

## 中文摘要

由于自动语音识别做不到 100%准确，就非常值得对识别结果加以验证。理想的验证器应该能够区分正确和错误的识别结果，或者指出某个识别结果正确的可能性。本论文以检验识别假设为背景研究汉语语音识别的说话验证和置信度估计，不特别针对普通语音识别或关键词识别。所研究的方法对普通语音识别和关键词识别的识别结果验证是通用的。

语音识别的错误可以分为两类，误识和非法声响造成的系统虚警。论文指出了误识与非法声响在验证任务中的不同地位，提出分别研究对它们的拒识，并采用相应的评价方法。本论文还强调了在评价验证方法时，必须考虑具有不同性质的非法声响。本论文采集了三个不同性质非法声响库，分别对应说话噪音，随意应答和无关长句。对非法声响拒识的研究将在这三个库上展开。

论文研究了可资验证利用的信息源及其综合方法，提出了归一化音节长度方差 (Normalized Syllable Length Variance)，研究了基于 MLP 和线性模型的信息源综合和验证，取得了良好的效果。信息源综合后的验证无论在对非法声响还是对误识的拒识能力上都比单独的信息源要好。

论文提出采用 MLP/线性模型估计的后验概率验证汉语数码语音识别，在拒识 4.9%的数码语音的情况下将识别的精度从 97.1%提高到 99.6%。其验证性能远远超过了常用的反词模型和前二选验证。

论文提出采用高斯混合模型作为垃圾模型，该模型在性能和训练速度上具有优势，便于实时的计算置信度。对电话语音识别系统，在拒绝 5%的合法语音的同时可以拒绝掉几乎 100%的短非法声响和 80%的无关长句。在线垃圾模型常常用来作为研究验证的基准方法，本论文通过直接计算半音节在线垃圾似然度，并从竞争集中去除模糊半音节，显著地提高了性能。另一方面，根据半音节模型的统计相似度来减小竞争集，使运算量下降到原来的 10%左右，而保持相当的验证性能。在研究中，词表无关一直被强调。

反词模型在说话验证中被广泛使用。本论文考察了反词模型在数码语音识别验证中的效果，对基于半音节模型的识别系统，根据汉语语音的特点，特别提出了基于反半音节模型的词表无关说话验证。由于语音数据与研究力度的关系，尚未取得预期的效果。

关键词：说话验证，置信度，拒识

## Abstract

As Automatic Speech Recognition is not perfect in performance, it is desirable to incorporate a verifier, which could discriminate correct and incorrect recognition or tell us the probability for certain recognition to be correct. This thesis studies utterance verification and confidence estimation for automatic mandarin speech recognition as verification of the recognition hypothesis and the spotting putative hit. Since the study is not recognizer or spotter-oriented, it would work for both speech recognition and keyword spotting.

Errors (incorrect recognition) that a recognizer or spotter makes can be categorized as mis-recognition and false alarm caused by out of vocabulary (OOV) utterances. It is pointed in the thesis that mis-recognition and OOV utterances should be treated separately as they play different roles in causing errors. Corresponding ways for evaluation are proposed. Also, it is emphasized that OOV utterances of different natures should be taken into account in evaluation. Three OOV utterance databases are collected. They contain utterance noises, short casual responses and irrelevant long sentences, respectively. Researches for rejection of OOV utterances are carried out with these three databases.

Various knowledge sources for utterance verification and ways they could be combined are studied. Normalized syllable length variance is proposed and demonstrated to be informative. MLP is adopted to combine several knowledge sources. It boosts rejection of not only mis-recognition but also OOV utterances.

Rejection based on a Posteriori Probability estimated by MLP/Linear Model is proposed for Mandarin Voice Dialer. Rejecting 4.9% of all the testing utterances, the MLP rejecter boosts the single digit accuracy from 97.1% to 99.6%. Such performance is much better than those of rejection based on anti-digit models and likelihood ratio.

Gaussian Mixture Models (GMMs) is proposed as the garbage model trained by general speech database to estimate the filler likelihood. This model is better with respect to verification performance and training. Moreover, it is ready to provide confidence real-timely. Rejecting 5% within domain utterances, it can reject almost all the short OOV utterances and about 80% irrelevant long sentences for the Telephone Speech Recognition System. Online garbage model is usually adopted as the benchmark method for studying verification. However, it is found that its performance could be improved by deleting the most confusable phoneme from the competing set or computing the phoneme online garbage likelihood directly. On the other hand, after competing sets are optimized according to statistical similarities between models, the computation load is reduced by 90% while the verification performance remains decent. Throughout these studies, Vocabulary Independence is emphasized.

Anti-word modeling is widely employed in utterance verification. It is employed in the Mandarin Voice Dialer. For the semi-syllable-based mandarin speech recognition systems, anti-semi-syllable modeling is proposed for vocabulary independent verification. Due to the lack of proper databases and research efforts, good performance expected hasn't yet been achieved.

**Key words: Utterance Verification, Confidence Measure, Rejection**

## 第一章 说话验证

本章将综述本论文的选题意义，概要介绍相关历史,研究现状以及论文的安排。

### 1.1 什么是说话验证

自动语音识别系统将输入的声音映射为文本，给出声音的内容，实际上是给出有关输入声音内容的假设。在许多情况下，我们不仅关心假设的内容，还关心假设有多么可靠，也就是说假设在多大程度上是正确的。这个概率就是严格意义上的置信度（Confidence Measure）。估计置信度，并据此对识别结果的正确性做出判断就是说话验证（Utterance Verification）。首先需要区别置信度与系统的识别率（或者称为精度）。识别率是指，在系统识别的语音中，识别结果是正确的所占的比率。而对一个输入识别系统的声响，会提供给我们一组观测值 $\Xi$ ，而当它经识别系统处理后，识别系统又会提供给我们另一组观测值 $\Psi$ ，这两组观测值构成了我们对该声响的全部知识 $K = (\Xi, \Psi)$ 。置信度就是指，当观察到知识 $K = (\Xi, \Psi)$ ，识别结果正确的后验概率。换句话说，有许多输入使观测值为 $K$ ，其中被系统正确识别的输入所占的比例就是这样的输入的置信度。广义的置信度可以是正确概率的任意一种单调映射结果。

### 1.2 为什么要说话验证

只要自动语音识别不是 100%的可靠，如果能给出识别结果的可靠性并对识别结果的正确性加以验证就会有利于减少识别错误。而在下列情况下置信度估计与验证是非常重要的。

第一，识别系统经常遇到非法声响（Out of Vocabulary Utterances, OOV Utterances）。诸如关键词识别系统（Keyword Spotter），口语对话系统（Dialog Systems）和使用环境恶劣的识别系统。利用验证可以使系统降低虚警率（False Alarm Rate），提高抗干扰和噪声能力（Resistance to Interfere and Noise）。

第二，识别错误代价非常高。诸如语音拨号，重要设备的语音操作（Maneuver by Voice）和语音确认（Voice Confirmation）系统。验证对误识可以起到“宁可错拒一千，不可放过一个”的作用，降低系统运转代价。

第三，需要利用识别结果进行下一步操作，而识别结果的正确性将影响下一步的在何种程度上依赖识别结果。比如，无监督的说话人自适应（Unsupervised Speaker Adaptation），文本相关的说话人识别（Text-dependent Speaker Recognition），自动语音翻译（Automatic Speech to Speech Translation）和多模式人机交互系统（Multi-Modal Human-Machine Interaction）。

第四，比较两个精度接近的语音识别系统。在识别精度接近的情况下，如果一个识别系统正确和错误的识别结果在置信度上有更大的区分性，这个系统显然要更好一些，因为它可以更可靠地告诉我们什么时候相信它。

以上几种情况对于语音识别技术的深入发展和应用都至关重要。因此本论文的选题具有深刻的理论意义和实用背景。

## 1.3 研究综述

### 1.3.1 历史

说话验证从关键词识别研究发展出来，现在已经用到几乎所有的语音识别问题中。下面通过对其产生和发展过程的回顾来综述国内外发展动态和文献。

根据（郑方，1997），关键词的研究始于1973年的Bridle，当时叫作“Detecting Given Words in Running Speech”，采用模板匹配。关键词（Keyword）的提法是1977年由Christiansen首先采用的。1985年，Higgins第一次在关键词识别中采用了filler（补白）方法。由于系统基于DTW模板匹配，此时的补白还是“补白模板”（Filler Template）。

随着HMM方法在语音识别中的流行，（Wilpon et al, 1990）提出了一个基于HMM方法的关键词识别系统，用来在交换机上自动识别用户的接通命令。这一个是5关键词的识别系统，假定输入语音中最多包含一个关键词。也就是说，这个系统一次只能检测出一个关键词。因此从这一点看来，它仍是基于孤立语音识别技术的关键词识别系统。作者提出了与补白模板对应的垃圾模型（Garbage Models，也

称为 Filler Models 和 Sink Models), 用来对非关键词语音建模, 区分关键词语音和非关键词语音。垃圾模型成为说话验证最重要的方法之一。

MIT 林肯实验室的 Rose 和 Paul 接着于 1990 年提了第一个基于连续语音识别技术的 HMM 关键词识别方法(Rose & Paul, 1990)。这个系统用于连续的对话语音 (Conversational Speech) 中关键词的识别, 由于采用连续语音识别技术, 补白模型 (Filler Model) 和部分维特比回溯技术 (Partial Viterbi Backtrace), 可以识别出语音流中的任意多个关键词, 用于语音监听。

随着 HMM 的 MMIE 训练算法的提出, (Rose, 1992) 第一次将基于 MMIE 的区分训练 (Discriminative Training) 技术引入到关键词识别中来, 这导致了区分技术后来在说话验证中的广泛应用 (Rahim, Lee and Juang, 1997) (Sukkar and Lee, 1996)。

至此的关键词识别系统及其说话验证都是词表/任务相关的 (Task Dependent or Vocabulary Dependent), 也就是说, 系统是针对特定的词表/任务训练和调试的, 如果要更新词表改变任务, 必须重新采集语音库, 重新训练。由于诸如音频信息检索这样的应用要求根据使用者需要迅速更新词表, 词表相关的说话验证就显得力不从心了。正是在这种需求的推动下, 关键词识别和说话验证研究迅速转向了词表/任务无关 (Task Independent or Vocabulary Independent) 系统 (Hofstetter & Rose, 1992) (Rose and Hofstetter, 1993) (James and Young, 1994) (Rose, 1995) (Sukkar & Lee, 1996) (Foote et al, 1997) (Sukkar, 1998)。直到今天这仍然是说话验证研究中的一个热点。

1993 到 1994 年 (Boite et al, 1993) 和 (Bourlard et al, 1994) 提出在线垃圾模型 (Online Garbage Models) 方法, 现在已经成为比较验证方法常用的基准方法 (Benchmark Method)。

当 HMM 的 MCE (Minimum Classification Error) 训练算法开始流行时, MIT 的 Lippmann 等人 (Lippmann et al, 1994) 提出了关键词识别的 Figure-Of-Merit Back Propagation 训练算法, 同样把优化目标转换成和系统性能直接相关的 FOM 值。这一思路直接产生了说话验证的最小验证错误训练 (Minimum Verification Error, MVE) 算法 (Sukkar et al, ICASSP 1996) (Rahim & Lee, 1997) (Sukkar, 1998)。

(Rahim, Lee and Juang, 1997) 提出反词模型 (Anti-Word Model) 来提高英语连续数码识别的验证。反词模型已经被证明十分有效, 并得到广泛应用 (Jouvet et al, 1999)。

同时，随着语音识别技术的成熟和投入实用，人们发现即使非关键词识别系统也经常遇到词表中没有的新词和无关的声响。如何检测出这些词表外的说话对提高识别系统的自然度和更新识别系统很重要。说话验证自然就被应用到普通识别系统中来 (Young, 1994) (Colton, 1997)。包括通过检测新词，更新大词表语音识别的词表和语言模型 (Kemp and Jusek, 1996) (Matsunaga and Skamoto, 1996)；拒绝会导致系统错误启动的非法声响 (Mathan and Miclet, 1992) (Villarrubia and Acero, 1993) (Rivilin et al, 1996) (Jitsuhiro et al, 1998) 等等。另一方面，语音识别也越来越多地与其他技术结合去解决比语音识别更难的问题，诸如 Spoken Language Understanding (Kawahara et al, 1997) (Bouwman et al, 1997) (Rose et al, 1998)，多模式人机交互 (Chen and Rao, 1998) 等。还有一些相关的任务如无监督说话人自适应和文本相关说话人识别，也需要语音识别提供的结果。在这些应用中，语音识别结果的正确性会影响整个任务的完成情况。在另一些系统中需要把多个语音识别器的结果综合起来 (Kirchhoff and Bilmes, 1999)，也需要评价各个识别器结果的可靠性。这些都成为了置信度估计与说话验证研究新的应用背景。

关键词识别，置信度估计和说话验证的研究在国外已经进行多年，正在进入高潮，而在国内的研究则刚刚起步 (郑方, 1997) (徐明星等, 1998) (刘加等, 1998) (韦晓东等, 1998)。其中 (郑方, 1997) (徐明星等, 1998) 提出了一个基于音节的汉语无限制语音流的关键词识别系统，采用了独特统计拒识方法。(刘加等, 1998) 采用了类似 (Foote et al, 1997) 音子网格 (Phone Lattice) 的方法，利用前二选识别结果进行拒识，取得一定的效果。(韦晓东等, 1998) 的报道了垃圾模型在拒识中的应用，这是国内见诸文献的第一家。

### 1.3.2 不同应用背景下的验证

下面将按不同的应用背景对当前的置信度估计和说话验证研究加以综述。

#### ✓ 对话系统 (Dialog System)

对话系统是目前语音识别研究与其他学科结合与应用最热门的领域。语音识别的结果需要与其他许多模块结合起来才能完成实时人机对话的任务，因此识别结果的可靠性非常重要。这样的系统包括 MIT 的天气报告 Jupiter 系统 (Zue et al, 2000)，AT&T 实验室的自动电话转接任务 (Automatic Call Routing Task) (Riccardi et al, 1997) (Rose et al, 1998)，Bell Lab, Lucent Tech 的汽车预定任务 (Car Reservation Task)

和电影查询任务 (Movie Locator Task) (Kawahara et al, 1997) (Kawahara et al, 1998), Philips 公司的欧洲自动铁路信息系统 (Automatic Railway Information Systems in Europe, ARISE) (Bouwman et al, 1999) 等等。

#### ✓ 监听系统 (Surveillance)

语音识别的自动监听因为其军用目的而很早就得以发展。自动监听需要从语音流中实时地报告关键词 (往往是敏感的军事, 政治, 经济话题) 出现, 而且漏报 (False Rejection) 的代价较高。R. Rose 从八十年代末开始在 MIT 的林肯实验室研究 (Rose and Paul 1989) (Rose, 1992), 后来将研究带到 AT&T Bell Lab (Rose, 1995) (Lleida and Rose, 1996)。BBN 系统与技术有限公司也一直在进行类似的研究 (Rohlicek et al, 1993) (Jeanrenaud et al, 1994)。

#### ✓ 语音数据库检索系统

这是由互联网发展产生的需求。由于互联网上大量音频 (包括语音) 数据的存在, 如何对它们进行内容标注和检索 (Indexing and Retrieving) 显得重要。这样的任务要求关键词识别和验证词表无关, 但是标注不需要实时进行。剑桥的 Olivetti Research Ltd, 很早就开发出一个实验系统—Pandora System (Hopper, 1990)。包括 MIT 林肯实验室 (Rose et al, 1991) 和 BBN 系统与技术有限公司 (Rohlicek et al, 1992) 都在进行这方面的研究。剑桥大学工程系在这一领域处于领先地位 (James and Young, 1994) (Foote et al, 1997)。

#### ✓ 大词表连续语音识别系统的置信度标注

对现有的大词表连续语音识别系统进行置信度标注有许多潜在的用途, 包括对识别系统进行自适应, 将识别系统加入到自然语言理解, 多模式人机交互中等。最成功的例子是美国 Carnegie Mellon 大学和德国 Karlsruhe 大学为他们合作开发的自动语音翻译 (Automatic Speech to Speech Translation) 系统 JANUS (现在版本已经到 JANUS-3) 研制的置信度标注器 (Confidence tagger) JANKA 系统。这个系统以 0/1 的方式给出对识别假设正确性的判断, 减小识别错误对翻译系统的困扰 (Schaaf and Kemp, 1997)。在英国, 研究者也为剑桥大学基于神经网络/HMM 的 ABBOT 大词表连续语音识别系统开发出了置信度估计系统 (Williams and Renals, 1999)。OGI 也一直在它的 MLP/HMM 混合大词表连续语音识别系统上展开

置信度与拒识的研究 (Colton, 1997)。

最后愿意概括一下说话验证领域主要的研究机构和研究者, 1) Bell Labs, Lucent Tech 的 M. Rahim 和 R. A. Sukkar 等; 2) AT&T 实验室的 R. Rose 等; 3) 德国 Karlsruhe 大学的 T. Kemp 和 T. Schaaf ; 4) 斯坦福研究院 (SRI) 的 M. Weintraub 和 Z. Rivilin 等人 (他们中的许多已经离开 SRI 加入了语音技术公司 Nuance Communications); 5) BBN 系统与技术的 J.R.Rohlicek, H. Gish 和 M. Siu 等。由于这些优秀的研究人员和研究机构的存在。语音识别的置信度估计和说话验证已经拥有了相对独立和稳定的学术环境。关于置信度与说话验证的论文已经多次出现在 IEEE Transactions on SAP, Speech Communications 和 Computer, Speech & Language 等语音识别界的国际权威刊物上, 语音识别界的权威国际会议 IEEE ICASSP, EuroSpeech, ICSLP 和 IEEE Workshop on ASRU 也每年辟出专题, 在会议 Proceedings 上收录论文报道这一领域的最新进展。置信度估计和说话验证正在进入研究的黄金时期。

## 1.4 本论文工作

本论文将以检验识别假设为背景研究说话验证和置信度估计，因此不特别针对普通语音识别或关键词识别，不涉及普通语音识别和关键词识别本身的问题。所研究的方法对普通语音识别和关键词识别的识别结果验证是通用的。许多验证的基本方法都得到了研究，将在论文的不同章节中出现。

第二章将对说话验证的数学原理进行分析，论述本论文评价说话验证的方法和语音数据库，并且引入本论文研究说话验证的识别系统，包括基于整词和基于子词的识别系统，包括基于孤立语音识别和连续语音识别的系统。

第三章将论述可资说话验证利用的信息源。提出了利用音节长度方差对错误识别结果进行拒识。强调了利用语音结构信息的重要性。

第四章以汉语数码语音识别为背景，研究了 MLP 估计后验概率在说话验证中的应用。提出了用 HMM 迹和 MLP 估计后验概率拒绝错误识别的方法。并将其与反词模型，线性模型和似然比等拒识方法进行比较。

第五章在基于半音节的识别系统：电话语音识别系统和语音确认系统上研究了垃圾模型和在线垃圾模型在任务/词表无关说话验证中的应用和改进。提出了采用高斯混合垃圾模型和直接从半音节计算的在线垃圾似然度，研究了优化计算在线垃圾似然度竞争集的方法。在研究过程中注意了方法的任务/词表无关性。对比实验证明了这些方法的有效性。

第六章研究了多个信息源的综合利用方法，包括基于规则和基于统计模型的综合方法。研究了利用 MLP 和线性模型综合信息源的方法。

第七章总结全文并给出对今后研究工作的展望。

## 第二章 数学与评价

统计假设检验(Statistical Hypothesis Testing)和贝页斯决策分析(Bayesian Decision Analysis)是说话验证和置信度估计的数学基础。对某个输入语音 $\vec{X}$ ，语音识别器(Recognizer)给出识别结果 $C$ ；这个结果实际上是关于输入语音的一个假设(Hypothesis) $H(\vec{X})$ 。而对此结果验证(Verification)的主要手段就是统计假设检验(Statistical Hypothesis Testing)。另一方面，如果将识别器的输出看成是验证器(Verifier)的输入，那么验证器实际是一个分类器(Classifier)，判断输入究竟属于类1(正确)还是类0(错误)。因此，说话验证又可以看成是模式分类(Pattern Classification)问题。而贝页斯决策分析正是模式分类的统计数学基础。本章将介绍说话验证的数学原理和评价方法，分析两个不同数学角度的内在联系，最后给出本论文将要用到的验证评价方法和研究采用的识别系统。

### 2.1 统计假设检验

根据 Neyman-Pearson 假设检验理论 (Bickel and Doksum,1976)，可以将说话验证归结为这样一个统计假设检验问题。零假设 $H_0$  (Null Hypothesis): 识别结果正确。与之对应是备选假设 $H_1$  (Alternative Hypothesis): 识别结果错误。说话验证就是对零假设进行检验。根据假设本身的性质(真/假)以及假设检验的结果(接受/拒绝)，有以下四种结果出现：正确接受( $H_0$ 真)，错误拒绝( $H_0$ 真)，错误接受( $H_0$ 假)和正确拒绝( $H_0$ 假)。因此假设检验可能出现两种错误：错误拒绝(False Rejection)和错误接受(False Acceptance)，分别称为第一类错误和第二类错误。出现两类错误的概率分别为 $P(I)$ 和 $P(II)$ 。假设检验的势(Power)为 $1 - P(II) = P(\text{拒绝}H_0 | H_0\text{假})$ 。设输入识别器的语音为 $\vec{X}$ ，而分布 $p(\vec{X} | H_0)$ 与 $p(\vec{X} | H_1)$ 已知，根据 Neyman-Pearson 引理，当

$$LR = \frac{p(\vec{X} | H_0)}{p(\vec{X} | H_1)} > \tau$$

时接受零假设  $H_0$  是优的。所谓优是指使在使  $P(I)$  受限的情况下, 检验的势(Power) 是所有可能检验中最大的。

这通常称为似然比检验(Likelihood Ratio Test)。其中  $\tau$  称为检验的临界阈值(Critical Threshold)。 $\tau$  取不同值, 假设检验将工作在不同的工作点(Operating Point)。

从似然比建设检验的角度来看说话验证, 对不同说话验证统计方法, 实际是从不同的角度来估计分布  $p(\vec{X} | H_0)$  与  $p(\vec{X} | H_1)$ 。

表 2-1

	接受 $H_0$	拒绝 $H_0$
$H_0$ 真	N(A, T)	N(R, T)
$H_0$ 假	N(A, F)	N(R, F)

设我们有  $N$  个识别结果来评测检验算法, 对应不同识别结果性质和不同检验结果的样本数见(见表 2-1)。其中  $N(X, Y)$  表示  $H_0$  为  $Y$  (T/F) 且假设检验结果为  $X$  (A/R) 的测试样本数, 而样本总数  $N = N(A, F) + N(A, T) + N(R, F) + N(R, T)$ 。用这些样本可以估计检验算法的性能参数。系统的(无条件)错误率(Unconditional Error Rate) 估计为:

$$\bar{P}_e = \frac{N(A, F) + N(R, T)}{N};$$

类似, 可以分别估计两类错误率如下:

错误拒绝率 (False Rejection Rate) :

$$\bar{P}(I) = \bar{P}(\text{拒绝}H_0 | H_0\text{真}) = \frac{N(R, T)}{N(A, T) + N(R, T)};$$

错误接受率 (False Acceptance Rate) :

$$\bar{P}(II) = \bar{P}(\text{接受}H_0 | H_0\text{假}) = \frac{N(A, F)}{N(A, F) + N(R, F)};$$

二者统称为条件错误率(Conditional Error Rate)。而检验的势为

$$1 - \bar{P}(II) = \bar{P}(\text{拒绝}H_0 | H_0\text{假}) = \frac{N(R, F)}{N(A, F) + N(R, F)}。$$

## 2.2 贝页斯决策分析

换个角度看说话验证。设识别器给出识别结果  $H$  以及识别过程中得到的特征矢量  $\vec{S}$ 。那么验证器的任务是根据  $\vec{S}$ ，把  $H$  分类为正确或错误，分别对应假设检验中的接受和拒绝。这样分类也有四种错误。从这个角度来看，不同的说话验证方法实际是从不同的角度来估计分布  $P(H\text{正确}|\vec{S})$  与  $P(H\text{错误}|\vec{S})$ ，或者是  $P(H\text{正确}|\vec{X})$  与  $P(H\text{错误}|\vec{X})$ 。设将正确识别拒绝的损失为  $A$ ，将错误识别接受的损失为  $B$ ，正确的拒绝和接受损失为  $0$ 。如下表

表 2-1

	接受	拒绝
正确	$\lambda(\text{接受, 正确}) = 0$	$\lambda(\text{拒绝, 正确}) = A$
错误	$\lambda(\text{接受, 错误}) = B$	$\lambda(\text{拒绝, 错误}) = 0$

设对输入  $\vec{S}$ ，采取的决策（接受/拒绝）为  $\alpha(\vec{S})$ ，则验证的条件期望风险为

$$R(\alpha(\vec{S})|\vec{S}) = \lambda(\alpha(\vec{S}), \text{正确}) \cdot P(H\text{正确}|\vec{S}) + \lambda(\alpha(\vec{S}), \text{错误}) \cdot P(H\text{错误}|\vec{S})$$

$$R(\text{接受}|\vec{S}) = B \cdot P(H\text{错误}|\vec{S})$$

$$R(\text{拒绝}|\vec{S}) = A \cdot P(H\text{正确}|\vec{S})$$

验证的期望风险为  $R = \int R(\alpha(\vec{S})|\vec{S}) \cdot p(\vec{S}) d\vec{S}$ ，理想的验证决策应该使  $R$  最小。如果在进行每个验证决策时，都使条件期望风险最小，就能使在对所有  $\vec{S}$  验证时，其期望风险也最小。这就是最小风险贝页斯决策。因此理想的验证决策是，

当  $R(\text{接受}|\vec{S}) = B \cdot P(H\text{错误}|\vec{S}) < R(\text{拒绝}|\vec{S}) = A \cdot P(H\text{正确}|\vec{S})$  时，接受识别结果。又因为  $P(H\text{错误}|\vec{S}) = 1 - P(H\text{正确}|\vec{S})$ ，上述准则可以简化为，

$$\text{当 } P(H\text{正确}|\vec{S}) > \frac{B}{A+B} = \frac{1}{1+A/B} \text{ 时,}$$

接受识别结果。将上式的右边用域值  $\tau$  代替，对应不同的阈值，也会有不同的验证工作点。当第一类错误的代价相对第二类错误的代价越大， $\tau$  越小；反之， $\tau$  越大。

直观地看, 当错误接受的代价相对大时, 要验证接受一个识别结果就困难, 接受的条件就越苛刻。

从贝页斯决策的角度来看说话验证, 实际上是对  $\vec{S}$  估计  $P(H\text{正确}|\vec{S})$ 。

## 2.3 联系

可以观察从似然比假设检验与从贝页斯决策角度解决说话验证问题的联系。

$$\begin{aligned} LR &= \frac{p(\vec{X}|H_0)}{p(\vec{X}|H_1)} = \frac{p(\vec{X}|H_0)p(H_0)}{p(\vec{X}|H_1)p(H_1)} \cdot \frac{P(H_1)}{P(H_0)} = \frac{p(\vec{X}, H_0)}{p(\vec{X}, H_1)} \cdot \frac{P(H_1)}{P(H_0)} = \frac{p(H_0|\vec{X})}{p(H_1|\vec{X})} \cdot \frac{P(H_1)}{P(H_0)} \\ &= \frac{p(H\text{正确}|\vec{X})}{p(H\text{错误}|\vec{X})} \cdot \frac{P(H_1)}{P(H_0)} \end{aligned}$$

当  $H_0$  与  $H_1$  互补时, 有  $P(H_1) = 1 - P(H_0)$ ; 而  $P(H_0)$  正是识别系统的识别率  $P$ , 因此有

$$LR = \frac{p(\vec{X}|H_0)}{p(\vec{X}|H_1)} = \frac{p(H\text{正确}|\vec{X})}{p(H\text{错误}|\vec{X})} \cdot \left(\frac{1}{P} - 1\right)$$

从而

$$LR = \frac{p(\vec{X}|H_0)}{p(\vec{X}|H_1)} = \frac{p(H\text{正确}|\vec{X})}{p(H\text{错误}|\vec{X})} \cdot \left(\frac{1}{P} - 1\right) > \tau \text{ 等价于 } p(H\text{正确}|\vec{X}) > \frac{\tau \cdot P}{1 - P + \tau \cdot P}$$

这个公式显示出两个解决问题不同角度的内在联系。

## 2.4 说话验证的评价

统计假设检验在信号检测理论中早就得到广泛应用。因此, 说话验证的评价与信号检测的评价(Evaluation)在原理上是完全一致的。说话验证器作为假设检验器, 与信号检测器一样, 可以在不同的工作点上工作。因此, 评价其特性就要考虑所有的工作点的特性, 也就是工作点组成的曲线特性。考虑整个工作特性曲线的评价方法称为动态方法, 而只考虑曲线上特殊点的方法称为静态方法/参数。下面将介绍主要的评价方法。

### 2.4.1 评价方法

ROC(接受机工作特性, Receiver Operating Characteristics)曲线是指以第一类错误率  $P(I) = P(\text{拒绝}H_0 | H_0\text{真})$  为自变量画出的  $P(\text{拒绝}H_0 | H_0\text{假})$  (势) 变化曲线; 或者以  $P(II) = P(\text{接受}H_0 | H_0\text{假})$  为自变量画出的  $P(\text{接受}H_0 | H_0\text{真})$  变化曲线。如图 2-1, 以第一种方式给出了四条 ROC 曲线。其中的粗实线由随机接受 (Random Guessing) 的假设检验得到, 它的含意是: 随机接受时, 假设检验对零假设无论真假, 拒绝的可能性完全相同。粗虚线是理想的假设检验 (Perfect Testing) 的性能, 总是能 100% 地拒绝错误零假设。另外两条细线由两个性能不同的实际假设检验产生。由于它们对错误假设的拒绝率比对真确零假设的要高, 它们的性能比随机接受好, 但比理想情况又差。其中的实线对应的假设检验又比虚线的好, 因为当对真确

零假设拒绝率相同情况下, 它对错误零假设拒绝率更高。

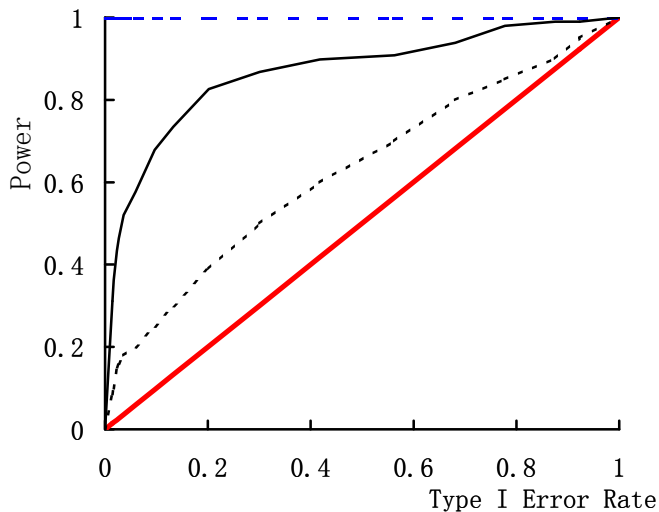


图 2-1 ROC Curves

FOM (Figure of Merit) 参数与 ROC 曲线密切相关。它是指在一定区间 ROC 曲线下的面积。它衡量假设检验方法在该区间中的整体性能。当考虑整个  $(0, 1)$  工作范围时, 随机拒绝的 FOM 为 0.5, 理想假设检验的 FOM 为 1, 而实际假设检验的 FOM 介于 0.5 到 1 之间。显而易见, 对于特定的工作范围, FOM 越大, 验证的

性能越好

DET (检测错误折衷, Detection Error Tradeoff) 曲线则是以  $P(I)$  为自变量画出  $P(II)$  的变化曲线, 并且取对数坐标的形式。由 DET 曲线可以确定另一个重要的性能指标, 等错点 (Equal Error Rate Point), 即  $P(I)$  与  $P(II)$  相等时的工作点。

## 2.4.2 比较基于不同性能识别系统的验证

说话验证研究总是在一定的语音识别系统上展开的。同样的识别系统可以采取不同的验证方法，而同样的验证方法可以用在不同的识别系统上。由于历史的原因，研究者总是在现有的识别系统上研究说话验证，因此导致识别系统往往不同。而要比较他们提出的说话验证方法，就需要有一种与识别系统性能无关的评价方法。

设随机变量  $A$  为零假设的性质（正确 1/错误 0），随机变量  $Z$  为假设检验的结果（接受 1/拒绝 0）。二者的互信息为

$$I(Z, A) = H(Z) - H(Z | A) = H(A) - H(A | Z)$$

$H(A)$  对应零假设性质的不确定度，反映假设检验的难度。

$$H(A) = -[P \log P + (1 - P) \log(1 - P)]$$

其中  $P$  为零假设（识别结果）正确的概率，即识别器的识别率（精度）；显然当  $p > 0.5$  时，识别器越精确，识别结果性质的不确定度就越小。

$H(A | Z)$  为加入验证和拒识后零假设性质的不确定度。

$$\begin{aligned} H(A | Z) &= -E \log P(A | Z) = -\sum_{a,z} P(a, z) \log P(a | z) = -\sum_{a,z} P(z | a) P(a) \log \frac{P(z | a) P(a)}{P(z)} \\ &= -\sum_{a,z} P(z | a) P(a) \log \frac{P(z | a) P(a)}{P(z | 1) P(1) + P(z | 0) P(0)} \\ &= -\sum_z \left[ P(z | 1) P \log \frac{P(z | 1) P}{P(z | 1) P + P(z | 0) (1 - P)} + P(z | 0) (1 - P) \log \frac{P(z | 0) (1 - P)}{P(z | 1) P + P(z | 0) (1 - P)} \right] \end{aligned}$$

其中， $P$  为识别器精度，而

$$P(0 | 0) = P(\text{拒绝 } H_0 | H_0 \text{ 假})$$

$$P(0 | 1) = P(\text{拒绝 } H_0 | H_0 \text{ 真}) = P(\text{I})$$

$$P(1 | 0) = P(\text{接受 } H_0 | H_0 \text{ 假}) = P(\text{II})$$

$$P(1 | 1) = P(\text{接受 } H_0 | H_0 \text{ 真})$$

它们的估计方法见本章的 2.1。

由于验证和拒识对零假设的性质做出了判断，零假设性质的不确定度应该减小。验证和拒识越可靠，减小得就越多。但是，减小的程度还与识别器自身的性能有关，也就是说，与验证和拒识任务的难度有关。当识别器性能差时， $H(A)$  大，一个简单的验证就使  $H(A | Z)$  比  $H(A)$  小得多。因此，采用互信息  $I(Z; A)$  来评价验

证严重依赖识别器的性能，即验证任务的难度。要减小这种依赖，一个简单的想法就是用任务的难度对验证取得的熵减小归一化，这样就得到归一化互信息 (Normalized Mutual Information)，也成为验证的效率 (Efficiency)。

$$E(Z; A) = \frac{I(Z; A)}{H(A)} = \frac{H(A) - H(A|Z)}{H(A)} = \frac{H(Z) - H(Z|A)}{H(A)}$$

其典型的曲线可以参见 (Williams and Renals, 1999)。

### 2.4.3 本论文的评价方式

本论文研究说话验证在两方面对识别系统的贡献：通过拒识提高系统对合法语音 (Within-Vocabulary Utterances) 的识别精度，即拒识误识 (Mis-recognition)；通过验证拒识非法声响 (Out of Vocabulary Sounds)。因此，需要从两个方面来衡量说话验证的性能。我们把零假设  $H_0$  错误的情况分为两类： $F_1$  指对合法语音的误识， $F_2$  指非法声响。拒识后系统对合法语音的识别精度 (Accuracy after Rejection) 为

$$AR = \overline{P}(H_0 \text{真} | \text{接受 } H_0, \overline{X} \text{合法}) = \frac{N(A, T)}{N(A, T) + N(A, F_1)} = \frac{\text{正确识别的语音数}}{\text{合法语音数}}$$

对非法声响的拒识性能可以用对非法声响的拒识率 (Rejection Rate) 来衡量：

$$RR = \overline{P}(\text{拒绝 } H_0 | \overline{X} \text{非法}) = \frac{N(R, F_2)}{N(R, F_2) + N(A, F_2)} = \frac{\text{拒绝的非法声响数}}{\text{非法声响数}}$$

$RR$  与验证的势  $\overline{P}(\text{拒绝 } H_0 | H_0 \text{假})$  的关系是：

$$\begin{aligned} RR &= \frac{N(R, F_1) + N(R, F_2)}{N(A, T) + N(R, F_1) + N(R, F_2)} \cdot \frac{N(R, F_2)}{N(R, F_1) + N(R, F_2)} \cdot \frac{N(A, T) + N(R, F_1) + N(R, F_2)}{N(R, F_2) + N(A, F_2)} \\ &= \overline{P}(\text{拒绝 } H_0 | H_0 \text{假}) \cdot \frac{N(R, F_2)}{N(R, F)} \cdot \frac{N}{N(F_2)} \cdot \frac{N(A, T) + N(R, F)}{N} \\ &= \overline{P}(\text{拒绝 } H_0 | H_0 \text{假}) \cdot \frac{N(R, F_2)}{N(R, F)} \cdot (1 - \overline{P}_e) / K \end{aligned}$$

其中， $\overline{P}_e$  为验证的 (无条件) 错误率， $K$  为非法语音在测试语音中所占的比例。不同的语音识别任务面对的非法声响在统计上也不尽相同。从研究验证方法对非法声响拒识的角度出发，我们更关心的是验证方法对各种可能遇到的非法声响的拒识能力，而不是验证方法对某个具体识别任务上面临非法声响的表现。因此，在本论

文的研究中，侧重于考察验证方法对不同性质非法声响的拒识能力。

采用  $AR$  和  $RR$  参数的优点是可以独立地考虑验证方法对误识和非法语音的拒识能力。 $AR$  只与合法语音相关，而  $RR$  只与非法语音相关。而且对于不同性质的非法语音，可以分别求出它们的  $RR$ 。

另一方面，说话验证必须考虑系统对使用者的友好程度 (Friendliness)。验证对合法语音的拒识率 ( $P(\text{拒绝}H_0 | \vec{X}\text{合法})$ ,  $RR$ ) 直接与使用者对系统友好程度的感觉相关。

$$\begin{aligned}\bar{P}(\text{拒绝}H_0 | \vec{X}\text{合法}) &= \frac{N(R,T) + N(R,F_1)}{N(T) + N(F_1)} = \frac{N(R,T) + N(R,F_1)}{N(A,T) + N(R,T) + N(A,F_1) + N(R,F_1)} \\ &= \frac{\text{被拒绝的合法语音数}}{\text{合法语音数}}\end{aligned}$$

对合法语音的拒识率与第一类错误率的关系如下

$$\bar{P}(\text{拒绝}H_0 | \vec{X}\text{合法}) = \frac{N(R,T) + N(R,F_1)}{N(T) + N(F_1)} = \frac{N(R,T) + N(R,F_1)}{N(R,T)} \cdot \frac{N(T)}{N(T) + N(F_1)} \cdot \bar{P}(I)$$

其中  $\frac{N(T)}{N(T) + N(F_1)}$  为识别器的识别率的估计  $\bar{P}$ 。

在本论文中，我们在大多数情况下将采用以  $P(\text{拒绝}H_0 | \vec{X}\text{合法})$  为自变量画出的  $RR$  和  $AR$  曲线来评价不同的验证方法。这样的做法与 (Mathan and Miclet, 1992) 相似。

图 2-2 画出了一个典型验证方法的合法语音的  $RR$  ( $RR$  of INV) 对合法语音的  $AR$  ( $AR$  of INV)，对非法声响的  $RR$  ( $RR$  of OOV) 的曲线。(INV 指 utterances withIN Vocabulary, OOV 指 utterances Out Of Vocabulary)。随着对合法语音的  $RR$  提高 (根本原因是验证的接受条件变得苛刻)，对合法语音的  $AR$  提高，对非法声响的  $RR$  也在提高。而且对不同的非法声响 (B 或 D)，验证的拒绝能力是不同的。

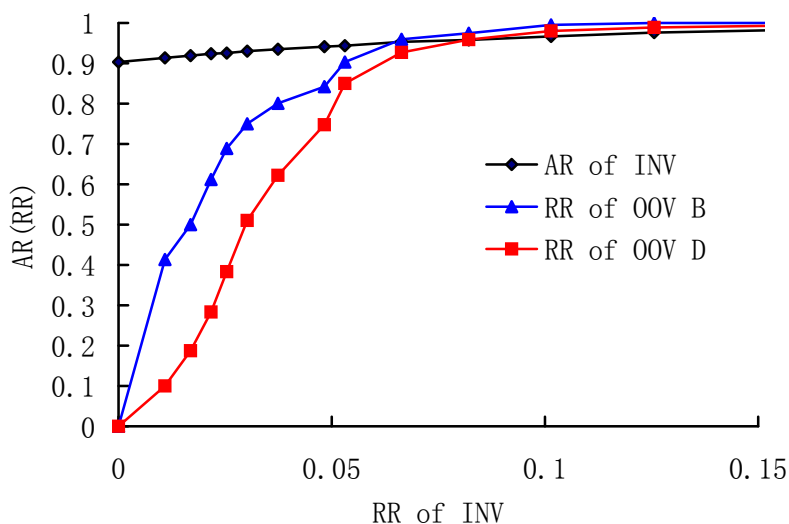


图 2-2

#### 2.4.4 评价样本的重要性

前面提到,不同的语音识别系统在实际使用中遇到的非法声响在统计上是不同的。当在研究词表/任务无关的说话验证的时候,就必须考虑说话验证对不同的非法声响的拒识能力。我们发现,一些方法可能对某些性质非法声响的拒识能力很高,但对另一些性质的拒识能力却不理想。因此,要全面评价一种验证方法,就必须对其多方面的拒识能力加以考察;另一方面,在为语音识别应用选取拒识方法时,也要根据识别在实际中遇到的非法声响在统计上的特点,有针对性的选取验证方法。

为了评价本论文研究的验证方法,根据实际的情况,我们采集了3种不同性质的非法声响库,其特点如表2-3,库的内容可以参见论文附录A。

表 2-3

库	性质	采样率	人数

B	使用者的短的随意应答	8k	20
C	说话噪声	8k	20
D	无关长句	8k	8

这三大类 OOV 声响在声学特点上分别代表了不同的三类非法声响，涵盖了大多数识别任务遇到的情况。对它们的拒识能力可以比较好地反映一般情况下，说话验证系统对 OOV 的拒绝性能。

## 2.5 论文研究采用的识别系统

本节将简要介绍三个语音识别系统。以下几章研究的说话验证都将在这三个系统上实现和评测。第一个系统是基于整词 HMM 与孤立语音识别的汉语数码语音识别系统。而后两个系统都是基于同一套子词模型 (Sub-word Models) 的。下面将先后介绍三个系统采用的声学模型，以及的结构和特点。

### 2.5.1 汉语数码语音识别系统

汉语数码语音识别具有广阔的应用前景，特别是汉语语音拨号，如果能够在低价专用集成电路(ASIC)上实现，将获得广泛的应用。由于语音拨号本身的特点，识别错误导致的拨号错误具有相对高的代价，因此以下两个方面的研究具有重要的意义。一方面是尽量提高识别精度，降低错误率;另一方面，争取能够区分正确和错误的识别结果，拒绝可能错误的识别结果也能降低错误率。这两方面的研究工作都在展开，前者可以参见(李虎生 刘润生, 2000)，而本章的工作就属于第二方面。

同英语数码语音识别不同，汉语数码语音识别有其特殊的问题(顾良, 刘润生 1997)。汉语数码识别很难达到英语数码识别的高精度。从高混淆的角度来说，汉语数码识别更接近英语字母表识别(English Alphabet Recognition)。针对英语字母表，已经有相当多的研究展开(Loizou and Spanias, 1996)。但是，绝大多数抗混淆的技术在计算上都是复杂的，不易于在 ASIC 上实现。因此，要将现有精度的汉

语数码识别器投入语音拨号实用，拒绝错误的识别结果有很重要的意义。加入拒识的语音拨号器的简单工作框图如图 2-3。

由于拒识器的存在，当拒识器认为识别结果不可靠时，就要求使用者将不可靠的数字再念一遍，从而避免误识导致的拨号错误。这样的机制既适用于孤立数码语音识别的拨号，也适用连续数码语音识别的拨号。

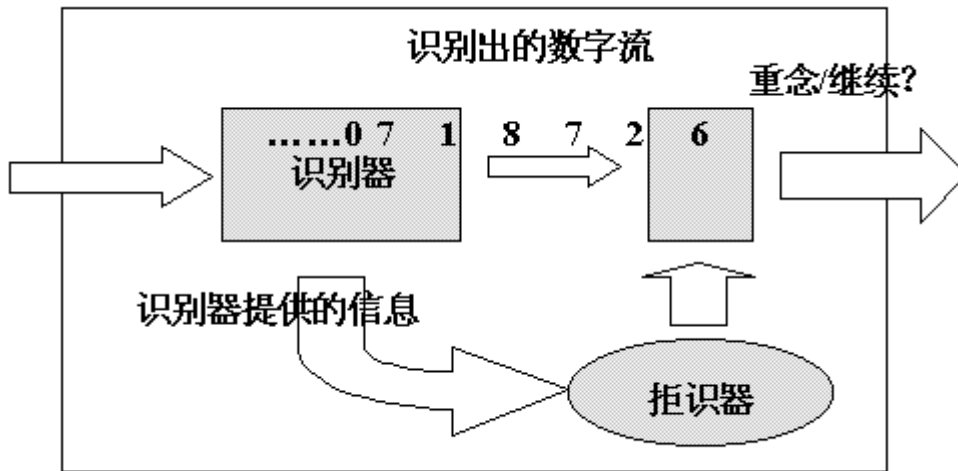


图 2-3

本论文采用基于连续 HMM 的孤立汉语数码语音识别器。每个数码由一个 6 状态 HMM 建模，每个状态的发射分布(Emission Distribution)用 7 个对角高斯分布的混合(7 diagonal Gaussian mixtures)来逼近。HMM 的拓扑结构为典型的无状态跨越和回跳自左往右结构，这里称为简单结构。识别用语音特征为 10LPCC，能量和它们的一阶差分。因此，一帧语音经过特征提取后成为 22 维特征矢量。为提高识别精度，在完成 HMM 的 MLE(Maximum Likelihood Estimation)训练后，采用 MCE (Minimum Classification Error)算法对模型进行进一步调整。

用来研究汉语孤立数码语音识别和拒识的语音库的情况如下表：

表 2-4

人 数	说话人特点	发 音 次 数 (0~9)	采样环境	采 样 率	量 化
80	四十岁以上和以下 的男性各 40 人	一次	普通办公室 环境	11, 025Hz	16bit 线性 量化

采用 Bootstrap 方法从统计上充分利用数据。每次实验，选取 10 人的语音对系统进行测试，用剩下 70 人的语音训练识别和验证的统计模型；下一次实验选取另外 10 人的语音测试，再用剩下 70 人的训练。这样一共进行 8 次实验，用 8 次实验的平均结果作为总的实验结果。

MLE 训练的识别系统识别率为 96.6%，经过 MCE 调整后识别率上升到 97.1%。这样的识别率对纯粹 HMM 的汉语数码语音识别来说已经是非常可观的（顾良，刘润生，1997）。另外，根据（李虎生等，2000）采用 MFCC 特征和两级识别，汉语数码孤立语音识别已经达到 98.8% 的识别率。但是求取 MFCC 特征的运算量更大，而且在 HMM 识别后引入第二级特殊的区分方法也使运算量大大增加，这样的识别系统目前还不易于在诸如 ASIC 这样的硬件上实现。因此，本论文选择上述利用 LPCC 的纯粹 HMM 识别系统研究拒识。

## 2.5.2 基于子词模型的识别系统

基于整词模型的汉语数码语音识别的验证只能是任务/词表相关的。因为不同的词表需要不同的底层模型（词模型）。而基于子词模型的识别，可以用同一组子词模型根据词法（灵活构建不同的词表，从而使任务/词表无关的说话验证成为可能。

### 2.5.2.1 语音特征与声学模型

后两个识别系统是在同样的声学底层上搭建起来的。汉语普通话为音节语言，

每个汉字在普通话中都发音为一个有调的音节。而每个普通话音节可以分割为无调的声母（包括为空的情况）和有调的韵母，在这里声母和韵母被统称为半音节，声母。根据（潘胜昔，1998），一共有 100 个声母，43 个无调韵母和 164 个有调韵母。子词模型实际上是用 HMM 为所有 100 个声母和 164 个有调韵母建模。对于无调韵母实际上是用对应的有调韵母模型来近似的。

声母和韵母 HMM 分别采用两状态和四状态的简单结构。由于历史的原因，每个 HMM 状态采用一个高斯分布估计发射概率分布（Emission Distribution），即

$$b(\vec{x} | \varpi_i, k) = \frac{1}{(2\pi)^{N/2} |\Sigma_{ik}|^{1/2}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu}_{ik})^T \Sigma_{ik}^{-1} (\vec{x} - \vec{\mu}_{ik})\right) \quad 2-1$$

其中为  $\vec{x}$  一帧语音特征向量（N 维）， $\varpi_i$  指示对应的半音节模型， $k$  指该模型的第  $k$  个状态；该状态对应的统计参数为  $\vec{\mu}_{ik}$ （均值）和  $\Sigma_{ik}$ （协方差矩阵）。通常的语音识别系统底层声学模型多采用多高斯混合（Gaussian Mixtures）估计状态发射概率分布，而其中的每个高斯分布的协方差矩阵为对角阵（非对角矩阵元素为零）。这样的高斯混合分布估计在性能很好而计算上又非常方便，所以得到广泛采用。而我们由于历史的原因，采用单高斯（协方差矩阵不限定为对角阵）估计发射概率分布。这样的模型在训练时有一定的优势（不需要 VQ），但在计算上稍微麻烦一些。训练语音为国家 863 大词表语音识别数据库的一部分（赵庆卫，1998）。对每个声母和有调韵母，都有一个 HMM。

采用有调韵母的 HMM 来近似对无调韵母建模。无调韵母 HMM 是一个虚拟的统计模型，和有调模型一样采用 4 状态简单结构，但是，一帧语音  $\vec{x}$  对应无调韵母  $\Omega_i$  第  $k$  状态的似然度为

$$p(\vec{x} | \Omega_i, k) = \max_{\varpi \in A_i} b(\vec{x} | \varpi, k) \quad 2-2$$

其中  $A_i$  表示  $\Omega_i$  对应的所有有调韵母组成的集合。可以发现这种虚拟 HMM 的状态发射概率分布估计与多高斯混合模型的相似之处。对多高斯混合模型，状态似然度为：

$$p(\vec{x} | \Omega_i, k) = \sum_{j=1}^M C_{ikj} b(\vec{x} | \varpi_{ikj}) \quad 2-3$$

其中  $\varpi_{ikj}$  为该模型第  $k$  状态对应的第  $j$  个高斯分布，其中  $C_{ikj}$  为该分布在混合模型

中占的权重,  $\sum_{j=1}^M C_{ikj} = 1$ ; 而  $b(\vec{x} | \vec{\omega}_{ikj})$  由公式(2-1)那样的高斯分布给出。比较公式(2-2)与公式(2-3), 可以认为(2-2)是(2-3)的一种特殊形式。由于虚拟无调模型在统计上更加精确 (有更多的参数, 更多的训练语音), 我们的识别系统采用虚拟无调韵母模型往往比采用有调模型识别率更高。但是, 付出的代价是运算量增加到原来的 4 倍左右 (一个无调韵母通常对应 4 个有调韵母)。

在 2.5.1 汉语数码语音识别中, 考虑降低语音特征提取的计算量, 采用了 LPCC 特征。但众所周知, 基于 MFCC 的识别通常具有更好的稳健性和精度, 这里子词模型的特征就采用 14 维 MFCC 及其一阶差分, 一维能量, 及其一阶和二阶差分, 为 31 维矢量 (江金涛, 1998)。

### 2.5.2.2 词表可更新说话人无关孤立语音识别系统

在许多中小词表语音识别应用中, 用户希望能够根据情况的变化灵活地更改识别系统的词表。有时候是识别的任务改变了, 比如某个机构的职员电话语音识别转接系统必须根据职员的增减更改词表; 有时候, 对同样的任务, 不同用户愿意设置不同的词表, 比如语音确认 (Voice Confirmation) 系统, 一些用户愿意说[正确/错误], 而另一些愿意选择[对/不对], 这样应该允许用户根据自己的习惯来设置词表。这些就是词表可更新的语音识别的应用背景。本论文工作以前一节论述的声学模型为基础构建了一个词表可更新的说话人无关的中小词表孤立语音识别系统。并以语音确认为应用背景研究了该系统上的说话验证。

在系统使用前, 用户需要输入任务的词表。系统自动将汉字转换为对应的拼音, 对每个词条生成对应的半音节词格 (Phoneme Lattice)。在识别中, 把输入语音与每个合法的半音节词格 Viterbi 对准, 得到对应的似然度得分。论文以语音确认为背景研究了该系统的说话验证。另一方面, 这个系统模块也被成功地用于智能娃娃的语音识别中。

对于许多语音识别 (包括关键词识别) 应用, 错误识别的代价非常高, 比如语音拨号和邮包的语音核对 (张昊天, 2000)。因此在识别之后加入语音确认非常重要。系统将识别结果回放或显示给使用者, 使用者在通过语音对识别结果进行确认。由于确认的词表非常小 (通常为 2), 而且可以选择为不易混淆的词条 (比如正确/错误), 识别率可以做到非常高。以[正确/错误]为例, 非特定人识别率在 99.3% 左

右。这样的识别要可靠得多。而另一方面，语音确认对非法语音的拒识就显得非常关键。如果无关的声响也能激活系统，那么也会造成重大的损失。与前面论述的汉语语音拨号系统和下面将要论述的电话语音识别系统不同，对这样的识别系统，识别率本身已经非常高，接近或达到了实用的要求，因此对非法声响的拒识就成为说话验证的主要任务。

本论文的验证研究主要在选取[正确/错误]作为语音确认用词的基础上进行。对于这样的语音确认系统应该选取确认词条使词条间的混淆最小，这正是本论文选取选择[正确/错误]而不是[对/不对]作为确认词条的原因。为这个系统采集了语音库A2，见表 2-5

表 2-5 合法语音库

语音库	内容	说话人数	采样率	其他
A1	电话语音 207 句	8	8k	通过电话线
A2	[正确/错误]	24	8k	模拟电话线低通滤波

下表给出了系统不同设置情况下的识别率：

表 2-6 识别系统的识别率 (%)

		有调韵母	虚拟无调韵母
语音确认 [正确/错误]		98.6	99.3
电话语音识别	前 4 人	84.7	90.3
	后 4 人	64.1	69.4
	总的	74.4	79.9

### 2.5.2.3 基于多子树结构的电话语音识别系统

本论文采用的电话语音识别系统已经被研究了多年(潘胜昔, 1998), (江金涛, 1998), (刘加 等, 1998)。这是一个接近于关键词识别的连续语音识别系统。通过对常用的电话转接用语的问卷调查, 得到 98 个基本句型, 代表了大多数常用的电话转接用语。用这些句型生成相应的基于规则的语言模型。在这个系统中, 规则语言模型用规则子树的形式实现。下图显示出一个句型(请接[地名]的[人名])对应的规则子树结构。

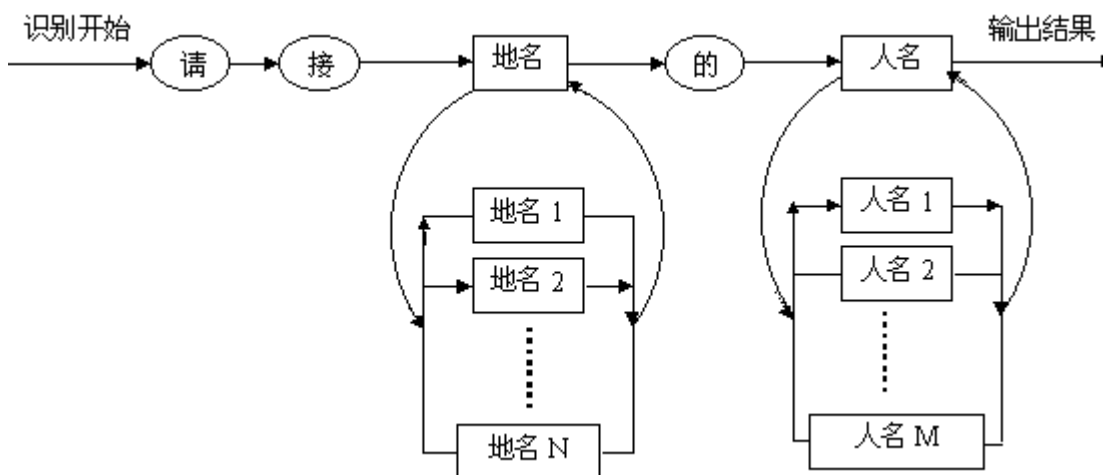


图 2-4 一个规则子树

规则子树中有对应着关键词的语意单元, 在电话语音识别系统中是, 地名, 人名以及电话号码。这些语意单元又分别由地名子树, 人名子树和号码子树的形式实现。例如, 地名子树给出所有地名关键词的词格。当识别搜索进入地名子树后, 就在地名子树规定的关键词中搜索。

采用语言模型和规则子树, 目的是用中小词表连续语音识别的方法实现关键词识别的任务。这样的电话语音识别系统对使用者的输入语音的限制以及识别的性能(精度和速度)介于普通语音识别和关键词识别之间。实际上, 只要将图中对应诸如[您好]/[请]/[找]的声学模型换成垃圾模型, 就实现了最基本的关键词识别系统。因此在此系统上研究的说话验证算法也在向关键词识别系统上移植时会更容易一些, 这是本论文选取它作为研究平台的重要原因。基于多子树结构的识别还有另

外一个好处，由于底层声学模型是与识别任务词表无关的（独立训练的），可以根据词法任意构建关键词的词格。因此，在关键词表发生变动时，只需更新相应子树的内容，而无需对整个识别系统进行改动。由于这个性质，这样的系统也非常适合用做研究任务/词表无关说话验证。这是本论文选取它作研究平台的另一个重要原因。

本论文采用（江金涛，1998）中的 8 人电话线语音数据库，在本论文中称为库 A1（见表 2-5）。这 8 个人中，有四个的语音录制条件与训练 863 语音数据库有较大差异，稳健语音处理（这里主要是指倒谱均值减）也没能较好地消除这种差异。这四人的语音识别率要比另外四个人的低得多。为了更有效地研究说话验证，孤立信道不匹配带来的特殊现象，下文中将单独研究这两组说话人语音的验证。

## 2.6 小结

本章论述了说话验证的数学原理和评价方法，分析了基于后验概率估计验证的物理意义及其同似然比假设检验之间的联系。从分别考虑误识和不同性质的非法声响的拒识出发，论述了本论文采用的评价手段，介绍了研究识别平台和语音数据库。特别强调了在研究对非法声响拒识能力时，考虑不同性质测试语音的重要性。

### 第三章 说话验证的信息源

在第二章 2.4.2 中, 提到以互信息  $I(A, Z) = H(A) - H(A|Z)$  来衡量验证使识别结果正确性的不确定度减小。那么验证过程必然需要额外的信息以减小这种不确定度。所谓额外是指识别过程没有利用或者没有充分利用的信息。本章将研究这样的信息源(Knowledge Sources), 主要是基于声学模型的信息源。其中的一部分将在后续章节中进一步研究。

论文语音识别系统在 HMM 的框架中实现, 对识别而言, 判决的根据是各个候选 (Candidate) 的似然度得分, 而且采用的是谁大取谁 (Winner-take-all) 的判决方式。要得到候选的似然度得分, 需要将输入语音与识别任务的声学模型和语言模型进行匹配。而对匹配的结果, 识别判决仅仅使用了最后的匹配得分, 而忽略了其他许多信息。另一方面, 识别任务的声学 and 语言学模型只对合法的语音是适用的, 也就是说, 它们只告诉我们正确的语音在统计上应该是什么样, 而没有关于非法声响的统计特征。另一方面, 它们也不能告诉我们, 如果合法语音被识别错误, 那么错误会有哪些统计特征。而这些统计特征对拒绝非法声响和误识是至关重要的。因此, 在考虑额外信息源时应该至少注意到两个方面, 1) 识别过程中忽略的信息; 2) 对误识和非法声响统计建模 (这两类信息源也并非泾渭分明, 例如在线垃圾模型就可以同时归到两个方面。)

### 3.1 识别过程中忽略的信息

识别过程中，语音与系统的声学模型（一般是 HMM）和语言模型对准（译码，Decoding），最后根据匹配的逐帧（Frame-wise）积累距离判断选取哪个匹配路径，哪个候选结果。而帧只是输入语音最基本单元，逐帧积累的匹配距离把各个语音帧独立并且等同考虑（Independently and Equally），实际上忽略了语音的结构信息（Structural Information）。因此在验证中，将尽量恢复和利用由译码产生的语音结构信息，从帧到 HMM 状态，从状态到 HMM（整词或半音节），在从半音节到音节，从音节到关键词，最后到整个输入语音。分层次地（Hierarchically）利用信息，这是贯穿本论文工作的一个基本思路。

#### HMM 迹 (Trace)

用 HMM 识别语音，由于其内在的 Viterbi 动态规划过程，把输入语音与 HMM 的某个合法的状态序列对应起来。对于通常采用的简单 HMM 结构（无状态跨越和回跳，自左往右），Viterbi 对准将语音分割为对应着各个状态的段，由此产生各个 HMM 的迹 (Trace) (Mathan and Miclet, 1992)，包括各状态分到的语音帧数，分到各状态语音段对应的平均值等。如图 3-1，一个 3 状态自左向右简单 HMM 及其产生的迹。

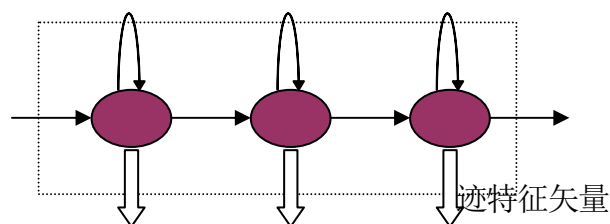


图 3-1

迹在状态的层次上考虑了语音的结构信息，而状态内的语音段求平均对语音模型特征的损失又不算太大，还把动态的语音映射成静态的模式，非常有利于采用一些常用的分布估计手段估计其概率分布。在第四章中将研究用人工神经网络估计迹特征的后验概率分布，并用后验概率进行数码语音识别的验证。

## 竞争模型的似然度得分

在语音与声学模型匹配的过程中，一帧语音需要和许多模型去匹配，然后选择将导致全局匹配距离最小的模型作为当前帧的模型。对识别来说，最后只考虑了最优的一条路径中模型与语音的匹配得分，而与这些模型竞争的那些模型的匹配得分（Scorings of Competing Models）被忽略。在线垃圾模型（On-line Garbage Model）(Boite et al, 1993)(Boulevard et al, 1994)利用这些信息来估计备选假设的似然度，在第 5 章中将研究这种方法，而且也将强调分层次地利用竞争模型得分的重要性。

## N-best 译码结果

语音与声学模型匹配后，往往可以给出不止一个候选结果（N）。识别一般选取匹配最好的结果作为输出。一种最简单且常用的验证方法就依据匹配结果前二选的差值（见第四章 4.3.1）。

现在，许多识别系统都能够给出前 N 选的词格（Word Lattices），词格不仅包含了前 N 优匹配距离，而且还包含了前 N 优匹配的对准信息，具有丰富的结构信息，许多研究证明前 N 选词格是最有效的验证信息源之一（Kemp and Schaaf, 1997）（Williams and Renals, 1999）。通过词格的对准信息，还可以得到识别结果在语音分割方式上的统计信息，有些可以被嵌入识别中，比如段长分布等。另一些则可以用来对识别结果进行后处理验证。例如 (Jouvet et al, 1999)根据识别假设中音节的个数来动态设置拒识门限。

对论文采用的基于多子树的电话语音识别系统，由于规则子树的限制，每次识别可以得到的合乎规则的结果数是不同。当输入为非法声响或者合法语音被错误对准的情况下，往往合乎规则的结果数很小，甚至为零。因此，可以采用一个简单的拒识准则利用这个信息：

如果合法结果数  $ResultNum < 2$ ，认为识别结果错误。

拒识的性能见表 3-1

表 3-1 验证的性能

拒识信息源	无	半音节长度	合乎规则结果数	同时使用
RR(%) of A1	0	3.1	0.6	3.7
RR(%) of B	0	67.4	82.7	95.4
RR(%) of C	0	52.4	73.8	88.1
RR(%) of D	0	45.3	0.6	45.7
AR (%) of A1	90.3	91.8	90.6	92.1

### 3.2 对误识和非法声响建模

#### 反词模型

如果词表为  $\{\varpi_j : j = 1, 2, \dots, C\}$ ，识别的结果为  $\varpi_i$ ，从 Neyman-Pearson 似然比假设检验出发，选择零假设和备选假设分别为：

$H_0$ ：输入语音是  $\varpi_i$ ；

$H_1$ ：输入语音不是  $\varpi_i$ 。

HMM 识别给出  $p(\vec{X} | \varpi_i)$ ，有  $p(\vec{X} | H_0) = \max_i p(\vec{X} | \varpi_i)$ ；因此一个很自然的想法就是直接用模型估计  $p(\vec{X} | H_1)$ 。如果对每个识别结果  $\varpi_i$ ，我们可以直接给出其备选假设的似然度  $p(\vec{X} | \overline{\varpi_i})$ ，那么可以用似然比

$$LR = \frac{p(\vec{X} | \overline{\varpi_i})}{p(\vec{X} | \varpi_i)}$$
 进行拒识。

反词模型 (Anti-word Model) (Rahim et al 1997) 正是试图直接用模型估计  $p(\vec{X} | H_1)$ 。所谓词 A 的反词模型就是用所有不是 A 的语音/声音训练的一个统计模型。和词模型一样，反词模型也可以采用 HMM。词模型给出  $p(\vec{X} | H_0)$ ，而反词模型给出  $p(\vec{X} | H_1) = p(\vec{X} | \overline{\varpi_i})$ 。

这样做的特点是对每个词或每个词聚类 (Word Cluster) 都有一个反词模型，一方面要求的存储量大；另一方面也使备选假设似然度估计可以做到更精确。如

如果在训练反词模型时加入非法声响， $H_1$ 也就涵盖了输入声响不合法的情况。反词模型对非法声响和误识都可以建模。如果不考虑非法声音，一般来说词 A 的反词模型就是用词表中其他词对应的训练样本训练出的模型。如图 3-2，输入样本空间（大椭圆）包括许多重叠的模式子空间（小椭圆），即不同词条所有可能的实现样本组成的子集；小椭圆之间的空隙对应着非法声音。A 的词模型对词条 A 对应的模式子空间建模，而 A 的反词就是对 A 在输入样本空间中的补集建模（对应图中 A 所处的小椭圆以外的部分）。如果不考虑非法声音，反词模型就是对 A 模式子空间以外的所有的模式子空间建模（对应图中所有有色的小椭圆部分）。（第四章采用的反词模型就属于这种情况。）

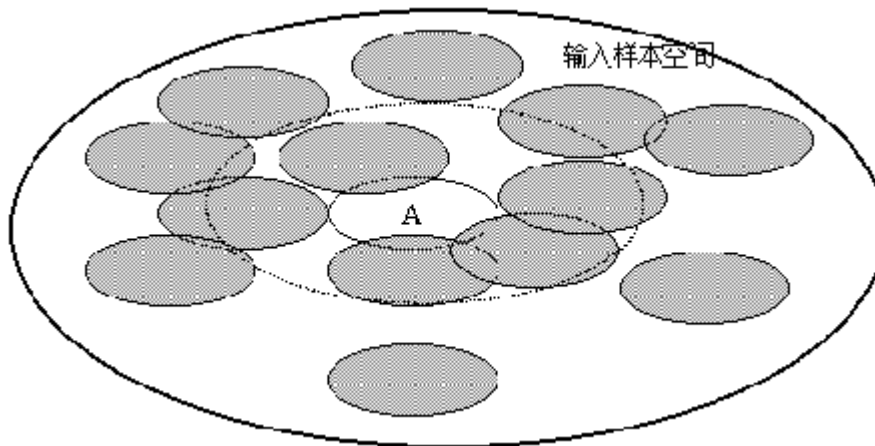


图 3-2

显而易见，对词 A 模式子空间的补集建模要比对其模式子空间建模要难。通常，一个模式子空间的补集要比该模式子空间复杂得多。为了降低反词建模的难度和减少反词模型的数量，可以根据混淆矩阵 (Rahim et al, 1997) 或采用模型聚类 (Model Clustering) 的方法，把词模型分为几个大的聚类，而后对每个聚类训练对应的反词模型 (反聚类模型, Anti-cluster Model)。如上图，虚线椭圆之内的模式子空间组成一个聚类，该聚类的反词模型就是对其外面的输入样本空间建模。如果不考虑非法声响，就是对虚线椭圆外的模式子空间建模，用对应聚类外的模式的语音来训练该聚类的反词模型。对与词表大或者基于子词模型 (Sub-word Models) 的识别系统，这样的做法是非常有效的。

在第四章中，将采用反词模型来对识别错误的汉语数码进行拒识。在第五章

中，将研究反半音节模型的词表无关的说话验证

## 垃圾模型

同样从 Neyman-Pearson 似然比假设检验出发，可以换一种方式来设置两个假设：

$H_0$ ：输入的是  $\varpi_i$ ；

$H_1$ ：输入的是非法声响，即不是来自  $\{\varpi_j : j = 1, 2, \dots, C\}$ 。

垃圾模型对应的正是这样的假设检验。零假设似然度用词模型来估计：

$$p(\vec{X} | H_0) = p(\vec{X} | \varpi_i)；$$

备选假设似然度就用垃圾模型（Garbage Model）来估计

$$p(\vec{X} | H_1) = p(\vec{X} | G)；$$

其中  $G$  表示垃圾模型。如果也采用 HMM 来对非法声响建模，同样可以得到  $p(\vec{X} | G)$ 。这样的 HMM 应该用非法声响训练得到。

垃圾模型早在基于 DTW 的语音识别流行时就已经被提出，当时叫垃圾模板（Filler Template）。到（Wilpon et al 1990）已经用 HMM 来对备选假设建模，因此称为垃圾模型或补白模型。在关键词识别和说话验证中，垃圾模型是一种非常流行方法(Bourlard et al, 1994)(Manos and Zue, 1997)。

在第五章中将研究垃圾模型和在线垃圾模型在说话验证中的应用。

## 3.3 声学统计信息

如果识别系统基于半音节/音位模型上，识别过程能够在输出词格中给出输入语音与各个半音节对准的信息。一个识别结果就是一个分配输入语音到各个半音节的假设方案。根据其分配，可以统计在该识别假设下半音节在声学上的统计特征。我们发现，在识别错误的结果中，这样的分配往往是病态的。也就是说，该识别假设下半音节的统计特征呈现异常。因此可以根据识别假设下的这些声学统计对识别结果进行验证。(Bartkova and Jouvét, 1997)采用高斯分布建立持续时间（Duration）模型，能量（Sound Energy）模型和浊化程度（Sound Voicing Degree）模型来估计

这三个音位参数的似然度，并将这些似然度用于估计置信度。在本论文中，仅仅使用更初等的方式利用一些简单的声学统计信息。

## 半音节长度

一般来说，汉语声母的持续时间短，而韵母的相对较长。可以统计语料库得到持续帧数分布。而对于非法语音和误识，半音节长度因病态分配出现异常。因此，可以根据词格计算识别结果中的最大声母长度  $Max_c$  和最小韵母长度  $Min_c$ ，

如果  $Max_c > 40$  或者  $Min_c < 10$ ，认为识别结果错误。

这样的拒识门限选得很宽松，主要是考虑到方法的普适性。对电话语音识别系统的拒识性能见表 3-1。

## 半音节和音节的长度方差

另一个重要的和半音节长度有关的参数是识别结果中半音节和音节的长度方差 (Variance)。计算方法如下，

1) 根据识别结果词格统计该识别结果中半音节/音节的均值：

$$M = \frac{1}{N} \sum_i Len_i$$

2) 计算方差：

$$Var = \sqrt{\frac{1}{N} \sum_i (Len_i - M)^2}$$

其中  $N$  为该识别结果中的半音节/音节个数。由于元音和辅音的声学特征具有较大差异，它们的长度方差是分别计算的。于是，可以得到三个参数  $Var_v$ ， $Var_c$  和  $Var_s$ ，分别对应声母，韵母和音节的长度方差。这样的参数适用于识别结果有较多半音节/音节的系统，诸如连续数码语音识别系统，电话语音识别系统等。因为只有在这种情况下这样计算出半音节/音节的长度方差才在统计上是可靠的。下面采用电话语音识别系统（见第二章 2.5.2）来研究这些参数对验证的作用。

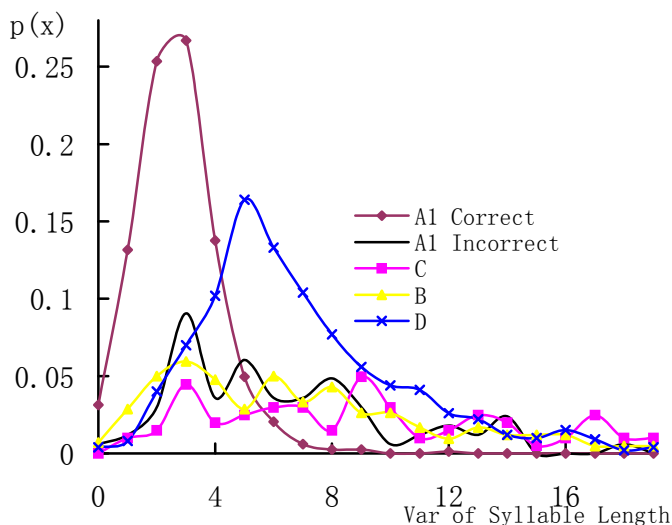


图 3-2

用语音库 A1, B, C 和 D(见第二章 2.5.2 和 2.4.4) 统计出  $Var_S$  对各种识别情况的分布如图 3-2 所示。正确识别 (A1 Correct) 的方差绝大多数集中在靠近坐标原点的地方, 而非法语音的却具有更宽的分布。显而易见, 如果选取域值  $\tau = 8$ , 当

$$Var_S > \tau \text{ 时}$$

认为识别结果错误, 可以拒绝掉相当多的识别错误, 而对正确识别结果几乎没有影响。改变拒识

门限  $\tau$  可以得到一系列工作点, 如图 3-3。同样对  $Var_V$  和  $Var_C$  也可以画出工作曲线, 如图 3-4 和 3-5。显然, 采用  $Var_S$  的拒识性能要好得多, 而采用  $Var_V$  的性能最差。这是因为音节的长度相对稳定, 而元音的程度变化最多。

另一方面, 注意到, 长的音节往往长度变化的绝对值大, 而且说话的语速会影响一句话的音节长度的统计特征, 特别是平均长度, 因此也就会影响长度方差的大小。语速慢的语音, 音节的长度均值  $M$  偏大, 而  $Var_V$  也往往偏大。因此, 本论文采用  $M$  对  $Var_S$  归一化, 得到  $Var'_S = \frac{Var_S}{M}$ 。用  $Var'_S$  进行拒识的性能见图 3-6, 其中带标记的曲线是  $Var'_S$  的, 无标记的是  $Var_S$  的。可以观察到采用  $Var'_S$  使性能得到的改善。

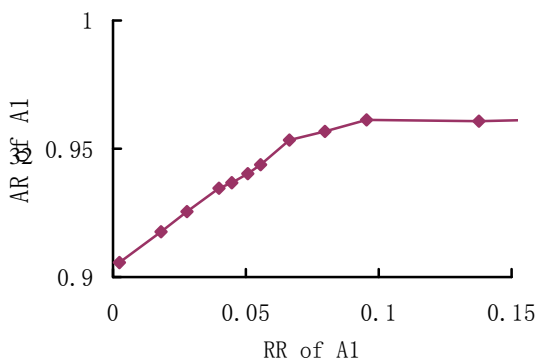
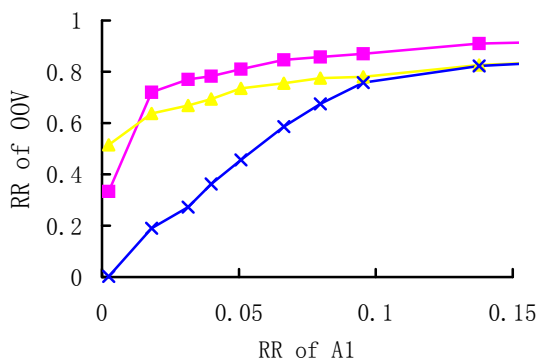


图 3-3 根据  $Var_S$  拒识

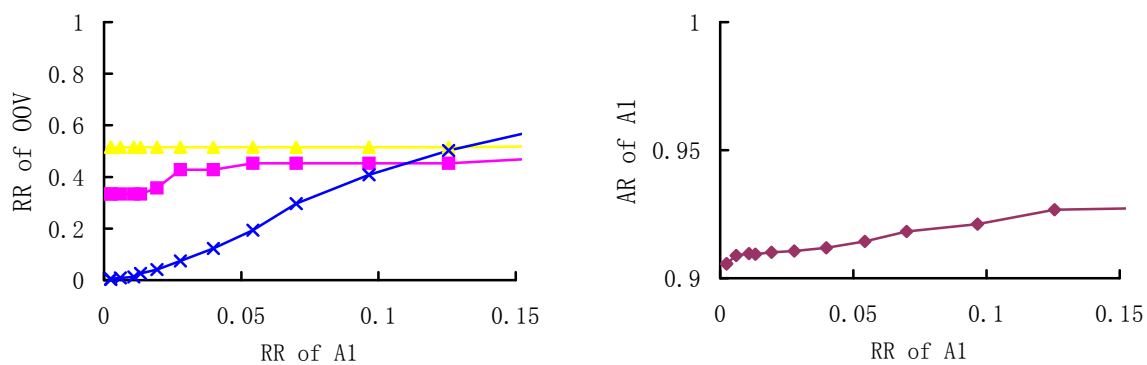


图 3-4 根据  $Var_V$  拒识

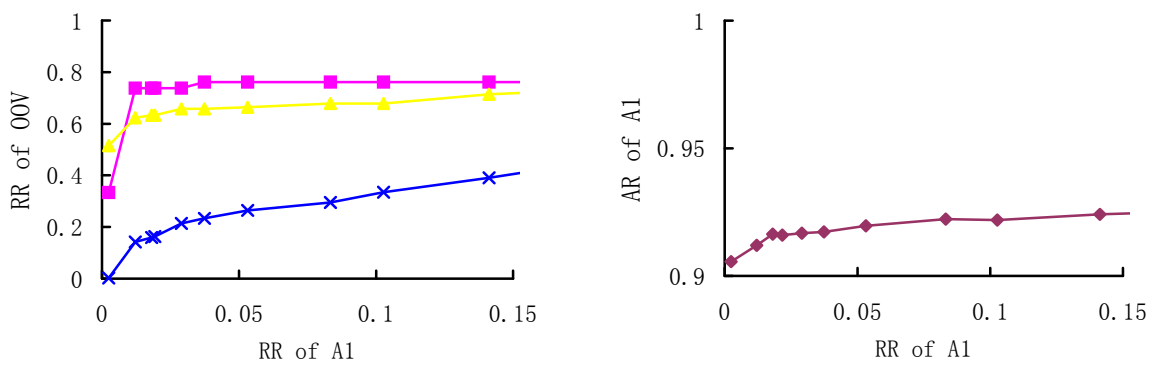


图 3-5 根据  $Var_C$  拒识

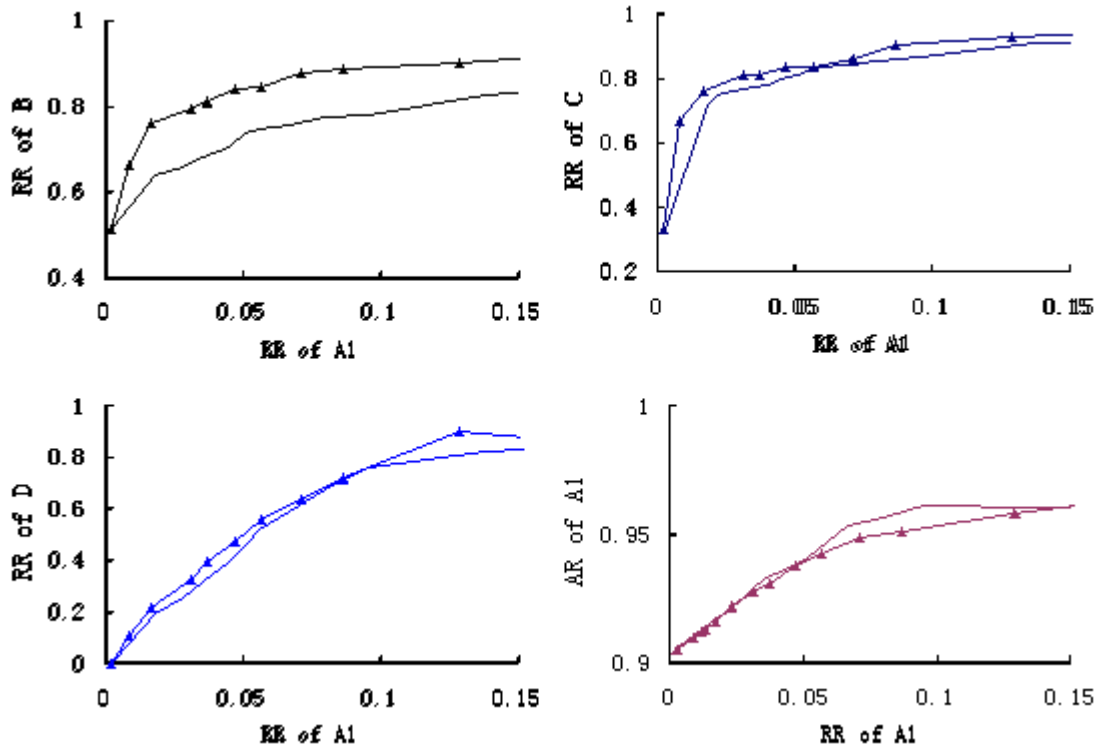


图 3-6

### 3.4 小结

本章论述了说话验证可资利用的信息源，提出了基于半音节长度和归一化音节长度方差的简单验证手段。这样的论述不是完备的 (Exhaustive)。针对特定的识别系统和任务，往往有特殊的信息源；而且对于大词表连续语音识别系统，语言模型 (Language Models) 也是非常重要的验证根据 (Schaaf and Kemp, 1997)。

本章论述的信息源及其基本方法大多数可以用到几乎所有的语音识别系统中，在后续章节中，将深入研究它们中的一些。第四章：基于 HMM 迹的后验概率估计的说话验证；第五章：基于垃圾模型和在线垃圾模型的说话验证；而第六章将研究如何将众多具有不同性质和不同量级的信息源加以综合利用提高验证性能。

## 第四章 基于MLP估计后验概率的拒识

由于流行的前向无反馈网络很难处理语音动态,神经网络多用于基于 HMM 的识别系统中,包括语音预处理/端点检测, HMM 状态发射分布估计,识别后处理等。本章基于多层前向感知机(Multi-layer Perceptrons, MLP)的拒识属于识别后处理的应用。本章以汉语数码语音识别为基本系统,研究了 MLP 估计后验概率在拒识中的应用。类似但又不完全相同的研究可以参考 (Mathan and Miclet, 1992) (Weintraub et al, 1997)。

### 4.1 MLP后验概率估计

MLP 指包括至少一个隐层的前向无反馈人工神经网络 (杨行峻 郑君里, 1992)。本文采用只有一个隐层的 MLP。

设 MLP(其参数集为 $\vec{W}$ )输入输出映射为:  $\vec{Y} = f(\vec{X}, \vec{W})$ 。每个训练样本 $\vec{T}$ 就是这样一对输入输出:  $\vec{T} = (\vec{X}, \vec{Y})$ 。 $\vec{X}$  为输入模式的一个实现,  $\vec{Y}$  为单位矢量, 它指示 $\vec{X}$  所属的模式类, 也是 $\vec{X}$  对应的理想网络输出。训练 MLP 就是寻找网络参数 $\vec{W}^0$ , 使得网络实际输出与理想输出间某种意义上的误差对一组训练样本来说最小。可以证明(Richard and Lippman, 1991), 当训练样本集统计上充分, 而网络具有足够的自由参数, 且训练没有进入局域最小点, 训练得到 MLP 的输出将是对输入 $\vec{X}$  类属的后验概率, 即

$$y_c = P(\vec{X} \text{来自类 } c | \vec{X}), \quad c=1, 2, \dots, C.$$

根据 (钟林, 1998), 选取误差函数为相对熵

$$E_{RE} = -\sum_{i=1}^M \sum_{j=1}^c \{d_{ij} \ln f_j(\vec{X}_i, \vec{W}) + (1-d_{ij}) \ln [1 - f_j(\vec{X}_i, \vec{W})]\}$$

其中 $M$  为训练样本数,  $C$  为语音类个数,  $d_{ij}$  样本 $\vec{X}_i$  对应的理想输出矢量的第 $j$ 个

元素。采用最陡梯度下降 (Steepest Gradient Descent) 调整网络参数, 时误差函数的值满足一定的要求 (学习停止条件)。如果采用最陡梯度下降, 参数是按如下进行迭代调整的:

$$\Delta w(t) = -\alpha \frac{\partial E}{\partial w(t)}; \quad w(t+1) = w(t) + \Delta w(t);$$

其中  $w(t)$  表示经过  $t$  步学习迭代后的网络参数  $w$  (包括连接权和神经元阈值),  $t$  是迭代步数;  $\alpha$  称为学习因子, 是大于 0 的实数。研究表明, 在许多情况下, 再加上一个惯性项有利于提高, 学习速度, 即:

$$\Delta w(k) = -\alpha \frac{\partial E}{\partial w(k)} + \eta \Delta w(k-1);$$

采用误差后向传播 (Error Back-propagation), 对输出层神经元连接权  $w_{ij}^{(2)}$  (可以将阈值看作一个输入恒定为 1 的连接权) 的调整, 要计算:

$$\frac{\partial E}{\partial w_{ij}^{(2)}} = \sum_{t=1}^M \frac{\partial E}{\partial I_{ii}^{(2)}} \frac{\partial I_{ii}^{(2)}}{\partial w_{ij}^{(2)}} = \sum_{t=1}^M \frac{\partial E}{\partial I_{ii}^{(2)}} o_{ij}^{(1)} = -\sum_{t=1}^M \delta_{ii}^{(2)} o_{ij}^{(1)}$$

$$\text{其中 } \delta_{ii}^{(2)} = -\frac{\partial E}{\partial I_{ii}^{(2)}} = -\frac{\partial E}{\partial o_{ii}^{(2)}} \frac{\partial o_{ii}^{(2)}}{\partial I_{ii}^{(2)}} = -\frac{\partial E}{\partial o_{ii}^{(2)}} o_{ii}^{(2)} (1 - o_{ii}^{(2)});$$

其中,  $o_{ik}^{(1)}$  和  $o_{ik}^{(2)}$  分别表示输入为第  $t$  个学习样本时隐层和输出层第  $k$  个神经元的输出 ( $o_{ik}^{(2)}$  就是式 4.1 和 4.2 中的  $f_k(\vec{X}_t, \vec{W})$ ),  $I_{ii}^{(2)}$  表示输出层第  $i$  个神经元的输入。对输入层到隐层的连接权的由误差反向传递,

$$\frac{\partial E}{\partial w_{ij}^{(1)}} = \sum_{t=1}^M \frac{\partial E}{\partial I_{ii}^{(1)}} \frac{\partial I_{ii}^{(1)}}{\partial w_{ij}^{(1)}} = \sum_{t=1}^M \frac{\partial E}{\partial I_{ii}^{(1)}} I_{ij}^{(0)} = -\sum_{t=1}^M \delta_{ii}^{(1)} I_{ij}^{(0)}$$

其中

$$\delta_{ii}^{(1)} = -\frac{\partial E}{\partial I_{ii}^{(1)}} = -\frac{\partial E}{\partial o_{ii}^{(1)}} \frac{\partial o_{ii}^{(1)}}{\partial I_{ii}^{(1)}} = -\sum_{l=1}^H \frac{\partial E}{\partial I_{il}^{(2)}} \frac{\partial I_{il}^{(2)}}{\partial o_{ii}^{(1)}} o_{ii}^{(1)} (1 - o_{ii}^{(1)}) = \sum_{l=1}^H \delta_{il}^{(2)} w_{li} o_{ii}^{(1)} (1 - o_{ii}^{(1)})$$

上述的最陡梯度下降算法是以网络对所有样本的总误差作为误差函数进行梯度下降的迭代, 对所有的样本计算过一次调整量后对网络参数进行更新。这样的做

法，一方面增加系统的存储要求，另一方面实验证明更容易进入局域最小点 (Local Minimum)。另一种算法是随机梯度下降(Stochastic Gradient Descent)，也就是每学习一个样本就对网络参数进行更新。对每个样本，可以有

$$E_t^{RE} = \sum_{j=1}^c \{d_{ij} \ln f_j(\vec{X}_t, \vec{W}) + (1-d_{ij}) \ln [1 - f_j(\vec{X}_t, \vec{W})]\}$$

于是有  $\frac{\partial E_t^{RE}}{\partial w_{ij}^{(2)}} = -\delta_i^{(2)} o_{ij}^{(1)}$ ， $\frac{\partial E_t^{RE}}{\partial w_{ij}^{(1)}} = -\delta_i^{(1)} I_{ij}^{(0)}$  从而对网络参数更新。由于在样本

学习的顺序上可以引进随机性，使网络易于摆脱局域最小点。因此，本论文的 MLP 都采用随机梯度下降-误差反向传播算法。

根据 (钟林, 1998)，在终止 MLP 学习时采用平方误差：

$$E_{SE} = \frac{1}{2} \sum_{i=1}^M \sum_{j=1}^c [d_{ij} - f_j(\vec{X}_i, \vec{W})]^2$$

当  $|E_{SE}(\vec{W}(t+1)) - E_{SE}(\vec{W}(t))| \leq \varepsilon$ ，结束训练 ( $\varepsilon$  是一个事先确定的正实数)。

## 4.2 基于后验概率的验证

### 4.2.1 估计数码语音后验概率

在用于语音识别时，MLP 最大的问题在于它缺少动态时间归整 (Dynamic Time Warping) 机制，不善于在语音持续的时间变化中抓住不变的特征 (钟林, 1998)。因此，要将 MLP 用于估计语音后验概率，最好先将语音变成静态模式。我们选择 HMM 的迹作为语音的静态模式 (见第三章)。

根据第三章，对于汉语数码语音识别采用的 6 状态简单结构 HMM，Viterbi 对准后，产生 6 个特征矢量。每个特征矢量包括 1) 该状态分到的语音帧数在语音总长中所占比例；2) 该状态分到的语音帧各个 LPCC 和能量的平均；3) 该状态分到的语音帧各个 LPCC 和能量一阶差分的平均。在下面将介绍减少特征的方法。

对某个输入语音  $\vec{X}$ ，10 个数码 HMM 都会生成对应的迹  $\vec{T}_i$ 。本论文采用这 10

个迹组成的特征矢量 $\vec{Y}$ 作为输入语音对 MLP 的静态输入。如图 4-1。

训练估计后验概率的 MLP 的样本与训练估计似然度的 HMM 的样本相同。对训练样本集，训练后的 HMM 并没有达到 100% 正确，因此对其中某些样本的识别是错误的。但是，这些被识别错误的样本仍然被正确标注并用来训练 MLP。因此，在训练 MLP 的样本中，既有被 HMM 很好建模的样本，也有一些 HMM 没有学好的 Outliers。这样，估计后验概率的 MLP 不仅学习了成功的识别例子，而且学习了错误的识别例子。这正是 MLP 估计的后验概率可以用来可靠拒绝识别错误的原因之一。

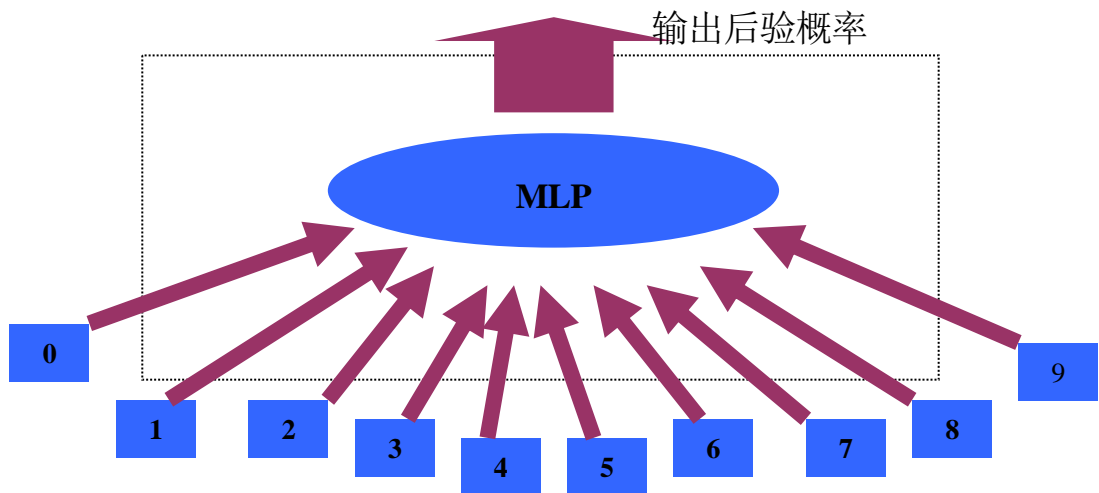


图 4-1

#### 4.2.2 基于后验概率的验证

训练后的 MLP 对输入样本 $\vec{X}$ 给出后验概率 $P(\varpi_j | \vec{X})$ ；设根据最大似然度准则 HMM 识别样本为语音 $\varpi_i$ ，于是识别正确的后验概率

$$P(H\text{正确} | \vec{X}) = P(\varpi_i | \vec{X})$$

根据第二章 2.2 对基于贝页斯决策验证的分析，当

$$P(H\text{正确} | \vec{X}) = P(\varpi_i | \vec{X}) > \tau \text{ 时，接受识别结果 } H$$

当第二类错误（错误接受）的代价相对第一类错误（错误拒绝）的越大， $\tau$  就越大，接受的条件也就越苛刻。由于  $\tau$  不同而工作在不同工作点的验证对应着不同的两类错误代价比。可以给出公式计算得到后验概率分布之后，由于额外的信息源，HMM 迹  $\vec{X}$ ，识别结果正确性的不确定性  $H(A)$  减小的量：

$$I(A, \vec{X}) = H(A) - H(A | \vec{X});$$

其中

$$H(A | \vec{X}) = \sum_a \int p(\vec{X}) p(a | \vec{X}) \log p(a | \vec{X}) d\vec{X} = \sum_j \int p(\vec{X}) p(\varpi_j | \vec{X}) \log p(\varpi_j | \vec{X}) d\vec{X}$$

### 4.2.3 优化存储和运算量

对于基于 MLP 后验概率估计的验证，存储和运算量主要由 MLP 的自由参数个数，特别是网络的连接权的个数决定。网络的连接权可以估计为：

$$N = I \cdot H + H \cdot O;$$

其中  $I$  为 MLP 的输入个数， $H$  为隐层神经元数， $O$  为输出数。由于  $O$  对应语音类的个数，可以改变的就只有  $I$  和  $H$ 。

#### ✓ 隐层大小影响

MLP 隐层神经元个数  $H$  总是影响 MLP 估计/逼近能力的重要因素。在一定程度上，增加隐层神经元个数可以提高 MLP 的估计/逼近能力；但是，一方面随着隐层神经元数的增加，运算和存储的要求也增加，另一方面，过多的网络参数又可能使 MLP 出现过训练。调整 MLP 的隐层神经元数，使之达到性能和运算/存储代价的折中是重要的。图 4-2 的拒识采用 HMM 整个迹作为拒识特征，可以看出 MLP24（有 24 个隐层神经元）的性能略胜过 MLP12（有 12 个隐层神经元）。

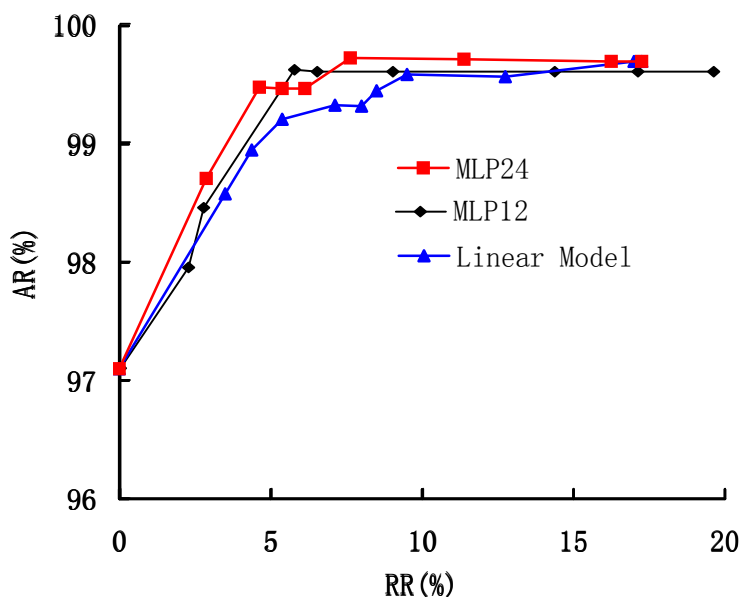


图 4-2

### ✓ 特征优化

对于输入采用的 HMM 迹，包括 3 部分，图 4-3 显示出各部分对拒识性能的贡献。图中的 MLP 具有 24 个隐层神经元，曲线 1 采用所有的迹特征（23 个/状态）；曲线 2 采用了迹特征的前两部分（12 个/状态）。也就是去掉了 LPCC 和能量差分项的均值；曲线 3 仅采用了第一部分即各状态分到的语音占总长的比例（1 个/状态）。可见，去掉 LPCC 和能量差分项对拒识性能并无明显影响，而另一方面又将输入特征数减小了几乎一半。因此，当对运算和存储要求的限制严格时，可以考虑在验证中去掉差分项，这里把这样的迹称为减迹（Reduced Trace）。

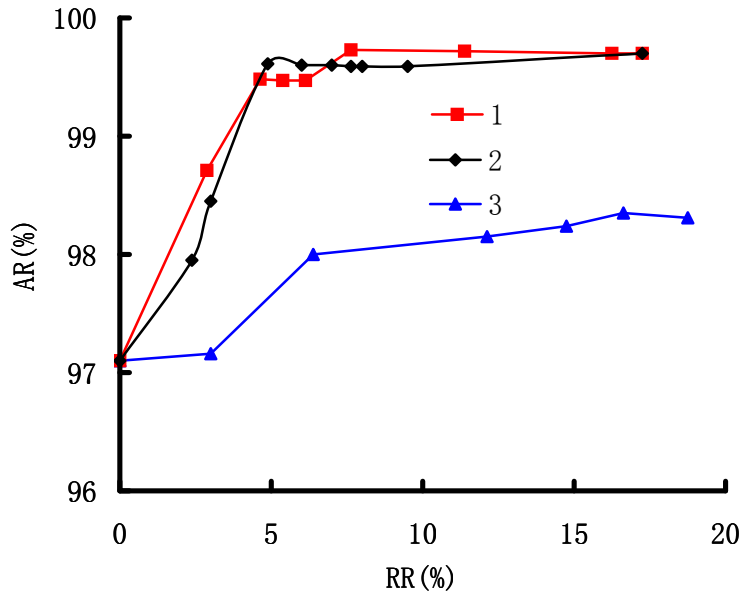


图 4-3

### 4.3 其他拒识方法

为了比较 MLP 拒识的性能，论文也实验了其他的常用拒识方法。

#### 4.3.1 前二选(Two Best)

对于基于 HMM 的识别系统，每次识别都给出输入语音对每个词条 HMM 的似然度得分。一种很简单且流行的做法是采用最高的两个得分进行拒识。由于识别已经选取似然度得分最大的模型  $\varpi_i$ ，因此当

$$LR = \frac{p(\vec{X} | \varpi_i)}{\max_{j \neq i} p(\vec{X} | \varpi_j)} > \tau \text{ 时，接受识别结果。}$$

比较第一章中提到的 Neyman-Pearson 似然比假设检验，

$$LR = \frac{p(\vec{X} | H_0)}{p(\vec{X} | H_1)} > \tau$$

由于识别结果为 $\varpi_i$ ， $p(\vec{X} | \varpi_i) = p(\vec{X} | H_0)$ 。因此，可以将 $\max_{j \neq i} p(\vec{X} | \varpi_j)$ 视为对 $p(\vec{X} | H_1)$ 的估计：

$$p(\vec{X} | H_1) = \max_{j \neq i} p(\vec{X} | \varpi_j) \quad 4-1$$

假设 $\vec{X}$ 都是合法语音（在本研究中为汉语数码语音， $\vec{X} = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_N\}$ ），于是有 $p(H_1 | \vec{X}) = \sum_{j \neq i} p(\varpi_j | \vec{X})$ ，因此

$$\begin{aligned} p(\vec{X} | H_1) &= \frac{p(H_1 | \vec{X})p(\vec{X})}{P(H_1)} = \frac{\left[ \sum_{j \neq i} p(\varpi_j | \vec{X}) \right] p(\vec{X})}{\sum_{j \neq i} P(\varpi_j)} \\ &= \frac{\left[ \sum_{j \neq i} \frac{p(\vec{X} | \varpi_j)P(\varpi_j)}{p(\vec{X})} \right] p(\vec{X})}{\sum_{j \neq i} P(\varpi_j)} = \frac{\sum_{j \neq i} p(\vec{X} | \varpi_j)P(\varpi_j)}{\sum_{j \neq i} P(\varpi_j)} \end{aligned}$$

再假设先验概率 $P(\varpi_j)$ 对所有类相同，则有

$$p(\vec{X} | H_1) = \frac{\sum_{j \neq i} p(\vec{X} | \varpi_j)}{N-1} \quad 4-2$$

对汉语数码语音识别， $N=10$ ；公式 4-2 正是第三章要研究的在线垃圾模型的形式。但是，对 HMM 识别系统，训练样本很难做到充分，因此假设

$$p(H_1 | \vec{X}) = \sum_{j \neq i} p(\varpi_j | \vec{X}) \text{ 不能成立}$$

另一方面，关于先验概率相等的假设也是不成立的。因此式 4-1 导致的验证并不能在 Neyman-Pearson 意义上做到最优。通过对识别错误的观察，我们发现识别错误总是发生在那些非常有个性化的样本上，也就是那些没有被训练样本在统计上体现的样本。而对于公式 4-1 的求和平均的形式，那些匹配特别差的 HMM 的似然度往往左右了最后的结果，反而使真正反映精细匹配的几个 HMM 的似然度提供的信息被模糊。可以采取两种方法来克服这个缺陷。

一，用几何平均代替算术平均：

$$p(\vec{X} | H_1) = \left\{ \frac{\sum_{j \neq i} \exp\{\gamma \cdot \log[p(\vec{X} | \varpi_j)]\}}{N-1} \right\}^\gamma \quad 4-3$$

当  $\gamma=1$  时，公式 (4-3) 等价于 (4-2)； $\gamma$  越大，匹配好的模型对结果的贡献越大；当  $\gamma \rightarrow \infty$  时，公式 (4-3) 等价于 (4-1)。

二，选取前  $M$  选估计：

$$p(\vec{X} | H_1) = \frac{\sum_{j \neq i} p(\vec{X} | \varpi_j)}{M-1} \quad 4-4$$

当  $M=N$  时，公式 (4-4) 等价于 (4-2)；当  $M=2$  时，等价于 (4-1)。

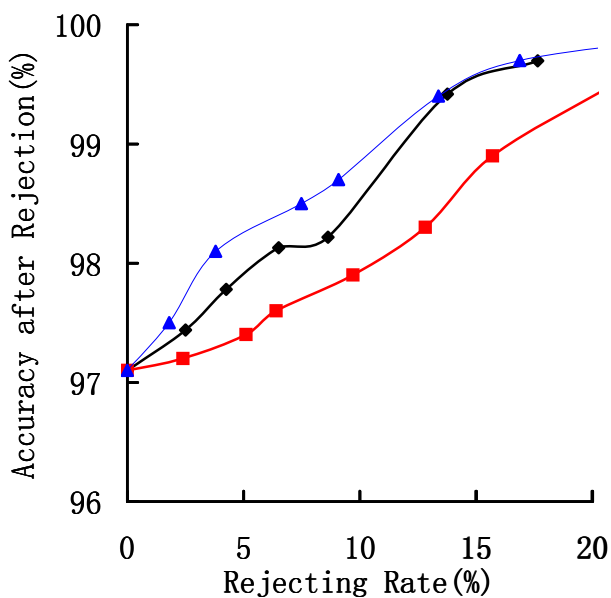


图 4-4

实验结果如图 4-4，曲线 1 对应公式 (4-2)，曲线 2 对应 (4-1)，曲线 3 对应 (4-3)。可以看出，用前二选拒识的性能介于几何平均和算术平均的在线垃圾模型之间。

这种方法的优点是几乎不需要额外的计算和存储量。但它的性能却不太令人满意。

### 4.3.2 反词模型 (Anti-word Models)

对于本章研究的数码语音识别任务，词表很小，因此采用最简单的反词模型，也就是对每个数码模型  $\omega_i$  训练对应的反词模型  $\bar{\omega}_i$ 。反词模型也采用 HMM，其结构和对应词模型完全一样。

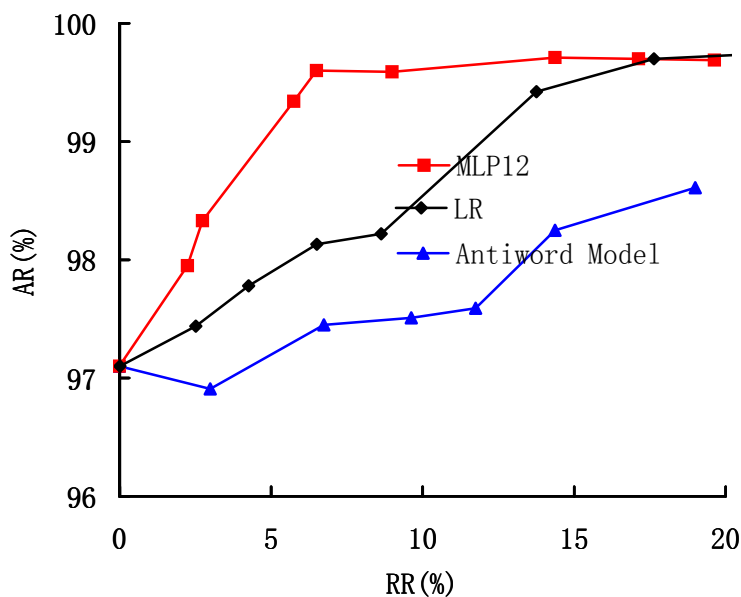


图 4-5

基于 MLP(12 个隐层神经元) 后验概率估计(采用迹), 反词模型和似然比的拒识性能如图 4-5 所示。显然, MLP12 的拒识性能远远超过后两种方法。但是, 在英语数码语音识别拒识中性能优异的反词模型对汉语数码的性能反而比不上

简单的根据前二选得分的拒识 (LR 曲线)。究其原因, 应是汉语数码语音识别中误识, 也就是混淆 (Confusion) 的不对称 (Asymmetries) 造成的。在汉语数码语音识别错误中有两种不对称: 第一, 识别错误主要发生在几个易混数码对内, 其他的数码相比之下, 很少发生错误; 第二, 即使在某些易混数码对内, 总是一个数码被误识成另一个, 而另一个数码却很少被误识为前者。由于这些不对称的存在, 用所有别的数码语音训练的简单反词模型估计的备选假设似然度远比词模型估计的零假设似然度要粗糙得多。因此, 这样的反词模型对语音验证的用处不大。要改进性能必须改变这种简单的训练方法。

### 4.3.3 线性模型

与 MLP 相对应，还有一种常用的统计模型：线性模型（Linear Model）

$$\vec{Y} = \vec{W} \cdot \vec{X} + \vec{b}$$

如果给线性模型的输出加上 Sigmoid 函数限幅使其输出在 (0, 1) 之间：

$$O_i = \frac{1}{1 + \exp(-y_i)}$$

就成为感知机（Perceptron）的形式，其训练算法与 MLP 的在本质上完全一样。本论文采用的是随机梯度下降算法。与 MLP 相比，感知机的估计和逼近能力要差，但是其模型简单，要求的运算和存储都比 MLP 小，所以也不失为一种折中的方案。其拒识性能与 MLP 的比较参见图 4-3。与 MLP 相比，线性区分的性能稍微差一些。这主要是线性模型在估计与逼近上的局限性造成的。

### 4.3.4 性能，运算和存储

在拒识性能上，采用减迹特征的 MLP24 在拒绝 4.9% 的输入数码语音之后，把正确率从 97.1% 提高到了 99.6%。这样的性能是非常优异的。

除了拒识性能外，上述几种方法对运算和存储的要求有显著的差异，如下表。如果要在诸如 ASIC 的低成本硬件上实现汉语语音拨号，性能和资源要求之间的折中非常重要。

表 4-1

拒识方法	存储量（自由参数个数）	运算量
MLP12/减迹	8,782	小
MLP24/减迹	17,554	较大
线性区分/减迹	7,210	小
反词模型	18,900	大且与语音长度成正比
似然比	0	可忽略

## 4.4 小结

本章提出采用 MLP 估计的后验概率对汉语孤立数码语音识别结果进行验证拒识。在拒绝 4.9% 的数码的同时，将单数码识别率从 97.1% 提高到 99.6%，这样的性能远远超过常用的前二选似然比和反词模型验证拒识。

用 HMM 迹和 MLP 估计后验概率的方法，可以用到连续数码语音的验证拒识上。但有一点需要注意的是，在孤立数码识别中，所有的误识都是替换错误 (Substitutions)。而连续数码识别误识的情况要复杂得多，包括：替换错误 (Substitutions)，删除错误 (Deletions) 和插入错误 (Insertions) (李虎生, 2000)。在后两种情况下，一个数码错误往往意味着与其相临的数码的分割也有问题，因此不能单独考虑数码串中某一个数码的声学特征而估计其置信度，还需要利用其上下文的声学特征。

本章没有考虑对非法声响的拒识。如果把非法声响看作是一类或几类模式，并用 HMM 对其建模，就可以把对非法声响的拒识按图 4-2 所示方式纳入到 MLP 估计后验概率的方法中。

另一方面，图 4-2 所示的后验概率估计方式不能推广解决大词表或词表无关的识别系统拒识问题。如果将 MLP 与反词模型相结合，对汉语语音识别可以得到一种基于半音节模型 (第二章 2.5.2) 的解决方案。假设系统将语音段  $\bar{X}$  识别为半音节  $i$ ，可以根据其对应半音节模型  $\omega_i$  和对应反半音节模型  $\bar{\omega}_i$  的迹，估计识别正确性的后验概率。这是论文的后续工作之一。

## 第五章 垃圾模型，在线垃圾模型及其性能优化

垃圾模型是说话验证中常用的方法。在线垃圾模型在概念上与之相关。在研究说话验证时，在线垃圾模型通常被选作基准方法。本章将以电话语音识别系统为平台研究基于垃圾模型与在线垃圾模型的说话验证，并将给出相应方法在语音确认系统上的性能。由于本章的验证方法基于半音节模型之上，验证的方法是词表/任务无关的。垃圾模型和在线垃圾模型也是关键词识别中最常用的方法，本论文工作虽然主要集中在验证上，但希望关于这些方法的研究和结论也会使关键词识别研究受益。

### 5.1 垃圾模型

#### 5.1.1 有回跳的HMM结构

(Bourlard et al, 1994) (Manos and Zue, 1997)比较了许多基于 HMM 垃圾模型的可能性。非法声响与语音相比，声学特征更加复杂，简单 HMM 结构并不能胜任。第一，由于非法声响的长度变化，因此，其 HMM 结构中应该有回跳(Back Jumping)。第二，许多非法声响，特别是非语音的短时变化剧烈，因此，HMM 结构中应该允许更多的跳转。有几种直观上很合理的 HMM 结构。结构 I (如图 5-1 左)通常用于对背景噪声建模；结构 II (如图 5-1 右)允许更多的跳转。采用训练系统半音节 HMM 的 863 语音数据库的一小部分 (16 个说话人，每人 45 句话)。两种结构各有 3 个状态，每个状态用 3 个高斯混合模型估计发射概率分布。



图 5-1 垃圾 HMM 的结构

### 5.1.2 高斯混合模型

极端的情况是基于高斯混合模型（Gaussian Mixture Model, GMM）和最小距离匹配的。这相当于可以从任何状态进入/离开的全连接 HMM 结构。通过聚类得到的每个 GMM 相当于 HMM 状态。对一帧语音  $\vec{X}$  与 GMM  $\varpi_i$  的距离定义为该帧语音对  $\varpi_i$  的似然度：

$$D(\vec{X}, \varpi_i) = p(\vec{X} | \varpi_i) = \sum_{j=1}^M C_{ij} N(\vec{X}, \vec{\mu}_{ij}, \Sigma_{ij})$$

其中  $\vec{\mu}_{ij}$  与  $\Sigma_{ij}$  分别为第  $j$  个高斯分量的均值与协方差矩阵。Viterbi 对准相当于为输入语音矢量序列找出最小距离匹配的 GMM 序列，路径的积累似然度对应路径的积累匹配距离。训练码字时采用普通的 K 均值聚类算法（杨行峻 迟惠生，1995），由于距离选取似然度的形式，实际上是对训练语音的最大似然度学习。为与有回跳的 HMM 垃圾模型比较，本论文采用 3 个 GMM 来对输入语音帧建模。每个 GMM 与有回跳的 HMM 状态一样，用 3 个高斯混合模型来估计，如图 5-2。

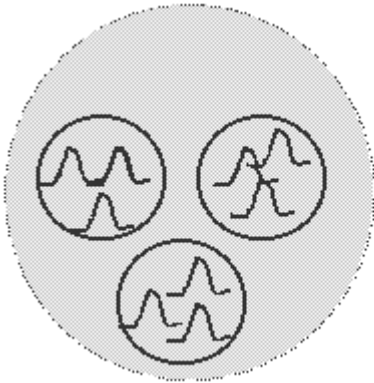


图 5-2 三个高斯混合模型的垃圾模型

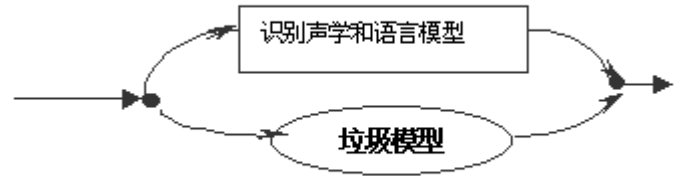


图 5-3 用垃圾模型验证

### 5.1.3 验证

验证的过程如图 5-3 所示。将垃圾模型与识别系统的声学 and 语言模型并行，识别结束时，同时给出了垃圾模型的似然度得分  $p(\vec{X} | G)$ 。当似然比

$$LR = \frac{\max_j p(\vec{X} | \sigma_j)}{p(\vec{X} | G)} < \tau \text{ 时}$$

拒绝识别结果。

图 5-4 中给出了 4 种不同结构垃圾模型的验证性能。没有标记的是 GMM，菱形标记的是 3 状态简单 HMM，三角标记的是结构 I，正方形标记的是结构 II。用一般语音 (General Speech Data) 训练的垃圾模型使系统对非法声响库 B 和 C 的拒识能力接近 100%，对无关长语音 (D) 的拒识效果也非常可观。而且，GMM 垃圾模型在这四种垃圾模型中是性能最佳的。值得指出的是，GMM 模型的训练也是这四种结构中最简单和最快的。

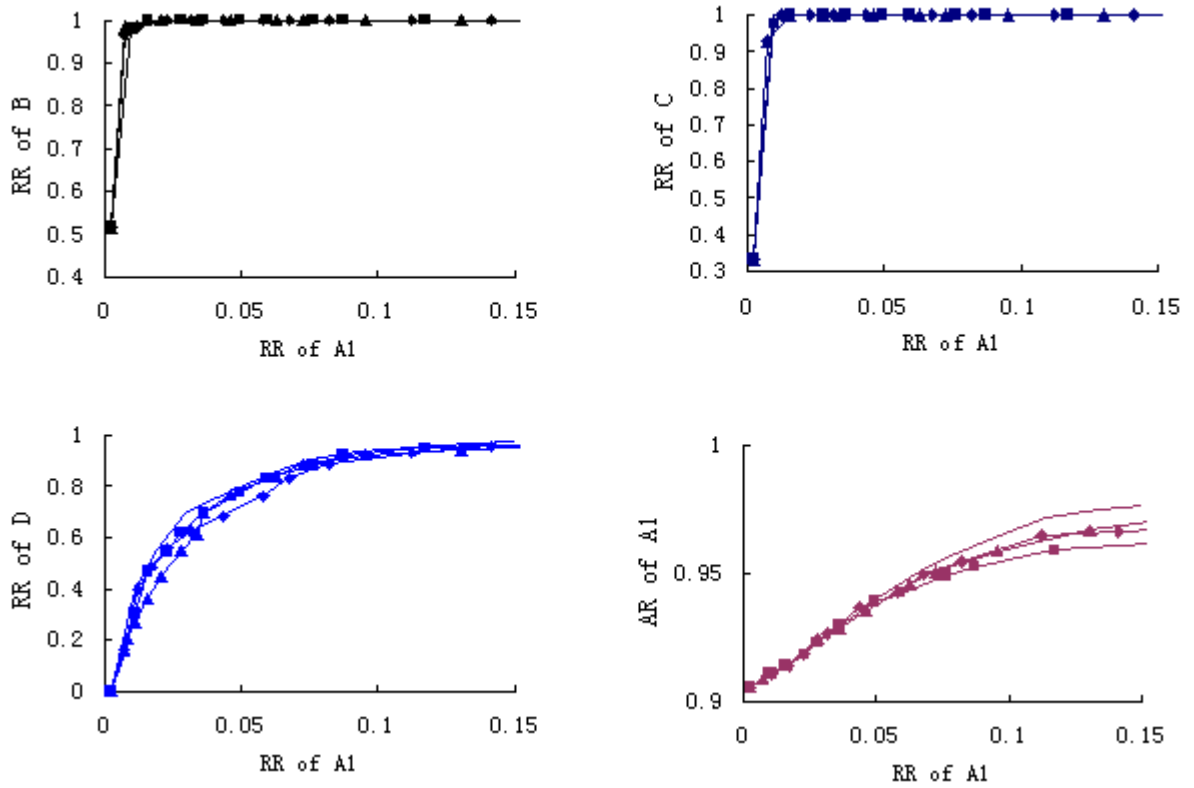


图 5-4 各种垃圾模型的验证性能

## 5.2 在线垃圾模型

### 5.2.1 在线垃圾模型

(Boite et al, 1993) (Boulevard et al,1994) 提出的在线垃圾模型实际上是利用词表中的统计模型估计备选假设的似然度。根据论文第四章 4.3.1 (注意那里  $\vec{X}$  指整个输入语音), 对一帧特征  $\vec{x}$ , 在假设

- 1)  $\vec{x}$  是合法语音的一帧;
- 2) 词表中词条的先验概率相同

成立的情况下, 有

$$p(\vec{x} | H_1) = \frac{\sum_{j \neq i} p(\vec{x} | \varpi_j)}{N-1}$$

其中  $\varpi_i$  为识别结果对应的模型， $N$  是词表的大小。即备选假设的似然度是竞争模型似然度的均值。一般把象这样根据竞争模型似然度估计的备选假设似然度称为在线垃圾似然度（Online Garbage Likelihood）。同样根据第四章 4.3.1 的分析，通常选取竞争模型似然度得分的前  $M$  个计算均值，也有人采用中数（Median）而不是均值。本论文采用均值的形式，即

$$L_{on-line}^i(\vec{x}_t) = \frac{1}{M} \sum_{j \neq i} L^j(\vec{x}_t) \quad 5-1$$

利用在线垃圾似然度，定义帧  $\vec{x}$  属于模型  $i$  的置信度为：

$$\begin{aligned} CM^i(\vec{x}) &= \log \left\{ \frac{p(\vec{x} | H_0)}{p(\vec{x} | H_1)} \right\} = \log p(\vec{x} | \varpi_i) - \log \left\{ \frac{\sum_{j \neq i} p(\vec{x} | \varpi_j)}{M-1} \right\} \\ &= LL^i(\vec{x}) - \log \left\{ \frac{\sum_{j \neq i} p(\vec{x} | \varpi_j)}{M-1} \right\} = LL^i(\vec{x}) - LL_{on-line}^i(\vec{x}) \end{aligned} \quad 5-2$$

如果在识别结果中语音段  $O_{t_1}^{t_N}$  对应模型  $i$ （在这里为半音节），可以计算模型（半音节）的在线垃圾似然度为

$$L_{on-line}^i(O_{t_1}^{t_N}) = p(O_{t_1}^{t_N} | H_1) = \prod_{k=t_1}^{k=t_N} p(\vec{x}_k | H_1) = \prod_{k=t_1}^{k=t_N} L_{on-line}^i(\vec{x}_k) \quad 5-3$$

由公式（5-2）和（5-3）可以得到语音段  $O_{t_1}^{t_N}$  属于模型（半音节） $i$  的置信度：

$$CM^i(O_{t_1}^{t_N}) = \frac{1}{N} \log \left\{ \frac{p(O_{t_1}^{t_N} | H_0)}{p(O_{t_1}^{t_N} | H_1)} \right\} = \frac{1}{N} \sum_{k=t_1}^{t_N} CM^i(\vec{x}_k)$$

对于识别结果中包含  $P$  个半音节的关键词  $i$ ，其对应语音为  $O^i = \{\Omega_1, \Omega_2, \dots, \Omega_p\}$ ，其中  $\Omega_k$  为半音节  $\varpi_k$  对应的语音段，长度为  $N_k$ ， $O^i$  的总长度为  $W$ 。

$$L_{on-line}^i(O^i) = p(O^i | H_1) = \prod_{k=1}^P p(\Omega_k | H_1) = \prod_{k=1}^P L_{on-line}^{\varpi_k}(\Omega_k)$$

词的置信度为

$$CM^i = \frac{1}{W} \log \left\{ \frac{p(O^i | H_0)}{p(O^i | H_1)} \right\} = \frac{1}{W} \sum_{k=1}^P N_k \cdot CM^{\varpi_k}(\Omega_k) = \frac{1}{W} \sum_{j=1}^W CM(\vec{x}_j) \quad 5-4$$

显而易见，这样的置信度仍然是从帧直接计算，它将各帧平等地对待，忽略了从帧到关键词的结构信息。考察下面两个电话语音识别的关键词及其相应的半音节组成：

线路教研组 [x-ian l-u j-iao y-an z-u]

电物教研组 [d-ian w-u j-iao y-an z-u]

它们各自包括十个半音节，而其中只有两个辅音是不同的 ([x/d]和[l/w])。而且，这几个辅音本来就相对较短，他们在整个语音中所占的帧数非常少。因此，如果平等对待各帧语音而不考虑结构信息，就会使不同的两个辅音提供的区分信息被其他相同半音节的良好匹配模糊掉，用公式 (5-4) 计算出的置信度也就无法区分识别结果究竟对不对。因此，本论文采用分层次的置信度整合 (Hierarchical Integration)，从帧到半音节整合时，平等对待帧置信度；从半音节到关键词整合时，平等对待半音节置信度；从关键词到整个说话 (Utterance) 时，平等对待关键词置信度。与第三章所强调的利用语音结构信息是一致的。这样关键词的置信度计算为

$$CM^i = \frac{1}{P} \sum_{k=1}^P CM^{\varpi_k}(\Omega_k) = \frac{1}{P} \sum_{k=1}^P \left\{ \frac{1}{N_k} \sum_{j=1}^{N_k} CM(\vec{x}_{kj}) \right\} \quad 5-5$$

其中  $\vec{x}_{kj}$  为半音节  $\varpi_k$  的第  $j$  帧语音特征矢量。同理，整个输入说话识别结果的置信度为：

$$CM = \frac{1}{Q} \sum_{i=1}^Q CM^i \quad 5-6$$

其中  $Q$  为说话中关键词的个数。

在文献(Rivilin et al, 1996)(Lleida and Rose,1996) (Kawahara et al, 1997) 中都证明分层次的整合计算置信度的优越性, 在绝大多数研究中对这种思想都有所体现(Weintraub et al, 1997) (Koo et al, 1998)。在本章 5.3 节中, 也将通过实验比较这种分层次整合方式的优势。

注意说话结果的置信度计算公式 (5-6), 只采用了识别结果中对应于关键词部分的语音计算置信度。这样做有两个原因, 第一是减少计算置信度的运算量; 另一方面, 由于人们在表达意思的时候, 倾向于将自己认为重要的部分发音清晰, 而把其他辅助的语言结构发得随意。例如在电话语音识别系统中, 一个使用者说: “请帮我接线路教研组的刘润生老师”, 他会将与自己目标相关的“线路教研组”和“刘润生”加以强调, 使这两个部分更加清楚明白, 而对其他辅助部分则会敷衍过去。因此, 采用对应着关键词发音清晰的部分验证会有更高的可靠性。再一方面, 即使识别系统把非关键词部分识别错误, 在关键词正确的情况下, 识别能够完成使用者的要求, 不应该把这个识别结果视为识别错误, 比如把上面那句话识别成“请给我接线路教研组的刘润生教授”。由此可见, 非关键词的错误或正确与最后的识别结果错误或正确无关, 因此在计算识别结果的置信度时应该忽略它们提供的信息。

### 5.2.2 帧上的与半音节上的在线垃圾模型

如公式 (5-1) 所示, 在线垃圾模型似然度通常是逐帧计算的, 帧上的在线垃圾似然度再分层次地集成为半音节, 词和输入说话的置信度。根据公式 (5-3), 半音节的垃圾对数似然度可以计算如下(Bourlard et al 1994) (Colton, 1997) (Jitsuhiro et al, 1998) (Sukkar, 1998) (Leung and Fung,1999) :

$$LL_{on-line, phoneme}^i(O_{t_1}^{t_N}) = \sum_{k=t_1}^{t_N} LL_{on-line}^i(x_k)$$

$LL$  代表对数似然度, 即  $LL = \log L$ 。  $O_{t_1}^{t_N}$  为输入语音段从第  $t_1$  帧到第  $t_N$  帧, 它被识别为对应着半音节  $I$ 。

然而, 我们发现, 直接计算半音节的在线垃圾似然度在效率和性能上都要好。设语音段  $O_{t_1}^{t_N}$  在识别结果中对应半音节  $I$ , 然后把该语音段与半音节  $I$  对应的竞争半

音节分别对准，得到与它们的长度归一化对数似然度  $LL^j(O_{t_i}^{t_N})$ 。半音节  $I$  的在线垃圾对数似然度就用竞争半音节的前  $M$  个似然度得分估计：

$$LL_{on-line}^i(O_{t_i}^{t_N}) = \log \left\{ \frac{1}{M} \sum_{j \neq i} \exp(LL^j(O_{t_i}^{t_N})) \right\} \quad 5-7$$

此后，半音节的垃圾似然度再用来计算词条和输入语音的置信度。我们把半音节  $i$  称为主半音节 (Master)，把它的竞争集里的半音节称为从半音节 (Cohorts)。按照通常的做法，一个半音节的从半音节包括除了主半音节所有的其他半音节。这里考虑声母和韵母的差异，声母的从半音节只包括其他所有声母，对韵母也一样。由于半音节数量很大，计算垃圾似然度的运算量也很惊人。运算负担主要是两个方面：1) 计算与从半音节的似然度；2) 对从半音节的似然度排序。直接计算半音节垃圾似然度将每帧语音排序一次减少为每个半音节排序一次，大大地减少第二方面的运算。在本章 5.2.3 中，将研究如何减少从半音节，从而减少第一方面的运算。下图显示出直接计算半音节垃圾似然度在验证性能上的优势，对非法声响和误识的拒识率都要高一些。对从帧计算的垃圾似然度， $M$  取 16，对直接计算半音节垃圾似然度， $M$  取 3。 $M$  的值是通过实验确定的，分别使两种情况达到各自的最优性能。标记着小方块的曲线为直接计算半音节垃圾似然度，标记着小三角的为从帧计算的垃圾似然度。

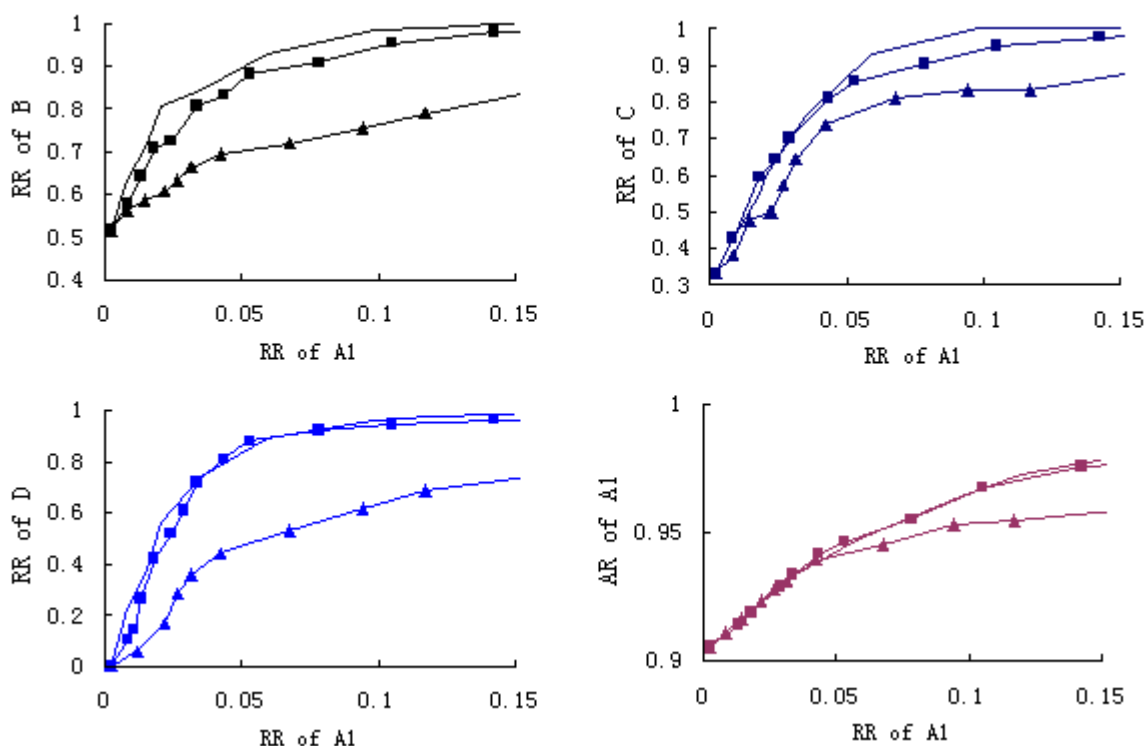


图 5-5

直接计算的半音节垃圾似然度与从帧计算的垃圾似然度相比,考虑了半音节的结构。例如,如果一个对应于半音节[ie]的语音段被识别成半音节[ie]或者[ei],这是两个不同的假设,前者正确后者错误。从帧计算的垃圾似然度对两个假设的几乎一样,这是由于从帧计算垃圾似然度没有考虑各帧语音出现的顺序,因此用这样的垃圾似然度不能较好地给出这两个假设的正确性。相反,对半音节[ie]和[ei]直接计算的半音节垃圾似然度就会有较大的差异,显然前者的会远小于后者的,因此对评价假设的正确性就很有用处。考虑了半音节的结构信息,这是直接计算的半音节垃圾模型的性能上优势的主要原因。

为了消除匹配极端情况(匹配过差)的负面影响,根据第四章 4.3.1 的分析,采用几何平均来计算归一化垃圾似然度,即

$$LL_{on-line}^i(O_{t_1}^{t_N}) = \frac{1}{\gamma} \log \left[ \frac{1}{M} \sum_{j \neq i} \exp(LL^j(O_{t_1}^{t_N})\gamma) \right]$$

权重  $\gamma$  应该正确设置,  $\gamma$  过大, 匹配好的从半音节将掩蔽掉其他从半音节的贡献;  $\gamma$  过小, 匹配不好的个别从半音节, 也会产生掩蔽效应。对于半音节垃圾似然度, 本论文取  $\gamma = 0.001$ 。如图 5-6 显示, 采用几何平均(无标记的曲线)使垃圾似然度对非法声响的拒识率得到提高, 对误识的拒识率几乎没有变化。

### 5.2.3 竞争集优化

减少每个半音节的从半音节数, 也就是减小对应的竞争集对提高验证速度至关重要。而减少从半音节就意味着忽略信息, 往往会引起验证性能的下降。因此, 需要将竞争集优化, 寻找与任务要求接近的折中。在这一小节, 将以前一节采用直接计算半音节垃圾似然度和几何平均的验证方法作为基本系统, 来比较不同的竞争集优化方法。

首先注意到, 由于口音的影响, 许多说话人将[n]发成[l], 翘舌音与对应的平舌音不分(如[sh]与[s]), 后鼻音与对应的前鼻音不分(如[en]与[eng])。也就是说, 在合法语音中, 有许多半音节的发音是“不合法”的。由于这些发音上混淆的半音节在声学特征上也是相似的, 发音真正对应的半音节此时就在竞争集中。于是这些不合法的半音节发音对应的垃圾似然度往往偏大, 接着导致置信度下降。使得对正确识别的拒绝率偏高。因此, 应该将这样的半音节(称为模糊音)从竞争集中去掉。图 5-6 显示出去掉模糊音带来验证性能的提高。其中, 有标记的为去掉模糊音后的验证性能,

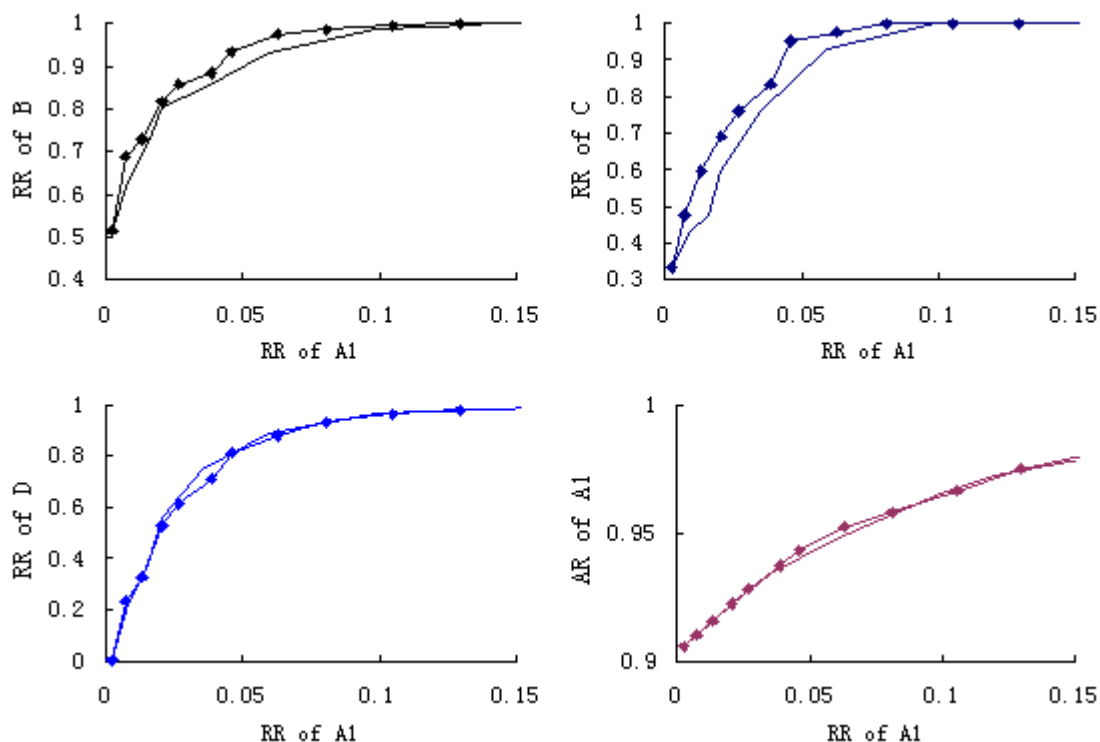


图 5-6

在(Sukkar and Lee, 1996)中, 作者根据训练语音中各个子词模型语音的混淆矩阵为每个子词找出一个固定大小的从模型集 (Cohort Set) 作为竞争集并且用所有从模型的似然度得分来计算在线垃圾似然度。从模型集大小远远小于子词模型的个数, 所以大大地减小了运算。本论文通过直接比较 HMM 的相似性 (Similarity) 为每个半音节找出从半音节和竞争集。根据 (Rabiner and Juang, 1993), 对两个 HMM  $\varpi_i$  与  $\varpi_j$  可以定义距离测度为

$$D(\varpi_i, \varpi_j) = E_{\vec{O}^i} \left\{ \frac{1}{T^i} \log \left( \frac{p(\vec{O}^i | \varpi_j)}{p(\vec{O}^i | \varpi_i)} \right) \right\};$$

其中  $\vec{O}^i$  为模型  $\varpi_i$  对应的语音 (特征序列),  $T^i$  指语音的长度。  $E_{\vec{O}^i}$  表示对模型  $\varpi_i$  对应的所有语音求期望。可以用训练语音估计其经验值:

$$\bar{D}(\varpi_i, \varpi_j) = \frac{1}{N_i} \sum_{\vec{O}^i} \left\{ \frac{1}{T^i} \log \left( \frac{p(\vec{O}^i | \varpi_j)}{p(\vec{O}^i | \varpi_i)} \right) \right\}$$

$N_i$  为  $\varpi_i$  对应的训练语音数目。同样考虑模型  $\varpi_j$  对应的语音，可以得到对称距离

$$\overline{D}_s(\varpi_i, \varpi_j) = \frac{1}{2} \{ \overline{D}(\varpi_i, \varpi_j) + \overline{D}(\varpi_j, \varpi_i) \} \quad 5-8$$

由于元音和辅音的相对独立性，只在元音中为元音寻找从半音节，在辅音中为辅音寻找从半音节。设置一个距离门限来判断某个半音节是否选为从半音节，也就是说：如果  $\overline{D}_s(\varpi_i, \varpi_j) < \delta$ ，那么确定  $\varpi_i$  与  $\varpi_j$  互为从半音节。由于  $\delta$  对所有半音节是相同的，最后每个半音节的从半音节数目也就是竞争集大小是不同的。这与(Sukkar and Lee, 1996)的做法正好相反，那里设置竞争集的大小为常数。我们认为，不同的半音节与其他半音节的相似程度是不同的，对于特征不够明显（与许多别的半音节相似）的半音节应该考虑采用更多的竞争模型。设置门限，使平均每个半音节有 10 个左右的从半音节。此时，竞争集的大小在 5 到 20 之间（附录 B 给出了 43 个无调韵母对应的竞争集）。因此，与原来采用所有模型作为从半音节相比，第一部分运算下将到原来的 10% 左右。与(Sukkar and Lee, 1996)的另一处不同是，我们并不采用所有的从半音节似然度得分来平均以计算在线垃圾似然度，而还是采用取从半音节似然度得分中的前  $M$  个。实验证明，这样做避免了竞争集中的不良匹配的过分影响，使对误识和非法声响拒绝的能力提高。竞争集减小后，验证性能如图 5-7 所示。

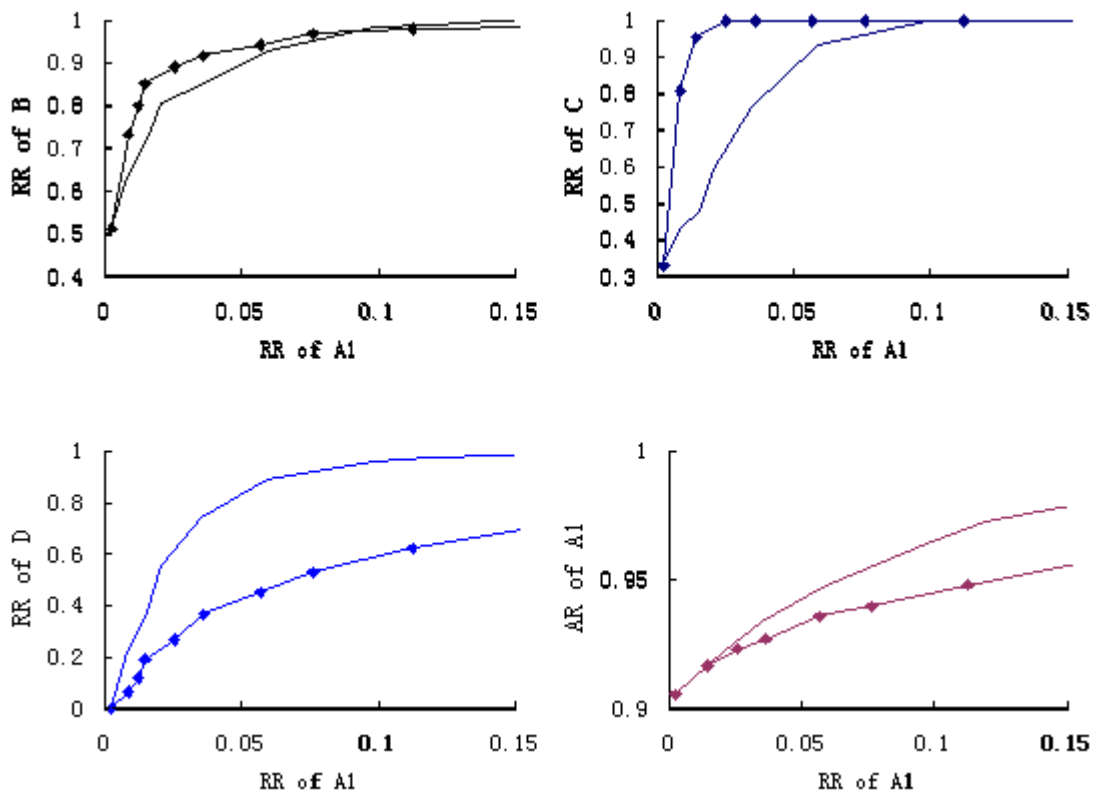


图 5-7

如图 5-7，减小竞争集后（有标记的曲线），对误识和无关长句子语音的拒识能力显著下降，对短小的非法声响的拒识能力明显提高。减小竞争集后，所有输入声响的垃圾似然度都会下降。对短小的非法声响，由于其声学上与底层的半音节模型差异较大，当把某个声响片段（Segment）识别成半音节 $\omega_i$ 时，这个片段和那些跟 $\omega_i$ 相似（在公式 5-8 定义的距离的意义上）的半音节的得分与和 $\omega_i$ 的得分相差不大。在竞争集中往往是这些半音节的得分排在最前面，因此减小竞争集后，它们的垃圾似然度下降与正确识别合法语音的下降相比要小得多。而另一方面，误识和无关长句子在声学上与底层的半音节模型差异不大，误识中有相当一部分是把合法语音在半音节层次上错误分割造成的；无关长句子一方面半音节层次上的错误分割更多，另一方面由于句子长，半音节丰富（Phonetically Rich），在与允许的搜索路径强行匹配时，往往有许多半音节是匹配正确的。这些现象都使得在匹配中，某个半音节的得分占较大优势，而在竞争集中排在前面的半音节与识别出的半音节也不

一定很相似，导致根据相似度减小竞争集后，其垃圾似然度下降与正确识别的合法语音的相比往往还要多些。这就导致了误识和无关长句子置信度增加比正确识别的置信度增加更多，而短小非法声响的置信度却增加更少。因此出现了上述的实验现象。

### 5.3 置信度分层次集成

在第三章，提到验证利用语音结构信息的重要性；在本章 5.2 节中，介绍了根据这种想法分层次地计算置信度的方法。这里给出两组实验结果以证明分层次集成置信度的优越。

#### 5.3.1 没有半音节层的集成

没有半音节层的集成是指直接从帧似然度直接计算关键词的置信度。这样，各个半音节对关键词置信度贡献不是相同的，而是越长的半音节贡献越大。则计算关键词的置信度由公式 (5-5) 变为：

$$CM^i = \frac{1}{W} \sum_{k=1}^P N_k \cdot CM^{\varpi_k}(\Omega_k)$$

其中变量的意义同 5.1 节。而其他的置信度计算方法不变。

#### 5.3.2 没有关键词层的集成

指直接从半音节置信度计算说话识别结果的置信度。识别假设输入说话中（的所有关键词）共有  $P$  个半音节，对应语音段  $\Omega_1, \Omega_2, \dots, \Omega_p$ ，对应半音节模型  $\varpi_1, \varpi_2, \dots, \varpi_p$ ，则计算识别结果的置信度由公式 (5-6) 变为：

$$CM = \frac{1}{P} \sum_{k=1}^P CM^{\varpi_k}(\Omega_k)$$

### 5.3.3 验证性能比较

实验结果如图 5-8，没有标记的曲线是完全分层次集成计算置信度，竞争集经过聚类优化。标记着小菱形记号的曲线是同样竞争集，但计算置信度时没有关键词层的集成。标记着小三角记号的曲线则没有半音节层的集成。分层次地集成置信度，大大地提高了验证对非法声晌的拒识能力。

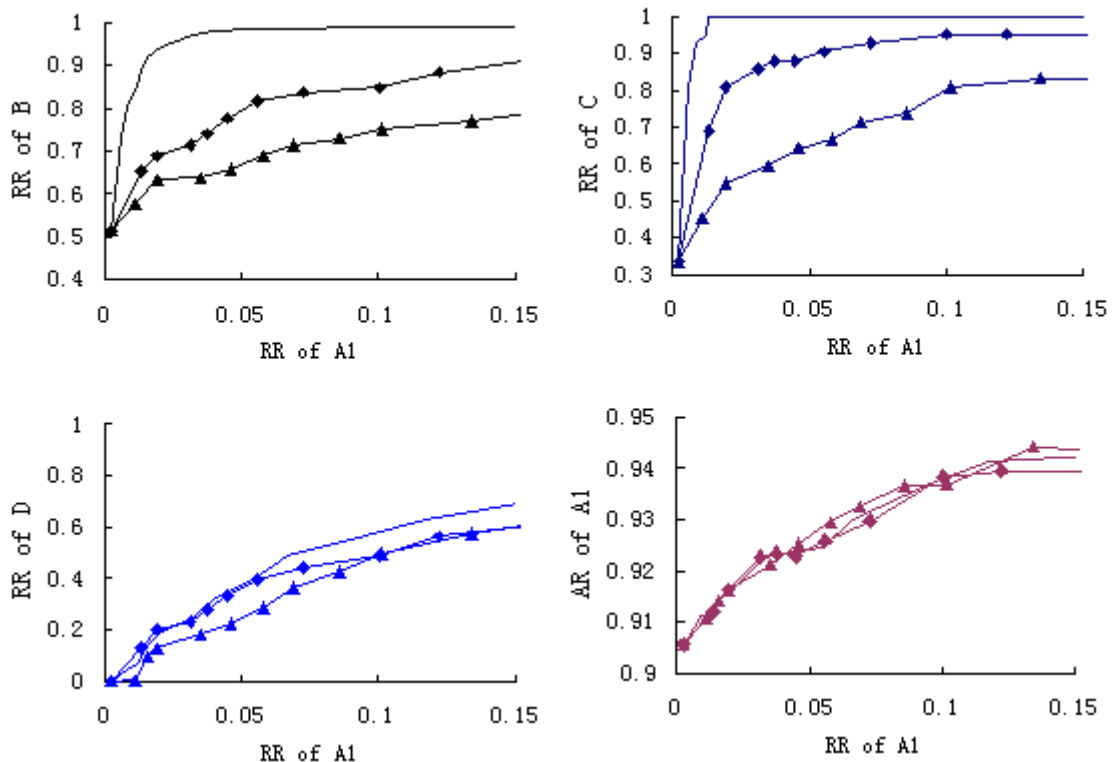


图 5-8

## 5.4 其他拒识方法和说话人

### 5.4.1 前二选

基于多子树的电话语音识别系统最后给出合乎规则子树的多候选结果，根据第三章 3.2，当合法结果数小于二，认为识别结果是错误的。当合法结果数不小于二时根据前二选的似然比验证拒识。

### 5.4.2 反词模型

用 863 语音数据库中的 16 个说话人的语音数据为每个半音节训练了一个反词模型，确切地说是反子词模型 (Anti-sub-word Model)。反子词模型和半音节模型一样，采用简单的 HMM 结构，单高斯，协方差矩阵不限制为对角阵。声母和韵母的反半音节模型训练数据是完全不同。训练声母  $\omega_i$  的反半音节模型  $\bar{\omega}_i$  的语音是所有其他声母对应的语音；对韵母的也是如此。反声母模型和声母模型一样，有 2 个状态，反韵母模型有 4 个状态。

通常，应该将训练语音与模型迭代对准并统计各状态的参数。从简化训练的角度出发，假设训练语音与反半音节模型的对准情况同该语音与对应的半音节模型的对准情况是一样的。也就是说，反半音节  $\bar{\omega}_i$  的某段训练语音  $\bar{X}$  对应着半音节  $\omega_j$  ( $j \neq i$ )，我们假设  $\bar{X}$  中分配给  $\omega_j$  第  $k$  个状态的语音也应该分配给  $\bar{\omega}_i$  的第  $k$  个状态 (对声母， $k = 1, 2$ ；对韵母， $k = 1, 2, 3, 4$ )。对韵母，我们将仅仅不同调的韵母看作同一种韵母，因此对仅仅不同调的韵母，它们的反韵母模型是相同的。训练的过程是首先用训练好的半音节模型把训练语音分割到状态一级，然后按照上述假设对反半音节各状态的参数加以统计，最后得到 100 个反声母 HMM 模型和 43 个反无调韵母模型。

得到反半音节模型后，就可以用反半音节的似然度代替公式 (5-7) 中的半音节在线垃圾似然度计算说话的置信度。

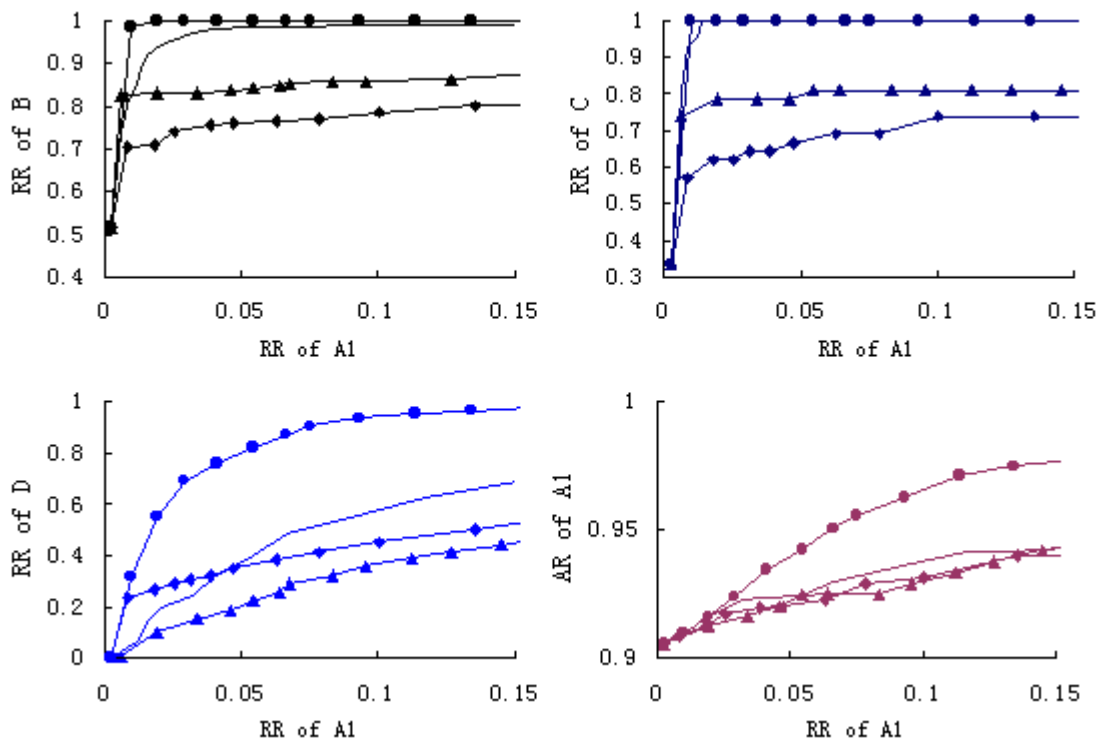


图 5-9

图 5-9 给出了采用反半音节模型（标记着菱形的曲线），前二选似然比（标记着三角形的曲线），竞争集优化后的在线垃圾模型（没有标记的曲线）和 GMM 垃圾模型（标记着圆点的曲线）的验证性能。这四种方法都是词表/任务无关的，不需要词表/任务相关的训练数据。在性能上，垃圾模型和在线垃圾模型具有较大优势；在参数存储要求上，在线垃圾模型和前二选似然比几乎为零；在计算量上，只有在线垃圾模型的额外增加非常可观。综合几个方面，认为采用 GMM 的垃圾模型的验证性能是最令人满意的。

### 5.4.3 对其他说话人的拒识

以上，采用电话语音识别系统测试语音库中识别率最高的 4 个说话人的各 207 句话研究了说话验证，这里将把以上几种主要方法应用到另外 4 个识别率很低的说话人的语音（图 5-10 中称为 A1 new）上，以说明这些方法对不同识别率说话人的有效性。

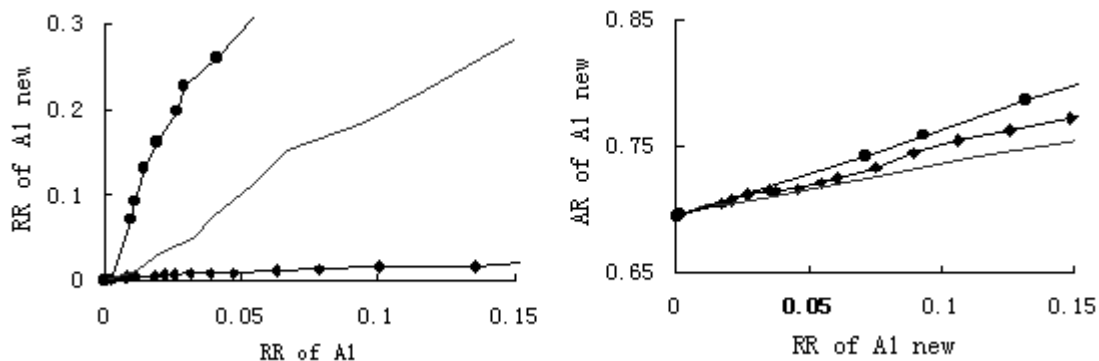


图 5-10

如上图，其中的标记与图 5-9 相同。左图显示出性能好的说话人某个拒识率对应拒识门限应用到 A1 new 上的拒识率。右图显示出 A1 new 的识别率随拒识率上升的变化，即对 A1 new 中误识的拒识能力。对非法声响，拒识方法的评价完全可以根据图 5-9 得到。从图 5-10 左图可以评价验证方法对不同说话人的稳定性。理想的方法，在工作点上对识别率高的说话人的拒绝率应该比对识别率低的说话人低。如图中所示，在线垃圾模型和垃圾模型是满足这个要求的。

## 5.5 在语音确认系统中的性能

将同样的方法在语音确认系统中实现并测试其性能，而且将电话语音识别系统的合法语音库 A1 在这里用做一种非法语音，以说明方法的任务/词表无关性。

图 5-11 显示出反半音节模型，垃圾模型和在线垃圾模型在语音确认系统中的验证性能，图中的标记同图 5-9，仅仅在对应非法语音 D 库的图中附加了对电话语音识别系统的合法语音库 A1 的拒识曲线（图中虚线所示）。三种方法的模型和参数与电话语音识别系统中使用的完全相同（当然，拒识门限是不同的）。

可以看到，电话语音识别系统上的研究结论到语音确认系统中仍然是成立的。需要注意的是三点。第一，电话语音识别系统的合法语音在这里是非法语音，各种拒识方法对它们的拒识能力同对无关长语音的拒识能力相当，这从一个角度证明了三种方法的任务/词表无关性；第二，语音确认系统的识别率高达 99.3%，因此对误

识的拒识不再是验证的主要任务，图中显示，加入拒识后，系统的识别率（AR）甚至出现了轻微的下降（对垃圾模型和反半音节模型），这是拒识非法语音的代价；第三，各种方法在确认系统中的验证性能都远远低于在电话语音识别系统中的性能。这是由于语音确认系统的声学 and 语言模型简单，[正确/错误]各自只有四个半音节。因此，可资验证利用的信息要少许多，合法语音与非法语音之间的区分信息也就要少许多。这是验证性能下降的基本原因。

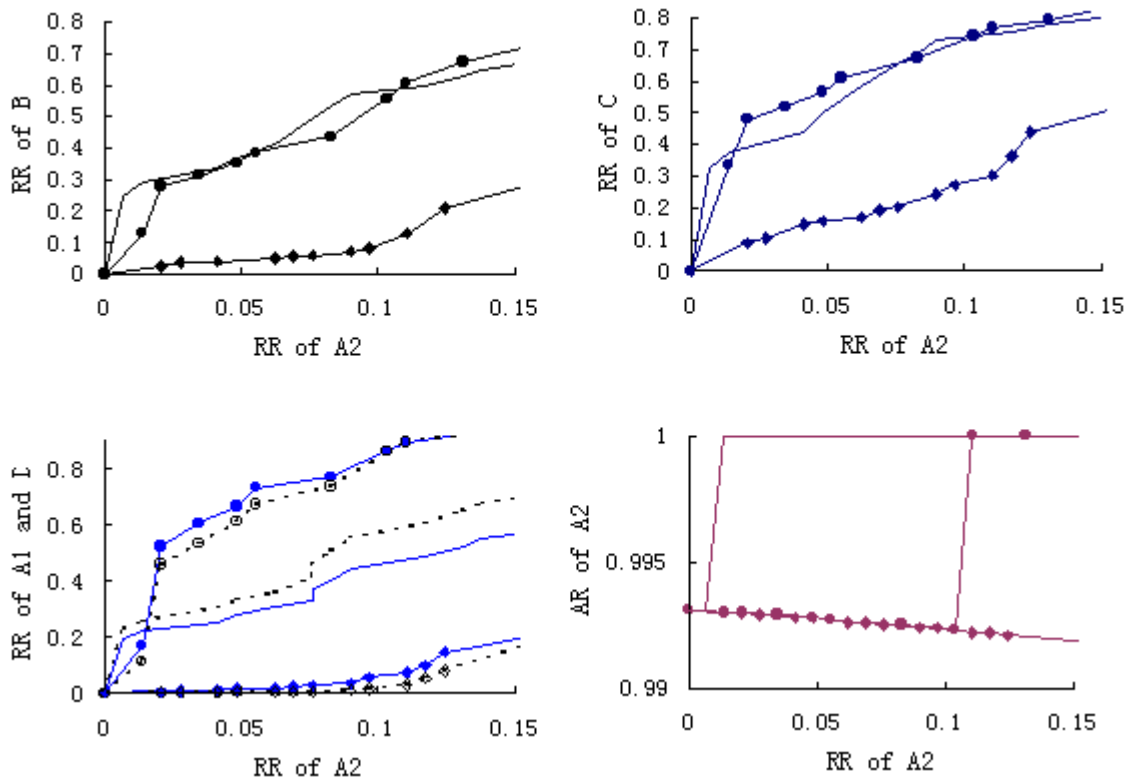


图 5-11

## 5.6 小结

本章研究了使用垃圾模型和在线垃圾模型对汉语语音识别进行词表无关验证，并将它们与前二选似然比和反半音节模型验证做了比较。提出了基于 GMM 的垃圾模型，这样的垃圾模型无论是在验证性能还是训练速度上都具有优势。

对在线垃圾模型和基于半音节子词模型的汉语语音识别，从考虑半音节结构信息出发，本论文提出直接计算半音节在线垃圾似然度。这样做显著地提高了在线垃圾模型方法对各种识别错误的拒识能力。

另一方面，为降低在线垃圾似然度的运算代价，本论文采用直接比较模型相似度的方法，将计算在线垃圾似然度的竞争半音节集减小到原来的 10% 左右，在保持相当的验证性能的同时，大大地提高了验证的速度。

针对基于半音节子词模型的汉语语音识别系统，本论文提出反半音节模型的方法估计识别结果的似然度。与垃圾模型相比，反半音节模型目前还不具有任何优势。期望深入的研究会进一步提高这种方法的性能。

表 5-1 列出了几个工作点上，竞争集减小后的在线垃圾模型(OLG-Reduced)，竞争集中去掉模糊半音节的在线垃圾模型(OLG)(这两者都直接计算半音节垃圾似然度)以及采用 3 个 GMM 的垃圾模型对不同识别错误的拒识能力。工作点 RR of A1=0.2% 对应直接拒绝没有合乎规则识别结果 (ResultNum=0) 的情况。

表 5-1

RR of A1(%)		0.2	5	10
AR of A1(%)	OLG-Reduced	90.6	93.3	94.6
	OLG	90.6	94.5	96.5
	GMM	90.6	94.1	95.7
RR of B(%)	OLG-Reduced	51.5	93.6	97.6
	OLG	51.5	94.4	99.2
	GMM	51.5	100.0	100.0
RR of C(%)	OLG-Reduced	33.3	100.0	100.0
	OLG	33.3	96.0	100
	GMM	33.3	100.0	100.0
RR of D(%)	OLG-Reduced	0.2	43.6	60.1
	OLG	0.2	82.4	96.0
	GMM	0.2	80.0	94.1

## 第六章 验证信息源综合

在第三章中，论述了许多可供验证识别结果正确性的信息源（Knowledge Sources），并研究了利用音节段长信息。在第四和第五章中则详细地研究了另一些信息源。面对着各种各样的信息，一个很自然的想法就是把它们结合起来。但是，各个信息源提供的信息如果以变量的形式有不同的形式和数量级，如何将这些变量综合（Combine）起来提高验证的性能是本章的研究内容。

### 6.1 基于规则（Rule Based）的综合

最简单直接的做法是制定一组规则 $\{R_i\}$ ，每个规则利用不同的信息源进行验证判断，按照一定的顺序使用。表 3-1 实际上给出了一个基于规则的简单信息源综合：

```
if [ (识别合法结果数<2) 或者 (最小韵母长度<10) 或者(最大声母长度>40)],  
    then 拒绝识别结果;  
else 接受识别结果;
```

在第五章 5.4.1 中，也利用了简单的规则综合：

```
if [ (识别合法结果数<2) ] then [拒绝识别结果];  
else  
{  
    if [ (前二选似然比< $\tau$ ) ] then [拒绝识别结果];  
    else 接受识别结果;  
}
```

第三章 3.3 中指出，当识别结果中音节数较多时（一般来说对应输入语音较短），应用归一化音节长度方差验证才可靠，而由第五章 5.2.3 可知，竞争集减小后的在线垃圾模型对短的非合法声响拒识性能优异。对于电话语音识别系统，识别结果关键词的音节数  $N$  从 2 到十几不等。可以制定简单的关于结合这两种信息源验证的规则如下：

```
if [N<4] then {
    根据在线垃圾似然度计算置信度  $CM_{OLG}$  ;
    if [ $CM_{OLG} < \tau_1$ ] then 拒绝识别结果;
    else 接受识别结果;
}
else
{
    计算  $Var'_S$  ;
    if [ $Var'_S > \tau_2$ ] then 拒绝识别结果;
    else
    {
        根据在线垃圾似然度计算置信度  $CM_{OLG}$  ;
        if [ $CM_{OLG} < \tau_1$ ] then 拒绝识别结果;
        else 接受识别结果;
    }
}
```

门限  $\tau_1$  和  $\tau_2$  的变化将使验证器的工作点组成一个面而不是一条线。为简单起见，根据第三章 3.3 的分析取  $\tau_2 = 8$ 。可以得到因  $\tau_1$  变化得到的验证工作曲线，如图 6-1。图中标记着菱形的曲线仅仅使用  $CM_{OLG}$ ；没有标记的曲线采用上述规则综合  $CM_{OLG}$  和  $Var'_S$ 。计算  $Var'_S$  需要的运算量很小，用它拒绝掉许多明显的错误而避免计算  $CM_{OLG}$  提高了验证的整体速度，而且使验证对无关长句和误识的拒识能力提高许多，这显示出综合信息源的优势。

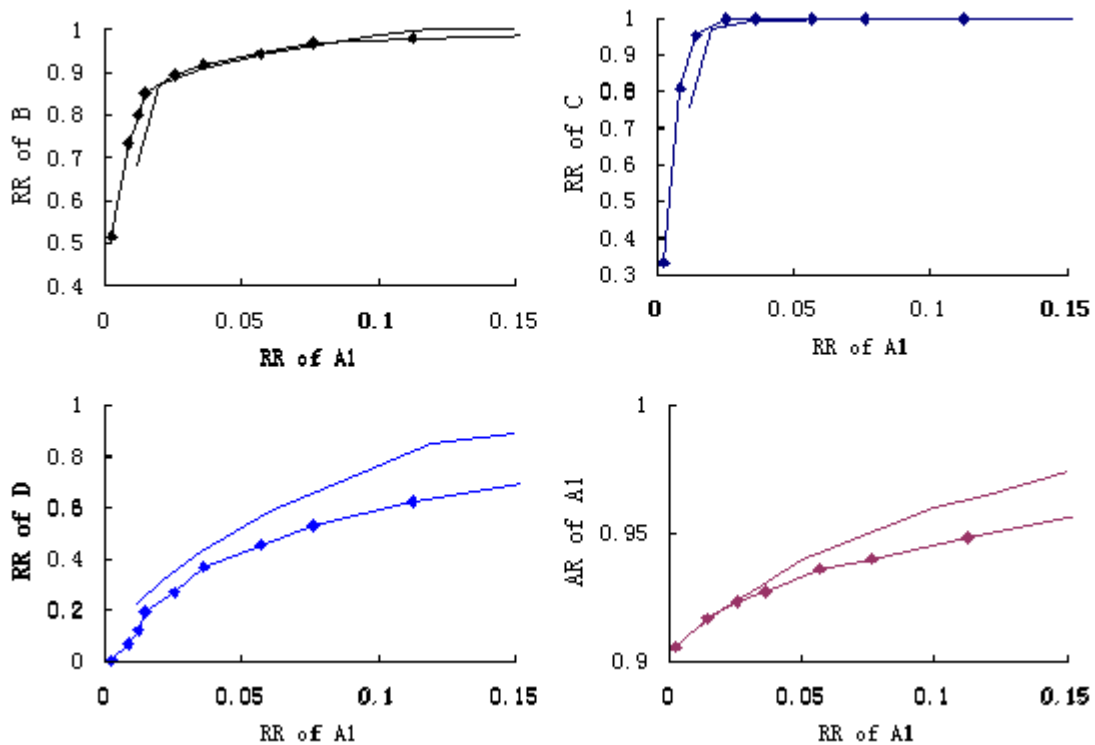


图 6-1

基于规则的综合适合于利用那些意义明确的“硬”信息，用来去除一些显而易见的识别错误。但是，对于诸如垃圾似然度，音节长度方差之类的“软”信息，就很难用规则把它们综合起来。难度存在于以下几个方面，第一是信息源之间存在统计相关性，使用规则不易于反映统计相关性；第二，使用规则意味着要对信息源提供的参数进行“硬”判决，但是一方面，信息源可能本身的信息就不够充分来做“硬”判决，另一方面，硬判决往往意味着门限的选择，如何在整体上优化规则组，为每个规则设置整体最优的判决门限，是非常困难的问题。因此在下一节中，将研究基于统计模型的信息源综合。

对于基于规则的信息源综合，考虑验证速度和可靠性，本论文认为，在使用规则顺序时应该遵循以下两个原则：

- 1) 拒绝条件越苛刻的规则越先使用。
- 2) 计算量越小的规则越先使用。

## 6.2 基于统计模型 (Statistical Model Based) 的综合

可以这样说，基于规则的信息源综合输出的是 0（拒绝）和 1（接受）。我们希望，利用各种不同信息源，也能给出一个“软”的置信度，表明识别假设正确的可能性的大小。这样，可以把寻找一组整体最优门限的问题转化为寻找一个整体最优的门限。换句话说，对两个不同信息源提供的两个参数  $X$  和  $Y$ ，希望能够找到一个影射

$$Z = f(X, Y)$$

使根据参数  $Z$  验证识别结果比分别使用  $X$  和  $Y$  都要好。这类似于模式识别中的分类器设计的问题 (Duda and Hart, 1973)。为简化问题起见，可以假设，这个映射的函数形式是已知的，需要确定的只是函数中的参数  $\Theta$ 。在这个假设之下，需要做的就是 1) 确定函数形式；2) 选取优化准则和手段估计  $\Theta$ 。

根据第二章 2.2 的分析，可以把验证按二类模式分类 (Binary Pattern Classification) 来处理。优化准则选取为最小分类错误准则。可以从分类器设计的角度解决这个问题。在 (Siu and Gish, 1999) 中，采用本质是线性模型和广义线性模型的 Logit Model 和 Generalized Additive Model 来综合信息源。(Kemp and Schaaf, 1997) 和 (Schaaf and Kemp, 1997) 采用 MLP 进行信息源综合进行词 (Word) 一级的正确性标注，用于对自动翻译系统 JANUS 语音识别的标注，并比较了其线性模型综合的性能。(Modi and Rahim, 1997) 也采用 MLP 来综合不同信息源估计的置信度，并应用于电话数字语音识别的验证中。在本论文中，将研究线性模型和 MLP 综合信息源估计说话一级的置信度。

### 训练与测试数据

对 A1 库采用识别率最低的四个说话人各 207 句话中的前 107 句为训练语音，另外四个说话人的后 100 句为测试语音；对库 B 和 C 取前 10 个人的声响为训练语音，后 10 个人的为测试语音。对这些语音标记好识别结果的正确性，正确为 1，错误为 0。由于条件的限制，用于训练和测试的数据都非常有限。期待着增加训练数据会带来性能的进一步提高。

## 信息源选择

在 (Siu and Gish,1999) 中, 采用贪心算法 (Greedy Incremental Selection Algorithm) 选择信息源。基于本论文前三章的研究, 考虑如表 6-1 候选信息源。选择它们的原因是考虑到它们所反映的信息在属性和计算来源上有较大差异。这四种参数从做到右分别以字母 V, L, O 和 G 表示。下面将比较采用不同信息源组合, 验证的性能。

表 6-1 信息源

信息源	音节长度归一化方差 $Var_s'$	识别结果的帧平均似然度 $L$	根据竞争集优化的在线垃圾模型计算的置信度	根据 GMM 垃圾模型计算的置信度
计算方式	识别给出词格, 从词格计算	识别给出	后处理计算	与识别几乎同时计算

## 线性模型

线性模型见第四章 4.3.3。其结构简单, 参数少, 但逼近能力有限。

### MLP

由于训练语音非常有限, 选取 MLP 隐层神经元个数为 3, 这样也限制了 MLP 的逼近能力。(Kemp and Schaaf, 1997) 发现隐层神经元数过少对综合性能的影响不大。本论文认为这只对信息源相对少时才成立。

## 实验结果

图 6-2 显示出用 MLP 综合信息源, 验证性能随信息源的增加而提高。其中虚线所示为根据 GMM 垃圾模型计算的置信度的验证性能; 有圆形标记的为 MLP 综合参数 G 和 V 估计的置信度的验证性能; 有三角标记的为 MLP 综合参数 G, V 和 L 估计的置信度的验证性能; 无标记的实线为 MLP 综合参数 G, V, L 和 O 估计置信度的验证性能。由于 GMM 估计的置信度验证对库 B 和 C 的拒识已经接近理想, 这四种情况对库 B 和 C 的拒识能力实际上无实质上的差别。信息源增加带来

的性能提高主要表现在对库 D 和误识的拒识上。考察一个工作点，在对合法语音的拒识率为 10% 时，与最好的单个信息源拒识性能（这里就是 GMM 垃圾模型）相比，MLP 综合四种信息源将对库 D 的拒识率提高了 5%（从 94.1% 提高到 99.1%），拒识后识别的精度提高了 1.5%（从 95.7% 提高到 97.2%）。而付出的运算和存储代价是微乎其微的（MLP 只有 23 个参数）。考虑到训练数据的

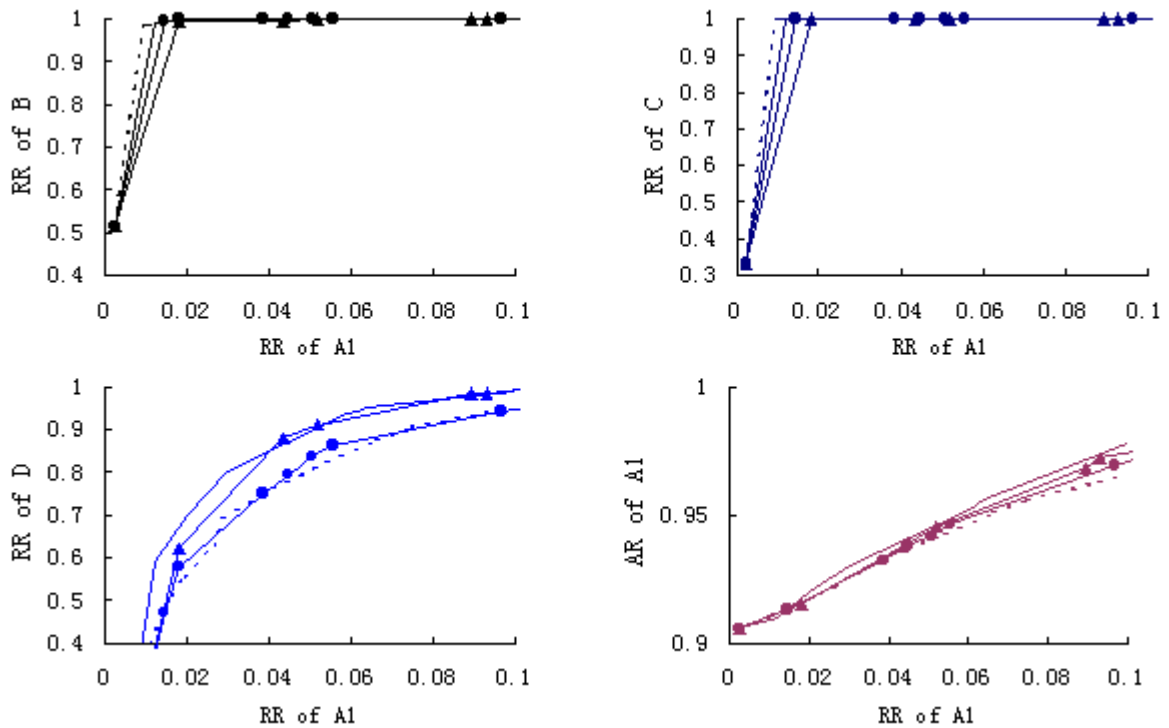


图 6-2

图 6-3 则显示出 MLP 综合信息源相对线性模型的优势。其中标记着菱形的曲线对应采用 GMM 垃圾模型，标记着三角的曲线对应采用线性模型综合 V, L, O 和 G，没有任何标记的曲线对应采用 MLP（3 个隐层神经元）综合 V, L, O 和 G。同样由于对库 B 和 C 的拒识本来就接近理想，MLP 的优势还是体现在对库 D 和误识拒识能力上。由于在综合信息源的问题上，MLP 与线性模型需要的存储和运算都微乎其微（这一点与第四章用 MLP 估计 HMM 迹的后验概率不同），本论文认为选择 MLP 是明智的。

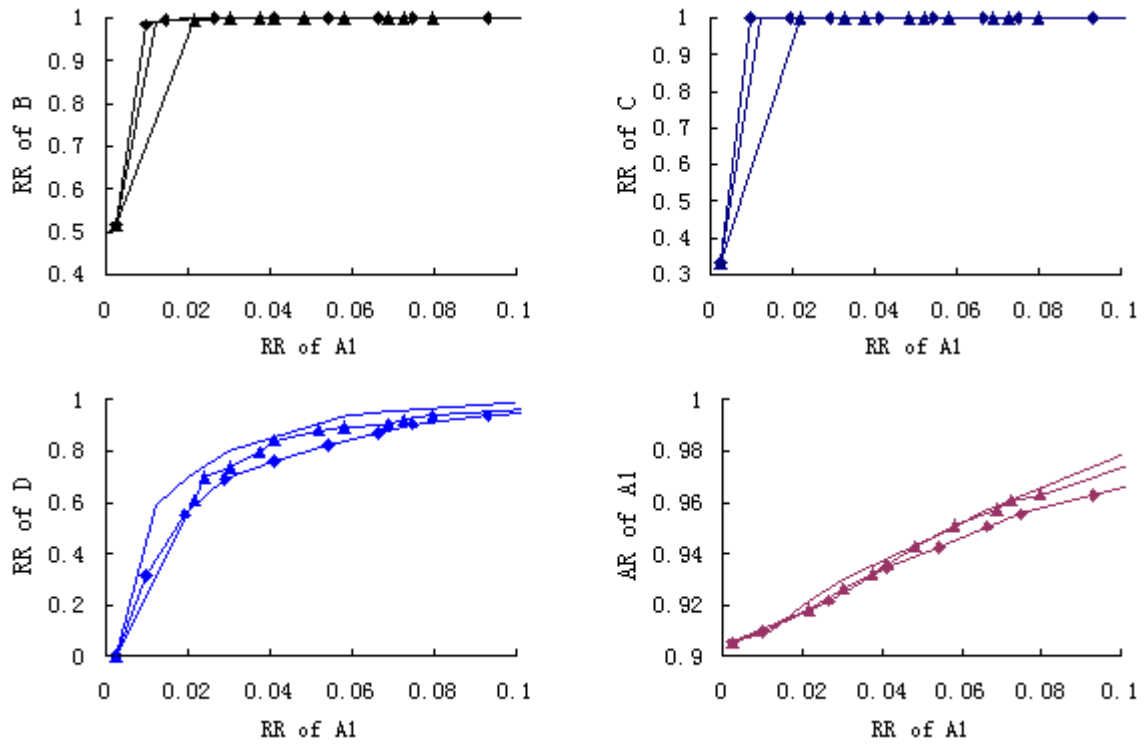


图 6-3

### 6.3 小结

本章研究了综合各种信息源提高验证性能的方法。基于规则的综合适合于意义明确的参数和显而易见的识别错误，给出的是关于识别假设正确性的硬判决。基于统计模型的综合需要额外的训练数据库以估计模型参数，但是统计模型易于优化，而且给出的是关于识别假设正确性的软判决，即置信度，易于调整工作点和结合需要置信度估计的其他应用。在实际使用中，往往是规则和统计模型同时使用互相补充的。在电话语音识别系统中，首先根据一些简单明确的信息（如合法识别结果数目，半音节长度极值，归一化音节长度方差等）将明显的错误拒掉；而对用于统计模型综合的信息源，也可以先设置一个较松的门限，拒绝掉明显的错误。只有当规则没有拒绝掉识别结果时，才动用信息源综合，估计置信度，做最终的判定。

本章提出采用 MLP 进行信息源综合，并比较了 MLP 与线性模型。由于 MLP 在性能上优于线性模型，而存储和运算代价又非常小，认为选择 MLP 是合理的。下表给出了综合四种信息源，MLP 估计置信度的在特定工作点的验证性能。工作点 0.2% 对应拒绝没有合乎规则的识别结果的情况。

表 6-2

工作点 (对合法语音的拒识率)		0.2%	5%	10%
MLP 综合 (VGOL)	AR 库 A1	90.6%	94.4%	97.2
	RR 库 B	51.5%	100%	100%
	RR 库 C	33.3%	100%	100%
	RR 库 D	0.2%	89.7%	99.1%
GMM 垃圾模型	AR 库 A1	90.6%	94.1%	95.7%
	RR 库 B	51.5%	100%	100%
	RR 库 C	33.3%	100%	100%
	RR 库 D	0.2%	80.0%	94.1%

## 第七章 结论

### 论文工作总结

置信度估计和说话验证对语音识别的实用化具有重要意义,而且是语音识别技术与其他智能系统(Intelligent System)结合的桥梁,因此具有很高的实用和学术价值。

本论文旨在为本实验室填补关于置信度与拒识工作方面的空白,期望验证与拒识能够使现有的识别系统/模块更加实用化。本论文探讨了评价说话验证手段,指出了误识与非法声响在验证任务中的不同地位,提出分别研究对它们的拒识。本论文还强调了在评价验证方法时,必须考虑具有不同性质的非法声响。在实验室已有的几个语音识别系统的基础上,论文较为全面的研究了汉语语音识别中的说话验证。在算法上,论文工作

- ✓ 研究了可资验证利用的信息源及其综合方法,提出了归一化音节长度方差和基于 MLP 的在说话一级的信息源综合和验证,取得了良好的效果;
- ✓ 研究了反词模型在汉语语音识别验证中的应用,根据汉语语音的特点,特别研究了基于反半音节模型的词表/任务无关说话验证;
- ✓ 提出 MLP/线性模型估计后验概率在汉语数码语音识别验证中的应用,这种方法的验证性能远远超过了常用的反词模型和前二选验证。
- ✓ 研究了基于垃圾模型和在线垃圾模型在词表/任务无关的说话验证,提出了高斯混合模型的垃圾模型,该模型在性能和训练速度上具有优势,便于实时的计算置信度。研究了改进在基于线垃圾模型验证的方法,从提高验证性能出发,提出直接计算半音节在线垃圾似然度,并从竞争集中去除模糊半音节;从提高速度出发,提出通过估计模型统计相似度来减小竞争集。

其中基于直接计算半音节在线垃圾似然度的验证方法已经在邮包语音校核系统的语音确认中使用,在实用中效果明显。

## 未来工作展望

在本论文工作中，由于实际条件和时间有限，在下面一些问题中留下了遗憾，希望未来的工作能够把它们弥补：

- ✓ 非法声响数据库太小，对开展对非法语音统计建模的验证方法研究限制很大。本论文因此没能涉及有关最小验证错误训练（Minimum Verification Error Training, MVE Training）方面的研究。
- ✓ 基于半音节的识别系统平台不够优化，特别是声学底层和搜索算法。对本实验室主要关心的拒绝非法声响问题，说话验证只有在识别系统具有一定的性能之后才能取得明显的效果。
- ✓ 基于HMM迹和MLP后验概率估计的验证方法可以应用到汉语连续数码语音识别中去，还可以把拒绝非法声响也考虑进去。对于基于半音节的系统，可以考虑MLP与反半音节模型的结合。
- ✓ 基于反词/反半音节模型的验证可以进一步改进。考虑到反词模型在多任务中的成功，有理由怀疑本论文为简化算法做出的假设。下一步的工作应该考虑优化反词/半音节模型的训练算法（包括采用MVE训练），优化反词/半音节模型的组成。
- ✓ 基于高斯混合的垃圾模型可以应用到识别系统搜索与剪枝中，用来剪掉错误的搜索路径和实时地拒绝掉非法声响。
- ✓ 希望能够确定关键词识别研究的长远计划，确定应用平台，采集相关语音库。有了平台和库，本论文研究的一些方法（垃圾模型，在线垃圾模型和反词/半音节模型等）很容易就可以把现有的半音节识别系统转化为关键词识别系统。

## 参考文献

- 江金涛 (1998). 用于用户交换机的电话语音识别系统的研究: [硕士学位论文] 北京: 清华大学电子工程系.
- 李虎生 刘润生(2000). 高性能汉语数码语音识别. 清华大学学报(自然科学版),40(1):32~34
- 李虎生(2000). 汉语数码串语音识别及说话人自适应: [硕士学位论文] 北京: 清华大学电子工程系
- 刘加 潘胜昔 江金涛 等(1998). 用 TMS320C31 实时实现电话语音识别系统. 清华大学学报(自然科学版), 38(1):51~54
- 潘胜昔 (1998). 电话语音识别的稳健性研究: [博士学位论文] 北京: 清华大学电子工程系
- 韦小东 朱杰 胡光锐(1998). 汽车噪声中自动语音的识别技术. 上海交通大学学报, 32(10):10~13.
- 徐明星 郑方 吴文虎 等(1998). 连续语音关键词识别系统的拒识方法研究. 清华大学学报(自然科学版), 38(S1):89~91
- 杨行峻 郑君理 (1992). 人工神经网络. 北京: 高等教育出版社
- 杨行峻 迟惠生 (1995). 语音信号处理. 北京: 电子工业出版社
- 张昊天(2000). 邮包校核语音识别系统的研究: [硕士学位论文] 北京: 清华大学电子工程系
- 赵庆卫(1998). 非特定人大词汇量连续语音识别系统的研究: [博士学位论文] 北京: 清华大学电子工程系
- 郑方 (1997). 连续无限制语音流中关键词识别方法研究: [博士学位论文] 北京: 清华大学计算机科学技术系
- 钟林 (1998). 人工神经网络小词表汉语语音识别: [本科毕业论文] 北京: 清华大学电子工程系
- Bartkowa, K and Jouvét(1997). D. Usefulness of phonetic parameters in a rejection procedure of an HMM Based speech recognition system. Proc. of EuroSpeech.
- Bickel, P.J. and Doksum, K.A.(1976). 数理统计. 第六章. 兰州大学出版社
- Boite, J-M, Boursard, H , D'hoore, B and Haesen, M(1993). A new approach towards keyword spotting. Proc. Of EuroSpeech, 1273-1276
- Boursard, H., D'hoore, B. & Boite, J.-M.(1994). Optimizing recognition and rejection performance in word-spotting systems. IEEE Proc. ICASSP, v.1, 373-376.
- Bouwman, G., Sturm, J. and Boves, L.(1999). Incorporating Confidence Measures in the Dutch

- Train Timetable Information System Developed in the ARISE Project. IEEE Proc. ICASSP
- Chen, T. And Rao, R.R.(1998). Audio-Visual Integration in Multi-modal Communication. Proceedings of the IEEE, v86,n5,837-852
- Colton, L.D.(1997). Confidence and Rejection in Automatic Speech Recognition. Ph.D. Dissertation, OGI
- Duda, R.O. and Hart, P.E.(1973). Pattern Classification and Scene Analysis. John Wiley, New York.
- Foote, J.T., Young, S.J., Jones, G.J.F. & Jones, K.S.(1997). Unconstrained keyword spotting using phone lattices with application to spoken document retrieval. Computer Speech & Language,11,207-224.
- Hofstetter, E.M. & Rose, R.C.(1992). Techniques for task independent word spotting in continuous speech messages. IEEE Proc. ICASSP, v.2,101-104.
- Hopper, A.(1990). Pandora-An Experimental System for Multimedia Application. Operating Systems Review, v24, n2.
- James, D.A. & Young, S.J.(1994). A fast lattice-based approach to vocabulary independent wordspotting. IEEE Proc. ICASSP, v.1 , 377-380.
- Jeanrenaud, P., Siu, M., Rohlicek, J.R. et al (1994). Spotting events in continuous speech. IEEE Proc. ICASSP, v.1 ,381-384.
- Jitsuhiro, T., Takahashi, S. & Aikawa, K.(1998). Rejection of out-of-vocabulary words using phoneme confidence likelihood. IEEE Proc. ICASSP.
- Jouvet D, Bartkova K and Mercier G(1999). Hypothesis Dependent Threshold Setting for Improved Out of Vocabulary Data Rejection. IEEE Proc. ICASSP
- Kawahara, T., Lee, C.-H. & Juang, B.-H.(1997). Combining key-phrase detection and subword-based verification for flexible speech understanding. IEEE Proc. ICASSP,1159-1162.
- Kawahara, T., Lee, C.-H. & Juang, B.-H.(1998). Flexible Speech Understanding Based on Combined Key-Phrase Detection and Verification. IEEE Trans. On SAP, v.6, n.6, 558-568.
- Kemp, T. & Jusek(1996), A. Modeling unknown words in spontaneous speech. IEEE Proc. ICASSP, v.2, 530-533.
- Koo, M.-W., Lee, C.-H. & Juang, B.-H.(1998),.A new decoder based on a generalized confidence score.IEEE Proc. ICASSP.
- Leung, L-K and Fung, P.(1999). A More Efficient and Optimal LLR for Decoding and Verification. IEEE Proc. ICASSP

- Lippmann, R.P., Chang, E.I. & Jankowski, C.R.(1994). Wordspotting training using figure-of-merit back propagation. IEEE Proc. ICASSP, v.1, 389-392.
- Lleida, E. & Rose, R.C.(1996). Efficient decoding and training procedures for utterance verification in continuous speech recognition. IEEE Proc. ICASSP, v.2, 507-510
- Manos, A.S. & Zue, V.W.(1997). A segment-based wordspotting using phonetic filler models. IEEE Proc. ICASSP, 899-902.
- Mathan, L. & Miclet, L.(1991). Rejection of extraneous input in speech recognition application using MLP's and the trace of HMM's. IEEE Proc. ICASSP, v.1, 93-96.
- Matsunaga, S and Sakamoto(1996), H. Two-pass strategy for continuous speech recognition with detection and transcription of unknown words. IEEE Proc. ICASSP, v.2, 538-541.
- Modi, P. and Rahim, M(1997). Discriminative Utterance Verification Using Multiple Confidence Measure. Proc. Of EuroSpeech.
- Riccardi, G., Gorin, A.L., Ljolje, A. And Riley, M.(1997). A Spoken Language System for Automatic Call Routing, IEEE Proc. ICASSP, 1143-1146
- Rahim, M.G., Lee, C.-H. & Juang, B.-H.(1997). Discriminative Utterance Verification for Connected Digits Recognition. IEEE Trans. On SAP, v.5, n.3, 266-277.
- Rahim, M.G. & Lee, C.-H.(1997). String-based minimum verification error(SB-MVE) training for speech recognition. Computer Speech & Language,11,147-160
- Richard, M.D. and Lippmann, R.P. (1991). Neural Network Classifiers Estimate Bayesian *a posteriori* Probabilities. *Neural Computation*,vol.3, 461-483.
- Rivlin, Z., Cohen, M., Abrash, V. & Chung, T.(1996). A phone-dependent confidence measure for utterance rejection. IEEE Proc. ICASSP, v.2, 515-518.
- Rohlicek, J.R., Ayuso, D., Bates, M. et al(1992). Gisting Conversational Speech. IEEE Proc. ICASSP, v2, 113-117
- Rohlicek, J.R., Jeanrenaud, P., Ng, K. et al.(1993). Phonetic training and language modeling for word spotting. IEEE Proc. ICASSP, v.2, 459-462.
- Rose, R.C. & Paul, D.B.(1990). A HMM based keyword recognition system. IEEE Proc. ICASSP, v.2, 129-132.
- Rose, R.C., Chang, E.I. and Lippmann, R.P.(1991). Technique for Information Retrieval from Voice Messages. IEEE Proc. ICASSP, 317-320
- Rose, R.C.(1992). Discriminant wordspotting techniques for rejecting non-vocabulary utterance in unconstrained speech. IEEE Proc. ICASSP, v.2, 105-108.
- Rose, R.C.(1993). Task independent wordspotting using decision tree based allophone clustering. IEEE Proc. ICASSP, v.2, 467-470.

- Rose, R.C.(1995). Keyword detection in conversational speech utterance using HMM based continuous speech recognition. *Computer Speech & Language*,9,309-333
- Rose, R.C., Yao, H., Riccardi, G. and Wright, J.(1998). Integration of utterance verification with statistical language modeling and spoken language understanding. *IEEE Proc. ICASSP*
- Sukkar, R.A. & Wilpon, J.G.(1993). A two pass classifier for utterance rejection in keyword spotting. *IEEE Proc. ICASSP*,V2,451-454.
- Sukkar, R.A.(1994). Rejection for connected digit recognition based on GPD segmental discrimination. *IEEE Proc. ICASSP*,V1,393-396.
- Sukkar, R.A. and Lee, J.-H.(1996). Vocabulary Independent Discriminative Utterance Verification for Nonkeyword Rejection in Subword Based Speech Recognition. *IEEE Trans. On SAP*,v.4,n.6,420-429.
- Sukkar, R.A., Setlur, A.R., Rahim, M.G. & Lee, C.-H.(1996). Utterance Verification of keyword strings using word-based minimum verification error(WB-MVE) training. *IEEE Proc. ICASSP*,V1,518-521.
- Sukkar, R.A.(1998). Subword-based minimum verification error(SB-MVE) training for task independent utterance verification. *IEEE Proc. ICASSP*
- Villarrubia, L. & Acero, A.(1993). Rejection techniques for digit recognition in telecommunication application. *IEEE Proc. ICASSP*, v.2, 455-458.
- Weintraub, M., Beaufays, F., Rivilin, Z, Konig, Y. & Stolckes, A.(1997). Neural-network based measures of confidence for word recognition. *IEEE Proc. ICASSP*.
- Williams, G. and Renals, S.(1999). Confidence Measures from Local Posterior Probability Estimates. *Computer Speech and Language*, 13, 395-411
- Wilpon, J.G., Rabiner, L.R., Lee, C.-H. & Goldman, E.R.(1990). Automatic Recognition of Keywords in Unconstrained Speech Using HMM's. *IEEE Trans. On ASSP*, v.38,n.11,1870-1878.
- Young, S(1994). Detecting Misrecognition and Out of Vocabulary Words. *IEEE Proc. ICASSP*, v.2,21-24
- Zue, V. et al(2000). JUPITER: A Telephone-Based Conversational Interface for Weather Information. *IEEE Transactions on Speech and Audio Processing*, Vol. 8 , No. 1, January 2000.

## 作者论文工作期间发表和拟发表论文

- [1]. Lin Zhong, Yuanyuan Shi and Runsheng Liu(1999).A Dynamic Neural Network for Syllable Recognition. the *Proceedings of International Joint Conference on Neural Networks*, Washington D.C., 1999.
- [2]. Lin Zhong and Runsheng Liu(1999). Training Strategies for a Dynamic Neural Network with Application in Speech Recognition. the *Proceedings of International Symposium on Signal Processing and Intelligent System (ISSPIS)*, Guangzhou, China, 537-540.
- [3]. 钟林 刘润生(2000). 一种新人工神经网络结构及其在数码语音识别中的应用. 清华大学学报（自然科学版）40(3):104-108
- [4]. 钟林 刘加 刘润生（2000）. 多层前向感知机汉语孤立数码语音识别. 电路与系统学报, 已接受。
- [5]. Lin Zhong, Jia Liu and Runsheng Liu(2000). Rejection Based On *a Posteriori* Probability Estimated By MLP With Application For Mandarin Voice Dialer On ASIC. Accepted by *IEEE ICASSP 2000*, Istanbul.
- [6]. Lin Zhong, Jing Liu, Jia Liu and Runsheng Liu. Improving Task Independent Utterance Verification Based on On-line Garbage Phoneme Likelihood. Submitted to International Conference on Spoken Language Processing 2000, Beijing
- [7]. Lin Zhong, Jia Liu and Runsheng Liu. Utterance Verification with Gaussian Mixture Models. Submitted to International Conference on Spoken Language Processing 2000, Beijing

## 附录

### A 非法声响库

库 B 包括	糟糕 好 好的 好哇 什么 怎么? 怎么回事! 胡说! 是吗? 恩--- 啊--- 呵呵 唉 嘿嘿 etc
库 C 包括	咳嗽 喘气 轻轻吹气 清嗓子 咂嘴 哼鼻子 咳嗽 清嗓子 轻轻咂嘴 etc
库 D 包括	邮包语音核对系统路单测试语音 (张昊天, 2000)

## B 根据模型相似度得到优化的韵母竞争集

韵母号	韵母	竞争集
0	i(1, ch,sh,zh)	i(2) i(3) i(4) e
1	e	ei en ve ie i(1) i(2) i(3) i(4)
2	u	v ui un uei uen uo uan
3	en	eng un in uen an
4	ai	an uai a ei ui
5	i(2,c,s,z)	i(1) i(3) i(4) e
6	iou	io iu ou iao
7	ing	in eng iong un vn ang ong
8	uan	an uang van uai ian ua
9	ong	ang eng iong ing un vn
10	uo	io u o ao ui
11	in	ing en un vn ian uen
12	ia	a ian ie ua io iu iou
13	iang	ian ang uang ing iong
14	ian	iang uan ia van
15	ei	uei ui e en ai
16	er	e a ie i(3) ei
17	eng	en ong ing ang ueng uen
18	uei(wei)	ei ui uen uai
19	i(3,r)	i(1) i(2) i(4) e er
20	ui	uei ei u un ua
21	uai	ai ua uan uei ui
22	uang	uan ueng ang iang
23	an	ang uan ai van ian
24	a	ai er o an ua ia
25	ang	an uang eng ong iang
26	ou	iou iu o ong
27	van	an uan v ve ian
28	vn	un v van in
29	iu	iou io ie ia
30	iao	ao ia iou io
31	ua	a uan uai ia ui
32	ve	v e ie vn i(4)
33	v	u ve vn i(4) e
34	ie	e i(4) ve iu ia
35	ao	ou an ai iao a
36	uen	ueng uei un uan in
37	un	uen ui vn in en
38	o	uo ao io a ou
39	iong	ong ing un iang
40	ueng	uen uang eng un
41	io	o iao ao i(4) ie ia iu
42	i(4)	i(1) i(2) i(3) ie in ia io iu

