鼓掌声，呼吸声，麦克风摩擦声等。

我们的语音/非语音的GMM训练使用的数据库为实验室采集的北京电视台新闻节目数据库中的2小时混合广播音频数据，组成成分与中央电视台新闻联播数据库类似。

5.7.2　实验参数以及实验结果分析

在基于BIC准则的分段中，我们选用了15维语音特征参数，其中包括14维MFCC、和短时归一化对数能量$\log E$。优化后的一系列搜索转换点使用的参数为（帧）：

$$\begin{aligned} N_{min} &= 500 \\ N_{max} &= 2000 \\ N_{grow} &= 100 \\ N_{shift} &= 300 \\ \delta &= 25 \end{aligned} \tag{5-12}$$

由于帧移为10ms，因此基于BIC准则的分段算法的准确度为$\delta \times 10\text{ms} = 0.25\text{s}$。

在语音/非语音GMM分类器中，我们选用了28维语音特征参数，其中包括14维MFCC、14维ΔMFCC（即MFCC的一阶差分）。语音（S）模型是由共计1小时的纯净语音与带噪语音训练得到；非语音（N）模型由共计1小时的音乐和噪声音频训练得到。语音和非语音GMM模型中的高斯混合分量数（Mixture）均为512。

由于本系统希望尽可能地保留语音，所以在测试中，我们将语音的集合定义为纯净语音，带噪语音、带有背景音乐的语音；将非语音的集合定义噪声，音乐以及静音。

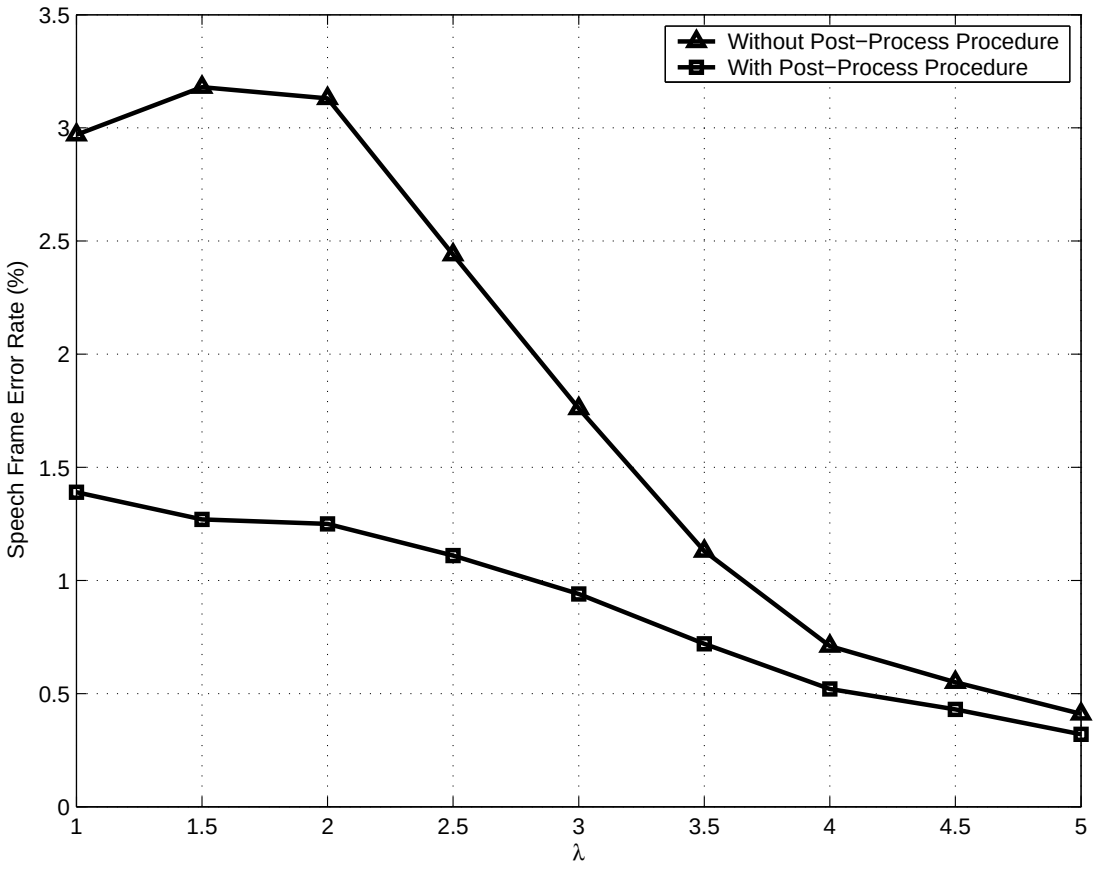


图 5.4　不同λ下语音/非语音分类器对所有语音帧的区分错误率

当基于BIC分段模块中使用的经验参数参数λ变化时，我们的语音/非语音分类系统在测试集上取得的性能如图 5.4、图 5.5以及图 5.6所示。随着λ增大，BIC准则中的惩罚因子增大，导致BIC准则更倾向于判定一段音频中不存在转换点，使得BIC分段模块输出的音频段的长度变长。由于广播音频中以语音音频居多，BIC分段模块输出的音频段的长度变长会包含入一些较短的非语音音频段，这样会造成非语音帧在后续的GMM分类模块中与所处的音频段一同被判定为语音，从而使得非语音的帧级别的区分错误率升高。

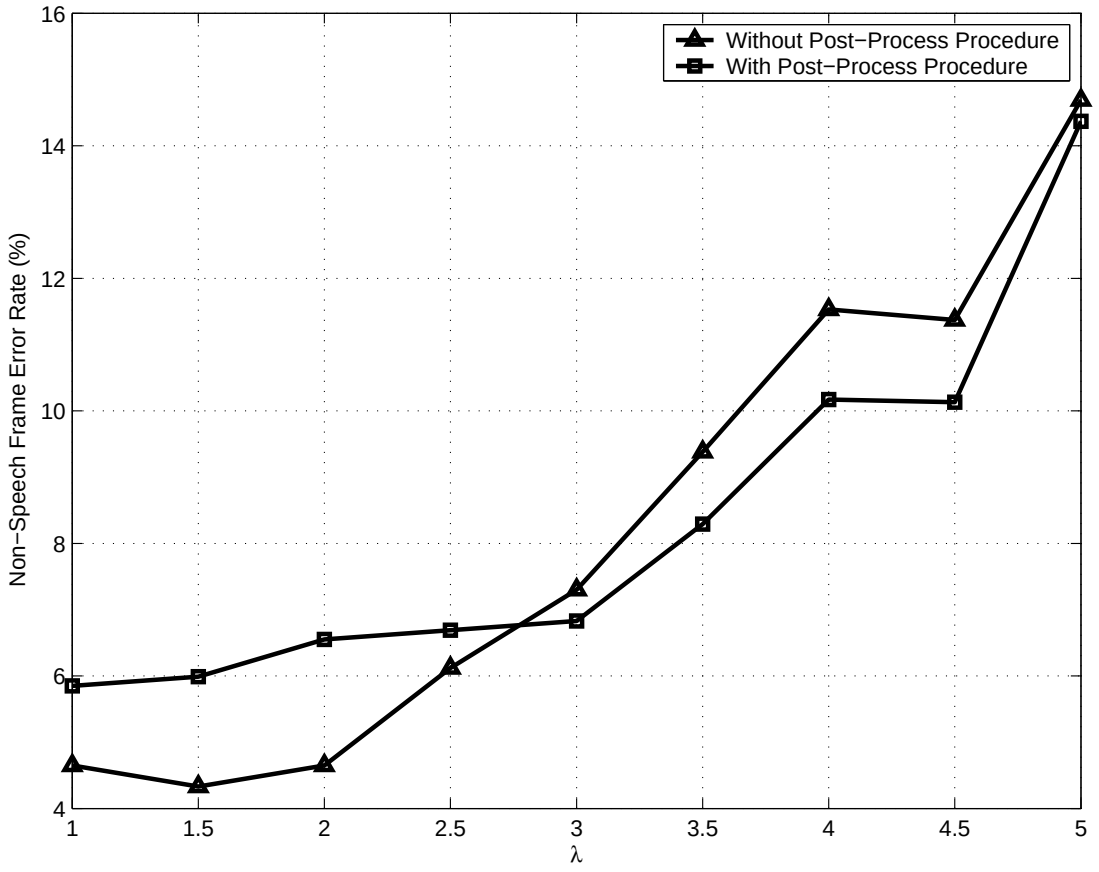


图 5.5　不同λ下语音/非语音分类器对所有非语音帧的区分错误率

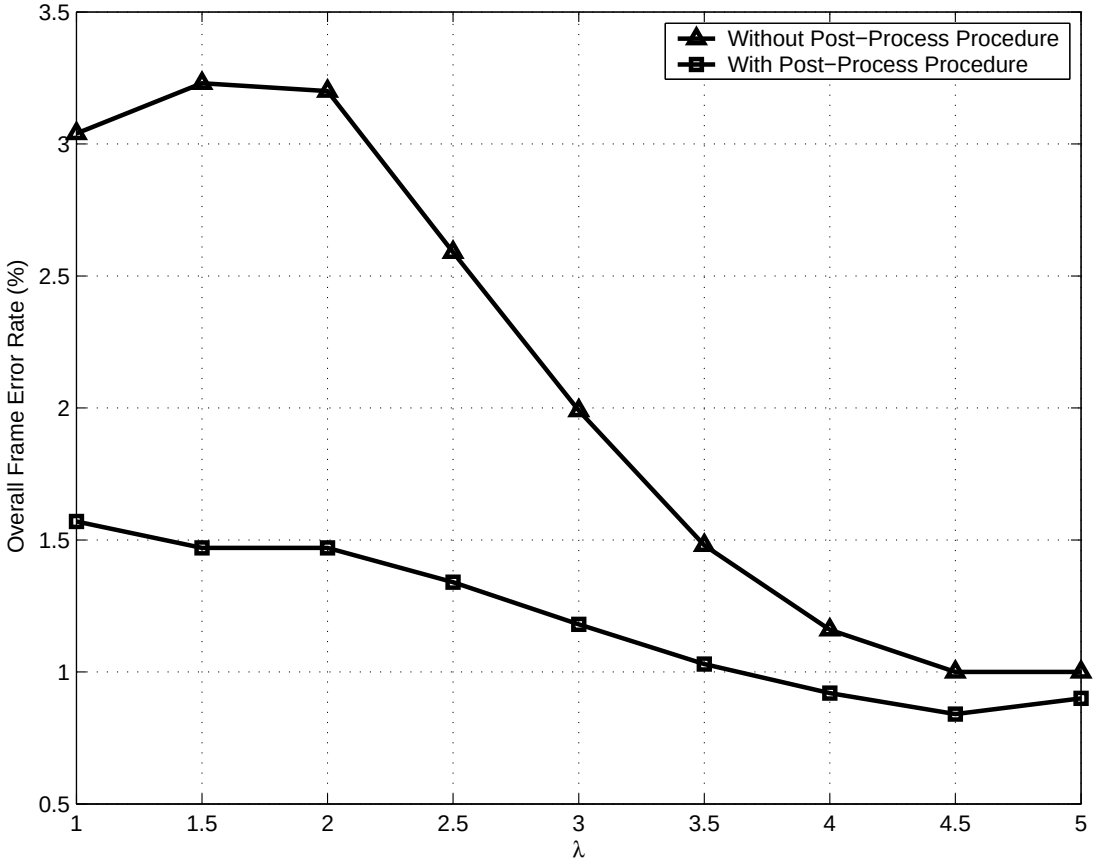


图 5.6　不同λ下语音/非语音分类器对所有音频帧级别的区分错误率

特别地，从图 5.4、图 5.5以及图 5.6中可以看出，后处理模块对系统性能的提高有一定帮助。出于稳健性的考虑，我们最终设定系统工作参数$\lambda = 3$。在$\lambda = 3$的时候，系统的性能如表 5.1所示：

表 5.1　系统工作点为$\lambda = 3$时的语音/非语音分类器的性能

	不使用后处理模块	使用后处理模块
语音帧错误率FER_S (%)	1.76	0.94
非语音帧错误率FER_N (%)	7.30	6.83
总体帧错误率FER_A (%)	1.99	1.18

特别地，虽然我们的训练数据中没有包含带有背景音乐的语音。但是在我们的系统中，绝大部分带有背景音乐的语音会被判定为语音。例如在使用后处

理模块，并且$\lambda = 3$的情况下，有93.88%的带有背景音乐的语音被归类为语音。

5.8　本章小结

在本章中，我们首先分析了当前音频分类算法的优点和不足；然后根据我们的语音/非语音的分类需要，设计并实现了我们自己的多步语音/非语音分类器。主要设计思想是先将通过BIC分段算法得到若干相同性质的音频段，然后使用GMM分类器判断BIC分段算法得到的同质音频段的类型，最后采用后处理算法调整音频类型转换点的位置，以降低分类结果的帧错误率。

我们的多步语音/非语音分类器在BIC分段算法的参数为$\lambda = 3$时，对于共计7小时的中央广播电视台新闻联播数据库，系统的总体帧错误率在1.2%以下。

第 6 章 结论

6.1 论文工作总结

本论文的主要工作在于研究如何消除无关输入音频对语音识别系统和语音信息检索系统的影响，并在片上实现了孤立词语音识别系统中的干扰噪声拒识模块，以及设计实现了针对连续广播音频流的多步语音/非语音分类器。

6.1.1 孤立词语音识别系统中的干扰噪声拒识

我们设计并实现了我们自己的干扰噪声拒识模块。我们的干扰拒识模块采用了HMM模型对噪声进行声学建模，并且采用了与正常语音类似的针对干扰噪声的线性识别网络，最后对网络输出的识别结果进行了置信度分析来确认系统输入是否为正常语音。

最后，我们将干扰噪声拒识模块整合到了片上孤立词识别系统中，使得原有的只能接受集内词条语音输入的系统也能够拒绝识别用户输入的无关语音和干扰噪声，增强了原有孤立词识别系统的可用性。

6.1.2 连续音频流中的非语音音频移除

我们设计并实现了我们自己的多步语音/非语音分类器。我们通过BIC分段算法得到若干相同性质的音频段，然后使用GMM分类器判断BIC分段算法得到的同质音频段的类型，最后采用后处理算法调整音频类型转换点的位置，以降低分类结果的帧错误率。

最后，我们将我们的多极语音/非语音分类器添加到关键词检索系统的前端，使得原有的只能在纯净语音流中检索关键词的系统也能接受广播音频流作为输入，拓展了系统的应用范围。

6.2 论文创新点

我们提出了从另一种角度处理在孤立词识别系统中出现的干扰噪声的方法，即采用一个自左向右HMM模型对出现的每一种噪声进行声学建模，并定义了

噪声解码使用的线性识别网络。我们还为干扰噪声拒识设计了4种置信度。正常语音与干扰噪声在这4种置信度构成的4维空间内有很好的区分度。我们将系统的DET曲线的等错点控制在2.5%以下。

我们还提出了一种先分段后分类的多步音频分类方法，即将说话人领域常用的BIC准则应用到音频分段中，然后使用GMM分类器判断分段模块输出的同质音频段的类型。我们实验结果也表明，我们自己的语音/非语音分类器能够比较好地完成从广播音频流中提取语音音频的任务。

6.3 未来工作展望

对于处理孤立词识别系统或者连续语音识别系统中出现的无关语音、干扰噪声以及集外词发音输入，如何建立更好的模型来描述这些干扰因素，并最终消除这些干扰因素对识别系统的影响，使得用户在使用系统时获得最大的舒适度，将成为一个研究热点。

对于音频分类和音频检索的研究也是受到了科研人员的广泛关注。音乐与噪声的产生机理、内在结构、表现方式以及对语音识别的影响，都是值得研究的题目。这说明研究人员正在致力于将原来只能在实验室环境下工作的语音识别系统或者音频信息检索系统应用到声学环境更复杂的现实生活中去。

参考文献

[1] Lee K F. Automatic Speech Recognition: The development of the SPHINX system. 1988

[2] Young S. The HTK hidden Markov model toolkit: design and philosophy. Technical report, TR.153, Department of Engineering, Cambridge University (UK), 1994

[3] Young S, Kershaw D, Odell J, et al. The HTK Book V3. Technical report, Cambridge University, 2000

[4] Huang X, Alleva F, Hon H W, et al. The SPHINX-II speech recognition system: an overview. Computer Speech and Language, 1993, 7(2):137–148

[5] Seymore K, Chen S, Eskenazi M, et al. Language and pronunciation modeling in the CMU 1996 Hub 4 evaluation. Proceedings of 1997 ARPA Speech Recognition Workshop, 1997

[6] Walker W, Lamere P, Kwok P, et al. Sphinx-4: A flexible open source framework for speech recognition. Technical report, Sun Microsystems Inc., 2004

[7] Huang X. Making Speech Mainstream. Technical report, Microsoft Corp., 2002

[8] Huang X, Acero A, Hon H W. Spoken Language Processing: a guide to theory, algorithm, and system development. 2001

[9] Boll S. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. on Acoustics, Speech and Signal Processing, 1979, 27(2):113–120

[10] Ephraim Y, Malah D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. on Acoustics, Speech and Signal Processing, 1984, 32(6):1109–1121

[11] Bell A J, Sejnowski T J. An information maximisation approach to blind separation and blind deconvolution. Neural Computation, 1995, 7(6):1129–1159

[12] Belouchrani A, Abed-Meraim K, Cardoso J F, et al. A blind source separation technique using second-order statistics. IEEE Trans. on Signal Processing, 1997, 45(2):434–444

[13] Acero A, Huang X D. Augmented cepstral normalization for robust speech recognition. Proceedings of IEEE Workshop on Automatic Speech Recognition, 1995

[14] Eide E, Gish H. A parametric approach to vocal tract length normalization. Proceedings of IEEE Int. Conf. on Acoustics, Speech, and Singal Processing, 1994. 346-348

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

[16] Leggetter C J, Woodland P C. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech and Language, 1995, 9(2):171–185

[17] Gales M J F. Maximum likelihood linear transformations for HMM-based speech recognition. Computer Speech and Language, 1998, 12:75–98

[18] Wang D, Narayanan S S. A confidence-score based unsupervised map adaptation for speech recognition. Proceedings of IEEE Int. Conf. on Signal, Systems and Computers, 2002. 222-226

[19] Hazen T J, Seneff S, Polifroni J. Recognition confidence scoring and its use in speech understanding systems. Computer Speech and Language, 2002, 16(1):49–67

[20] Cox S, Rose R. Confidence measures for the SWITCHBOARD database. Proceedings of IEEE Int. Conf. on Acoustics, Speech, and Singal Processing, 1996. 511-514

[21] Hetherington I. The problem of new out-of-vocabulary words in spoken language systems: [博士学位论文]. Cambridge MA: Massachusetts Institute of Technology, 1994

[22] Garofolo J S, Voorhees E M, Stanford V M, et al. TREC-6 1997 spoken document retrieval track overview and results. 1997. 83-91

[23] Spina M, Zue V. Automatic transcription of general audio data: Preliminary analyses. Proceedings of IEEE Int. Conf. on Spoken Language Processing, 1996. 594-597

[24] Atal B S. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. Journal of the Acoustical Society of America, 1974, 55(6):1304–1312

[25] Furui S. Cepstral analysis technique for automatic speaker verification. IEEE Trans. on Speech and Audio Processing, 1981, 29(2):254–272

[26] Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. on Speech and Audio Processing, 1980, 28(4):357–366

[27] Hermansky H. Perceptual linear predictive analysis of speech. Journal of the Acoustical Society of America, 1990, 87(4):1738–1752

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

[29] Moon T. The expectation-maximization algorithm. IEEE Signal Processing Magazine, 1996, 13(6):47–60

[30] 杨行峻, 迟惠生. 语音信号数字处理. 北京：电子工业出版社, 1995

[31] 吴宗济, 林茂灿. 实验语音学概要. 北京：高等教育出版社, 1989

[32] 王洪君. 汉语非线性音系学. 北京：北京大学出版社, 1999

[33] 朱璇. 基于子词的嵌入式语音识别系统: [博士学位论文]. 北京: 清华大学, 2003

[34] 丁玉国. 语音识别片上系统中噪声和无关语音处理方法研究: [硕士学位论文]. 北京: 清华大学, 2005

[35] Duda R O, Hart P E, Stork D G. Pattern Classification (2nd Edition). New York: Wiley-Interscience Press, 2000

[36] Bishop C M. Neural Network for Pattern Recognition. Oxford：Oxford University Press, 1995

[37] Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines. UK：Cambridge University Press, 2000

[38] Seber G. Multivariate Observations. New York: Wiley-Interscience Press, 1984

[39] 刘镜. 语音识别中置信度分析的理论和应用: [Master Thesis]. 北京: 清华大学无线电系, 2000

[40] Wessel F, Schluter R, Macherey K, et al. Confidence measures for large vocabulary continuous speech recognition. IEEE Trans. on Speech and Audio Processing, 2001, 9(3):288–298

[41] TMS320VC5509 Fixed-Point Digital Signal Processor. Technical report, Texas Instruments, 2003

[42] Saunders J. Real-time discrimination of broadcast speech/music. Proceedings of IEEE Int. Conf. on Acoustics, Speech, and Singal Processing, 1996. 993-996

[43] Scheirer E, Slaney M. Construction and evaluation of a robust multifeatures speech/music discriminator. Proceedings of IEEE Int. Conf. on Acoustics, Speech, and Singal Processing, 1997. 1331-1336

[44] Li Y, Dorai C. Instructional Video Content Analysis Using Audio Information. IEEE Trans. on Audio, Speech and Language Processing, 2006, 14(6):2264–2274

[45] Williams G, Ellis D. Speech/music discrimination based on posterior probability features. Proceedings of EuroSpeech, 1999. 687-690

[46] Chen S, Gopalakrishnan P. Speaker, environment and channel change detection and clustering via the Bayesian information criterion. Proceedings of DARPA Speech Recognition Workshop, 1998

[47] Cettolo M, Vescovi M, Rizzi R. Evaluation of BIC-based algorithms for audio segmentation. Computer Speech and Language, 2005, 19(2):147–170

致　谢

衷心感谢我的导师刘加教授三年来对我的悉心指导和关怀。他严谨求实，平易近人，无论在学业上还是生活上都给予了我莫大的关心和帮助。他的言传身教将使我终身受益。

感谢信息教研组的其他老师，包括邓北星老师，肖熙老师，欧智坚老师，他们与我进行了许多有益的讨论，并一直鼓励我勇敢向前。

感谢丁玉国师兄，是他的铺垫工作使我少走了许多弯路；感谢姚竞和宋辉同学，他们为我的算法的硬件实现提供了很有价值的建议；感谢李曜、孟莎、侯韬、钟山、路向峰、张卫强等实验室同学对我论文工作的热情帮助和支持。

最后感谢我的父亲母亲，没有他们二十多年如一日的无私哺育和爱护，就没有我今天的一切。我要把这篇论文献给他们。

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____ 日　期：_____

个人简历、在学期间发表的学术论文与研究成果

个人简历

1981 年 10 月 19 日出生于辽宁省大连市。

2000 年 9 月考入北方工业大学电子工程系，2004 年 7 月本科毕业并获得工学学士学位，

2004 年 9 月考入清华大学电子工程系攻读硕士至今。

目前已正式发表的论文

[1] Wei Chu, Xi Xiao, Jia Liu, Confidence Score Based Unsupervised Incremental Adaptation for OOV Words Detection, In Proceedings of 6th International Workshops on Statistical Techniques in Pattern Recognition (S+SSPR'06), pp. 723 - 731 (SCI检索).

[2] Wei Chu, Jia Liu, Subband Energy Distance Measure Applied in Multi-Pass Speech/Non-Speech Discrimination, In Proceedings of 20th IEEE International Conference on Information Sciences, Signal Processing and its Application (ISSPA'07) (EI检索).

[3] Wei Chu, Jia Liu, Using Confidence Measures to Evaluate the Speaker Turns in Speaker Segmentation, In Proceedings of 20th IEEE International Conference on Information Sciences, Signal Processing and its Application (ISSPA'07) (EI检索).