口语对话系统中语音识别的研究

张国亮

Research on Speech Recognition in Spoken Dialogue System

Dissertation Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the degree of

Doctor of Engineering

by

Guo-liang ZHANG

(Computer Science and Technology)

Dissertation Supervisor: Professor Wen-hu WU

Associate Supervisor: Associate Professor Fang ZHENG

October, 2003

摘　要

近年来，口语对话系统的研究得到广泛重视，其中语音识别算法的性能直接影响整个系统的性能，但在某种意义上，现有语音识别算法的性能不能满足需要。为了提高口语对话系统中语音识别性能，本文从识别算法和语音确认两个主要方面出发，在关键词搜索性能、语境知识对关键词识别的指导以及新语音确认特征诸方面进行了研究，提出如下方法、策略：

1. 关键词动态确认。针对关键词的漏检错误大多数都与另一错误关键词的误警错误有关这一现象，提出关键词动态确认的概念。在搜索过程中引入虚拟 OOV 模型来对产生的关键词候选进行确认，及早地将不正确的关键词候选剪除，从而避免其对正确关键词候选的干扰影响。实验表明：在误警率相同的条件下，误识率下降了约 10%。

2. 语境知识指导下的关键词识别策略。针对目前的各种语音识别方法在对话系统中性能不佳的现实，为改善对话系统的整体性能，提出对话语境知识指导下的关键词识别策略：利用对话管理器给出期待焦点信息，确定对应焦点下的活动词表、活动规则集，生成相应的识别自动机，并用其来指导关键词识别。实验表明：语境知识指导下的关键词识别算法具有很高的识别性能和鲁棒性，基本能够满足口语对话系统的需要。

3. 上下文相关语音确认策略。通过分析识别结果中互相干扰现象的发生，发现待确认结果前后序词的确认度会对待确认词本身的确认度有一定的指示作用，提出上下文相关语音确认策略：引入上下文知识来指导语音确认，改善语音确认的性能。利用待确认词与其前后序词的确认特征组成待确认词的上下文相关确认特征，体现出识别结果中的前后序词对当前待确认词的影响。实验表明：该算法的拒识性能明显好于传统语音确认算法。

关键词：关键词识别，语音确认，语音识别，口语对话系统

Abstract

In order to overcome some difficulties in the task of Speech Recognition in spoken dialogue systems, some new methods and new strategies, which are founded on recognition algorithm and speech verification, are proposed in this dissertation involving the keyword search algorithm, dialogue context guided on keyword recognition and the new features for speech verification:

1. Keyword dynamic verification. Based on the fact that the miss error of a keyword is mostly due to the false alarm of another keyword, virtual OOV model is proposed and adopted to verify the hypothesis immediately in dynamic programming and to prune the wrong hypothesis as soon as possible, so as to avoid the wrong hypothesis from having bad influence on the right hypothesis. Experimental results show the keyword error rate is reduced by 10% at the same false alarm rate by using the keyword dynamic verification technique.

2.Recognition strategy directed by the dialogue context knowledge. In order to improve the recognition performance, a keyword recognition strategy with the dialogue context embedded in it is proposed, including the expected focus given by the dialogue manager, the active lexicon/rules determined by the understander, and the recognition automata passed to the recognizer to direct the keyword recognition. The strategy can achieve high recognition performance and good robustness proved by the experimental results, which shows that the strategy will work very well in spoken dialogue system.

3. Context dependent speech verification strategy. Though the analysis of interaction effect of speech recognition results, it is found that the confidence of the previous word and that of the following word will be a guide to the confidence measure of the middle word, in this way, the context dependent speech verification strategy will be concluded. A method is

adopted that is integrating the confidence features of the middle word and its context words into the context dependent confidence features of the middle word, which reflects the context influence in speech verification algorithm. Experimental results show the false rejection performance of the new strategy is obviously better than the traditional context independent speech verification strategy.

Keyword：Keyword Recognition, Utterance Verification, Speech Recognition, Spoken Dialogue System

目　录

第一章　绪　论

语言作为人类最重要最自然的交流工具，是人类获得信息的最重要的来源之一，让计算机能“听懂”人类的语言，也是人与计算机之间进行沟通的最方便的形式之一。用语音来实现人与计算机之间的交互，包括三项技术：语音识别、自然语言理解和语音合成。语音识别(Auto Speech Recognition)的主要任务是完成语音到文字的转变；自然语言理解(Natural Language Understanding)则是完成文字到语义的转换；语音合成(Speech Synthesis)是用语音方式输出用户想要的信息。随着计算机能力的迅猛提高，这三项技术得到了飞速发展，今天集语音识别、自然语言理解和语音合成三项技术于一身的口语对话系统(Spoken Dialogue System)受到了国内外研究机构的广泛关注和高度重视，其应用也必将带来很好的社会、经济效益。

目前一批研究或实际系统已经出现，常见的比如旅游信息查询、电话客票服务、语音呼叫中心、天气预报信息查询等。作者所在的课题组从一九九九年起，开始对话系统方面的研究与实现，目前有两个系统原型正在研究中，一个是“航班信息系统 *EasyFlight*”，另一个是“语音自动总机”。构建一个完善的口语对话系统，需要应用语音信号处理、语音识别、语言理解、知识表示、对话管理、文语转换和语音合成等多项技术。本论文的研究工作主要集中在口语对话系统中语音识别方面的研究工作。

本章的内容安排如下：首先对口语对话系统进行简介并简述其诸方面的研究现状，然后介绍对话系统条件下语音识别研究所面对的困难和研究现状，最后是论文的研究工作整体描述和组织。

1.1 口语对话系统简介及研究现状

1.1.1 口语对话系统简介

口语对话系统，可以简单地定义为：以自然语音为输入输出接口，通

过与用户进行交谈，实现自动服务的系统，其最终目的是让计算机与人可以像人与人一样自由的用口语语言进行交谈，并完成用户希望计算机所做的各种任务。要达到这些目标，就要求口语对话系统首先能够听懂用户所说的语言，并且理解用户所说的内容，最后智能地用自然语言的方式进行回答。所以口语对话系统主要涉及到计算机人工智能领域中语音识别、自然语言理解、语音合成三个重要领域，属于三者的交叉学科，可以说是这三个领域最新研究成果和具体应用的有机结合体，是语音识别技术、自然语言理解技术和语音合成技术走向实用化的有机载体。

口语对话系统的输入输出为自然语音，这里的自然语音是相对于朗读式语音而言的，它是指连续非朗读式的、自然的口语或对话发音。基于标准发音、标准口音、朗读式的关键词识别系统和连续语音识别系统在过去的几十年里，已经从理论研究、技术实现和推广应用方面取得了巨大的成功，但上述限制条件也制约了其应用的进一步发展。人们期待更为自然和方便的人机交互方式，提出了基于自然语音或电话语音的听写录入、信息查询、预约、订票、导游等新的需求，并在这些方面进行大量的研究和开发工作。在这种大环境下，面向自然语音的口语对话系统也就成为当前语音领域中的研究热点。在国外，有专门的研究计划开展这项研究，像美国的 TRAINS 计划[1]和 DARPA-Communicator 计划[2]、欧洲的 ARISE 计划[3]、REWARD 计划[4]、VERBMOBIL 计划[5]等，有许多著名的学府和研究机构从事这项研究，比如 MIT 的 SLS 实验室[6]、CMU 大学[7]、Lucent-Bell 实验室[8][9]、OGI 的 CSLU 中心[10]、法国的 LIMSI[11][12]、德国的 Erlangen-Nuremberg 大学[13]、日本的 ATR 实验室[14]和 Philips 公司[15]等，在国内，清华大学[16]、中国科学院[17]、中国科学技术大学、香港大学、香港中文大学[18]、香港科技大学、台湾大学[19]等也都投入了相当大的精力开展这项研究。可以预计，在刚刚到来的二十一世纪，语音技术将得到前所未有的发展。

图 1-1 给出了基本的口语对话系统结构略图，图中表明通常必须含有四个主要功能模块：语音识别器、语言理解器、对话管理器和语音合成器，四个模块都以前一级模块的输出为输入，互相依赖缺一不可。

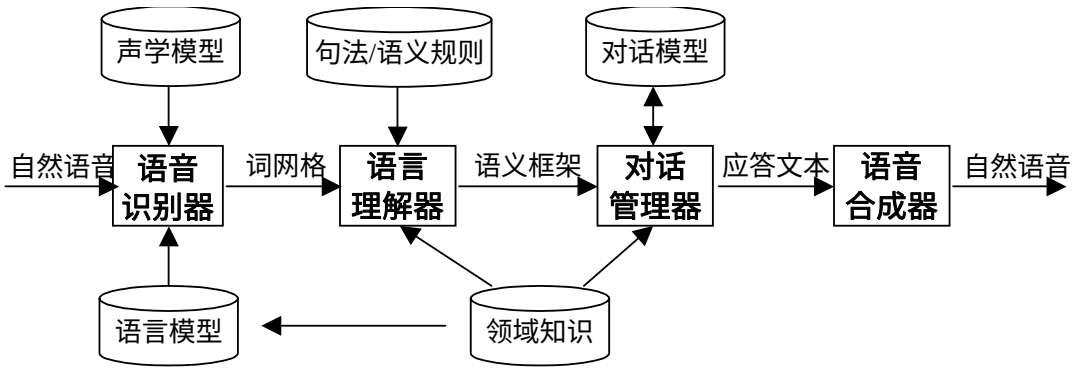


图 1-1 对话系统模块与模型略图

语音识别模块的主要目的是听清楚人的语言，即把人的语音转换成文字。类似于其它许多语音系统，例如听写机、语音导航系统等，这一模块是整个系统的关键核心，语音识别性能的高低将直接影响到系统的整体性能。语言理解模块的主要目的是理解人语言的内容，把用户的真实意思用计算机的内部方式表达出来。和一般的语言理解任务不同，在口语对话系统中，语言理解模块接受的输入通常是语音识别模块给出的关键词候选，可能出现多候选的情况，而且关键词候选可能出现误识和漏识的情况，这无疑将给语言理解模块提出很大的挑战。对话管理器也是口语对话系统中不可或缺的模块之一，它的任务是根据语言理解模块的理解结果以及对话的上下文语境、历史信息，进行综合分析以确定用户的意图，根据需要查询后台数据库，并组织适当的应答语句，以保证计算机与人的谈话可以有效并友好地继续下去，一直到用户的目的得以实现。最后，由语音合成模块将应答文字转换为语音再输出给用户。当前，无论在国内还是国外，语音合成的研究已经相对成熟，已经有投入市场的实用产品，其主要挑战在于使生成的语音更加自然与生动。有时根据具体的需要，口语对话系统可以不包括语音合成模块，可以直接由文本输出结果。

系统的运行往往还依赖于一些模型或数据库，比如是声学模型、语言模型、句法/语义规则、领域知识、对话模型和领域数据库等。它们都记录了口语对话系统运行中所必须的数据和知识，在系统运行时由各个模块进行调用。

1.1.2 口语对话系统的研究现状

本节将从几个不同的方面对口语对话系统当前的整体研究现状做一个简要综述。

宿主平台：口语对话系统构建依赖的平台，通常根据具体的应用有着不同的表现形式。例如 Hugunin et al[20]设计的基于 Microsoft Excel 软件的嵌入式电子表格系统；Issar 在[21]中描述的用于在 WWW 页面上填充表格的语音接口；移动式办公室机器人 Jijo-2[22]，它们都是口语对话系统在不同应用背景下的具体表现。如今，随着大量公有信息的出现(订票、信息查询等)以及电话通讯的广泛普及，越来越多基于电话的口语对话系统面世，比如欧洲的自动铁路信息研究计划[3]。基于电话的系统应用前景广阔，有很好的社会效益和经济效益；其技术挑战主要在于电话信道的窄带特性、信道之间的差异以及现实世界中的噪音问题。

语音识别：语音识别算法根据口语对话系统的需求也可分为多个类型。早期的系统灵活性需求较低，采用孤立词或模板匹配算法即可获得较好的识别效果。现今，随着人们对口语对话系统输入自然语音的灵活性要求越来越高，语音识别算法中的各个模块：特征提取、声学模型训练、搜索算法和语音确认等都面临着前所未有的挑战。而且，因为口语对话系统通常都有很强的应用背景，这就更加要求语音识别算法在各种实际的应用环境下仍然能够稳定的工作。在本章的节 1.3 将详细介绍口语对话系统中语音识别算法的研究现状。

语言理解及语义分析：语言层的处理，包括获取句法结构、语义表示，解决自发语句中的口语现象和指代(anaphora)消解。一般结构比较严谨的形式语言通常可以用基于规则的上下文无关文法进行表述，通过自顶向下法或自底向上法来进行分析。众所周知，与书面语不同，口语对话系统中用户的语句是很随意的，其中充满了垃圾、碎片、犹豫、纠正、重复、指代(anaphora)、省略(ellipsis)、词序混乱和病句等现象，这些现象均是口语对话系统中的语言理解必须面对的一些难题。所以，理解的鲁棒性是现今研究工作关注的焦点。

Sasajima 在开发口语对话系统框架 EUROPA[23]时，采用了关键词检出策略，在识别得到的关键词网络(lattices)上，使用一种高效的 BTH[24]分析器进行实时分析，口语现象中的词序混乱现象可以被很好处理。杨开城[25]提出句法分析应该是一种实体推理，增强语义信息是实现句法分析实体推理的有效手段。利用基于词的兼类处理规则，大大提高了句法分析的效率；利用词静态和动态句法语义特征来限制句法规则过强的生成能力，取得了较好的效果。在 TRIPS9 8[26]中，一种称为谈话上下文(Discourse Context)的数据结构被提出用来保存指代对话历史中的前辈候选(candidate antecedents)以解决指代消解问题。Boros[27]提出的短语检出(phrase spotting)的方法是一种部分分析(partial parsing)技术，在解决自发语音中的口语现象和识别错误等问题方面表现了很好的鲁棒性。Wang 在[28]提出了 LEAP(Language Enabled Application)的概念，它建立在语义类之上，不仅能有效地分析病态(ill-formed)语句，而且对非明确(under-specified)语法也有很好的效果。

对话管理：对话管理也是口语对话系统的核心，它使用对话模型来描述对话状态，决定对话状态的转移和上下文语境下的应答。对话管理面临的一个主要问题是如何利用恰当的确认策略(confirmation strategies)和混合主导的方式，提高对话效率和用户的满意度。

Denecke[29]提出交谈目标(communicative goal/CG)的概念，认为用户与机器交互的目的是达成一个交谈目标。Zanten 在[30]中描述了一个自适应的对话管理方法，它使用一种层级槽结构(hierarchical slot structure/HSS)描述对话，提示问题在这种情况下也是层级结构的，避免用户被动地逐个填写槽值的乏味(rigid)过程。这种方法具有比较强的灵活性，它的自适应表现使得对话更有效和更富智能。Papineni[31]介绍了一种基于表格(forms)的混合主导的对话管理方法，该文认为一个对话过程由若干任务构成，可以用一个表格对应一个任务的形式来描述对话状态，而整个应用可以用一个表格集合来描述。动态调整可容许的表格列表，可以在系统主导和混合主导之间进行切换。Lin 在[32]中使用推理树来表示对话主题需要的所有信息，一个主题对应一个推理树，并最终生成一个自动机来描述主题信息。随着

对话的进行，主题状态在自动机的状态间进行切换。Pargellis[33]探索了自动对话管理生成的用户定制的方法。该文设计了一个自动对话产生器(Automatic Dialogue Generator)，可以根据用户的任务描述表格，自动生成一个有限状态对话管理模型。

语音合成：当最终应答文本由对话管理器生成后，要通过语音合成来生成最终的应答语音，播放给用户，使用户感觉到彷佛是与人交谈而不是计算机。现阶段，国内外均有很多语音合成的实用产品，这些产品基本都可以将文字转化为清晰并流利的语音。其主要挑战在于如何使生成的语音更加自然生动、需要的语音数据库更小、成本更低。

1.2 自然语音：语音识别的难点

已有实验表明，对于采用标准发音的、仔细朗读的语音，传统的连续语音识别器一般都能够达到很高的词或音节识别率。然而，对于无准备的、自然随意的发音，这些识别器的性能却会急剧下降[34]，这在很大程度上是由于自然语音本身的特殊性决定的。

在声学层面，自然语音往往包含了多变的语速、语气、韵律和真实的情绪，以及严重的协同发音，这就会造成大量的音素级的插入、删除和替换现象。此外，不同的人具有不同的口音背景和发音习惯，因此即使说话人努力去按照标准的读音去发音，实际的音素序列也不会完全相同[35][36]。汉语相对于西方语言来说，又具有其特殊性。汉语的发音是基于标准音节的声韵结构（包括零声母现象），这种结构非常短，因此在自然发音中很容易受到口语上下文的影响而发生畸变[37]。

在语言层面，自然语音通常伴随着大量的口语现象，例如垃圾、碎片、犹豫、纠正、重复、指代、省略、咳嗽和吹气等现象，这些口语现象在日常生活人与人交谈中普遍存在。虽然这些口语现象主要由自然语言理解模块来进行处理，但由于现今存在的系统中多使用一定的语言知识来指导语音识别，所以口语现象的发生实际上也会影响语音识别模块的总体性能。表 1-1 给出了若干汉语自然口语现象的例子，从中可以初步体会到自然语

音识别的复杂性。

表 1-1 汉语自然语言中的口语现象

序号	口语现象	原始文本	包含口语现象的文本
1	咳嗽	今天五点的飞机	今天<咳嗽>五点的飞机
2	呼气	今天五点的飞机	<呼气>今天五点的飞机
3	垃圾	今天五点的飞机	今天“那个”五点的飞机
4	重复	今天五点的飞机	今天五点、五点的飞机

由于口语对话系统的应用背景复杂，导致输入的自然语音的语音质量不高，通常还伴随着一定的环境噪音。而且这些负面影响往往和对话系统的领域相关，不同口语对话系统可能出现的噪音类型也不尽相同，这也给自然语音的识别增添了难度。例如，在嵌入式设备上的口语对话系统中，录音设备与一般麦克风的录音质量相差很大；使用电话作为输入设备的口语对话系统，同样也面临着电话信道畸变、移动电话和固定电话信道特点不同以及户外使用电话时背景噪音大等一系列问题的困扰。

总之，自然语音识别是语音识别技术的主要难点之一，也是口语对话系统的主要难点。它给全世界的语音研究工作者都提出了巨大的挑战。

1.3 口语对话系统中语音识别的研究现状

1.3.1 特征提取

特征提取是语音识别的重要环节，现有的语音识别系统中以 Mel 频率倒谱系数(MFCC，Mel-Frequency Cepstrum Coefficient)[38]为主，配合上能量参数以及它们的一阶、二阶差分组成一个多维的特征向量。因为口语对话系统往往要面临很多信道畸变、环境噪音等实际应用中出现的问题，所以特征去噪必不可少。比较经典的去噪算法为倒谱均值归一化(CMN，Cepstral Mean Normalization)算法[39][40]，该算法有很好的适应性，可以很好地处理通常可能出现的环境噪音。但当环境噪音具有明显特点或信噪比较小时，需要使用更为复杂的去噪算法。

1.3.2 声学模型建模

声学模型建模的基础是识别基元的确定和大批量高标准的训练数据。在英文口语对话系统中常见的识别基元为音素和词，而在汉语口语对话系统中音素、声韵母、半音节、音节和词最为常见。为了尽可能地取得好的声学模型，绝大多数说话人无关(Speaker Independent)的系统都采用上下文相关(Context Dependent)的声学模型建模法。声学模型采用半连续或连续隐含马尔可夫模型(HMM，Hidden Markov Model)拓扑结构，用状态之间的驻留和转移来描述语音样本的产生。在汉语中，有 35 个音素、57 个声韵母、400 个无调音节和 1300 多个有调音节，当使用上下文相关建模法时会遇到数据稀疏问题，通常采用状态共享(State Tying)法使多个状态共享训练数据，有效地解决了数据稀疏问题。

因为口语对话系统中要处理的自然语音往往与朗读式语音区别较大，Zheng[41]定义了更加全面地覆盖汉语中可能出现的音变现象的广义声韵母集，并提出了基于广义声韵集的精细建模方法，有效提高了对自然语音的识别正确率。今天，自然语音的发音建模仍旧是一个有待探索的研究课题，它的研究成果将直接关系到口语对话系统中语音识别性能的好坏。

1.3.3 识别策略

三种主要识别策略在口语对话系统中被采用：有限状态语法、基于小领域 N-Grams 语言模型的连续语音识别和关键词检出。有限状态语法和 N-Grams 语言模型都用来描述语言知识信息，两者区别在于前者是通过专家总结出来的，而后者是从训练数据中统计出来的。许多系统使用基于 N-Gram 的连续语音识别策略，它们的差异往往表现在词表和 N-Gram 的规模上。在 August[42]中，Gustafson 使用了规模为 500 的词表，以及基于 70 个词类和 229 个词类对的词类二元文法 *bigram*。旅游信息系统 LOADSTAR[17]使用了词和词类的混合模型，词类的规模为 733。相对基于词的模型来说，词类语言模型的物理(句法)意义更明显，而且能够避免出现数据稀疏问题。

关键词/关键短语的识别策略也可用于对话系统，其特点在于不考虑输入语音中与系统领域无关(或影响不大)的语言成分，提高语义要素的识别率。因为其具有忽略关键词以外的语音部分，很多口语现象可以被忽略掉而不影响关键词的识别性能，所以现有系统多采用此策略。

能否将上述三种识别策略结合在一起，取长补短，发挥各自的优点，以提高系统的性能。本论文的第四章将介绍作者在这方面的研究成果。

1.3.4 语音确认

语音确认步骤在口语对话系统中不可或缺，语音确认给出的置信度得分不仅能有效地去除错误候选，而且还可以为后续的语言理解提供信息。现有的置信度特征主要分为三类：声学层面、词图层面、语义层面。声学层面通常使用词级的后验概率、搜索空间路径条数、词候选驻留时间等特征；词图层面通常使用候选词图中和词候选并列的其它候选的个数等特征；语义层面要结合语言理解，根据语言理解的结果对整句候选进行确认。现有的系统中置信度模型主要采用 fisher 线性分类器、人工神经元网络、决策树，近年来随着支持向量机(Supported Vector Machine)研究的兴起，有的系统中支持向量机也被用作置信度模型，这几种分类器都可以联合多路置信度特征并给出最终的确认结果，但也都需要大量的训练数据来训练模型内的各种参数。

由于对话状态不是一成不变的，因此各种声学参数可以根据当前对话所处状态进行动态调整。Lopez-Cozar et al [43]描述了一个电话快餐预定系统，在该系统中，他们使用了一种 Adaptive Confidence Threshold 的策略，其主要特点是，在交谈过程中可以根据环境条件的变化，动态调整置信度阈值。这样做，可以提高对话中的句子理解率(Understanding Rate)和减少对话回合(turn)数。

众所周知：在语音识别中，搜索空间的大小决定着识别率的高低，搜索空间越小识别率越高。有些口语对话系统中采用了一定的策略保证搜索空间在一定的规模之下而取得比较好的识别性能。PADIS-XL[44]是一个大规模的自动地址名录信息系统，人名和街道名的识别在该系统中具有核心

地位。在姓名数达到 35,000 的情况下，Seide 使用了一种词典切换(lexicon-switching)的策略，即根据当时的对话任务使用不同的词表，而且参照先前回合的识别结果对当前词表作进一步的限定；这样，词表的规模在一个完整的对话进程中不断减小，不仅降低了误识率，而且节省了搜索空间，保证了系统的实时响应。

另外有些文献提到在使用传统的基于 HMM 的识别方法的基础上，应用韵律(prosody)[45]或语音基频轮廓(F0 Contour)[46]等其它语音特征来提高识别性能。针对自发语音中的口语现象，有些文献在语音修改(Speech Repairs)方面提出了比较系统的方法，比如 Spilkery[47]认为语音修改可以根据声学/韵律线索、词碎片、编辑项和句法/语义异常等四种特征来定位。

1.4 背景口语对话系统介绍

作者所在的课题组从一九九九年起，开始口语对话系统方面的研究与实现，目前有两个系统原型正在研究中，一个是“航班信息系统 *EasyFlight*”，另一个是“语音自动总机 *dEar-Attendant*”。本文的研究工作均是以这两个口语对话系统为背景进行的，以解决这两个口语对话系统中在语音识别方面出现的实际问题为目标。下面对这两个口语对话系统进行简要介绍。

航班信息系统 *EasyFlight* 是清华大学“985”重点研究课题之一，它的目的在于以语音为接口，通过公用电话网络，向大众提供航班信息查询及订票服务。系统的数据库中预先存放国内每条航线、每次航班的各种信息，用户通过电话进行查询感兴趣的航线和航班信息，最终通过和计算机的多次信息交互完成订票业务。这包括以下条件的确定：(1) 起点城市，(2) 终点城市，(3) 出发时刻，(4) 到达时刻，(5) 航班号，(6) 机型，(7) 票数，(8) 身份证号码；数据库要求得到以下信息：(1) 有无航线，(2) 有无航班，(3) 航班时间信息，(4) 机型信息，(5) 有无余票及票数，(6) 价格。此口语对话系统采取用户主导和系统主导相结合的方式，其语音识别的主要难点在于：一、电话信道上的信道畸变与信道噪声；二、对话中的口语

现象；三、大量的时间、时刻、航班号、身份证号码等数字和字母的识别；四、用户说话主题的不规律跳跃。

语音自动总机 *dEar-Attendant* 相对于 *EasyFlight* 功能比较简单，主要是完成电话语音自动接线员的任务。该口语对话系统也是通过电话接收用户的输入，用户说出被拨叫人的姓名，对话系统识别并予以确认后，自动接通用户和被拨叫人的电话。如果数据库中有同名的人名时，系统会根据附加信息提示用户予以确认。系统认为用户总是配合的，输入语句中肯定包含一个被拨叫人的姓名，所以在进行识别时只关心关键词部分即被拨叫人人名的识别，其余部分不予考虑。虽然这个口语对话系统非常简单，几乎没有涉及到自然语音理解的内容，但其在口语语音识别模块的可研究性和 *EasyFlight* 系统没有本质区别，而且因为其对话比较简单，所以更容易实用化、产品化。

1.5 研究工作概述

本论文将研究焦点定位于口语对话系统中的语音识别算法，针对口语对话系统中语音识别需面对的语音质量差、语句的自发性、随意性、有识别错误等特点，本文尝试用语言理解指导语音识别的思路来解决口语对话系统中口语现象带来的识别困难，并在关键词检出、识别框架和语音确认三个方面都提出了一些行之有效的新思路和新方法。

1.5.1 研究目标

1) 分析关键词检出算法中存在的问题和汉语口语的特点，提出描述能力更强的补白模型来提高关键词检出算法的识别性能。

2) 面对目前实际应用中语音识别性能不佳的现实，尝试将对话语境知识运用于识别的机制，以提高小领域系统下语音识别的准确性。

3) 提出新的置信度特征和新的置信度模型，使得语音确认结果更加准确，有效剔除语音识别结果中置信度较低的候选，以保证向后续语言理解模块提供质量更可靠的语音识别结果。

4) 探索小领域口语对话系统语音识别模块的构建机制，提供可移植性、结构化的一般构建方法。

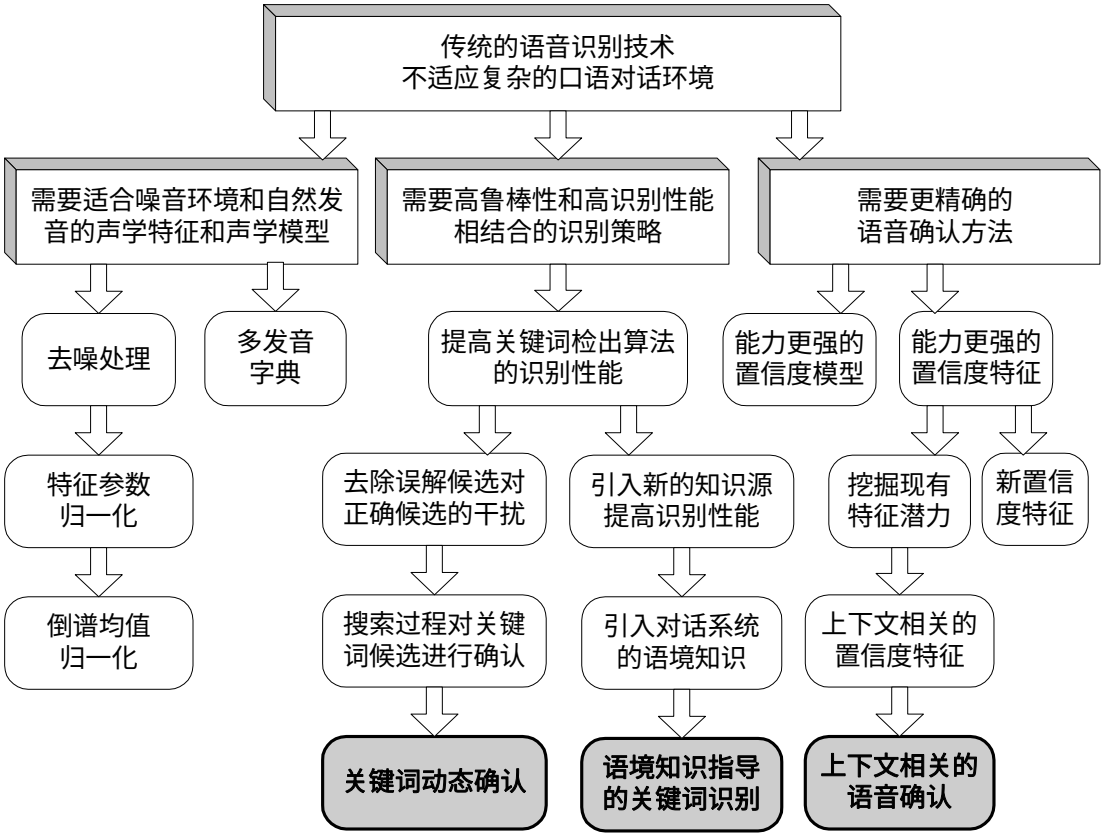


图 1-2 论文的研究思路

1.5.2 研究思路和研究内容

本论文的研究工作涉及口语对话系统语音识别领域的多个方面，图 1-2 从总体上反映了论文研究的思路和出发点，以及各部分主要工作之间的联系。

本论文在口语对话系统中的语音识别算法各个方面均进行了广泛的探索和认真的研究，提出若干新思路并通过实验证明了其有效性，同时也为这方面后续研究积累了一定的经验。主要研究工作可以概括为如下三点：

1) 关键词动态确认技术

通过详细分析关键词检出算法的具体实验结果，发现关键词的漏检错误大多数都与另一错误关键词的误警错误有关这一现象，进而试图通过减少误警关键词候选的思路来提高关键词识别正确率。提出关键词动态确认技术，在搜索过程中引入关键词确认模型来对产生的关键词候选进行确认，及早剪除不正确的关键词候选，从而避免其对正确关键词候选的影响。实验表明：采用关键词动态确认技术后关键词检出算法的识别性能有明显提高，在误警率相同的条件下，误识率下降了约 10%。

2) 语境知识指导下的关键词识别

现有的口语对话系统的语音识别算法以关键词检出、模板匹配和连续语音识别为主，关键词检出算法以其良好的鲁棒性而获得更多关注，但现有的关键词识别性能仍不能满足口语对话系统的需要。本文试图从使用语境知识来指导关键词检出这一思路出发来提高关键词检出算法的识别性能。提出语境知识指导下的关键词检出策略，用识别自动机来表述对话管理器给出的期待焦点下的语境知识，并用其来指导关键词检出。实验表明：语境知识指导下的关键词检出算法具有很高的识别性能和鲁棒性，明显好于其它算法，且具有很好的可移植性。算法整体性能基本满足口语对话系统的需要。

3) 上下文相关的语音确认

大量文献指出在关键词检出后再加入语音确认步骤可以更进一步地降低误警率，并且语音确认步骤给出的置信度得分可以为后续的语音理解提供一定的信息，所以语音确认必不可少。为了进一步地提高语音确认的性能，需要找到更多具有区分度的知识源，本文即从这一出发点进行研究，提出了上下文相关的语音确认方法。将待确认词候选的前后序词候选的置信度特征也作为待确认词候选置信度特征的一部分，体现出识别结果的上下文对当前待确认词的影响。实验表明：该算法的拒识性能明显好于传统的语音确认算法。

1.6 论文的组织结构

论文的内容和结构是按照上述三个方面的主要工作来组织的。

在第二章中，介绍了本论文声学层面的研究背景，包括特征提取、去噪处理、声学基元选择、声学模型训练等一系列和本论文研究工作相关的背景知识；第三章从如何提高关键词检出算法本身的识别性能出发，提出了关键词动态确认技术，避免了误警候选对正确候选的干扰，有效地提高了关键词检出算法的整体识别性能；第四章介绍了如何引入语境知识来指导关键词识别，提高小领域中关键词检出算法的识别性能；在第五章中，将对上下文相关的语音确认方法进行详细的阐述；最后，在第六章中对论文的研究进行总结和展望。

在附录中，简单介绍了在博士就读期间完成的关于大词表连续语音听写系统的研究工作。

第二章　特征提取和声学模型建模

语音识别有三个基本问题：特征提取与选择、模式划分和时间对准。特征提取与选择既是语音识别的第一个基本问题也是一个重要环节。特征提取与选择的好坏直接影响到识别器的性能。同样，声学模型建模，即模式划分是语音识别的另一个重要环节，其任务是用一种比较好的方法把不同的识别基元在特征空间中划分开来。所谓比较好，指的是既简单又有效，是复杂度与性能的综合考虑。特征提取与声学模型建模方法的优劣将直接影响到口语对话系统中语音识别的整体性能，所以为了能够更准确的描述本文的研究内容，在详细介绍作者的研究工作之前，本章先对两个口语对话系统 *EasyFlight* 和 *dEar-Attendant* 中在特征提取和声学模型建模两个方面选用的方法作一个简要介绍，以作铺垫。

本章的内容安排如下：第 1 小节介绍特征参数的提取与选择，主要介绍 MFCC 和能量参数的提取；第 2 小节介绍特征参数的鲁棒性处理，重点介绍 CMN 算法；第 3 小节介绍声学识别基元的选择；最后第 4 小节描述声学模型的上下文相关建模算法。

2.1 特征参数

语音信号的特征主要有时域和频域两种。时域特征如短时平均能量、短时平均过零率、共振峰、基音周期等；频域特征有傅里叶频谱等。现在还有结合时间和频率的特征，即时频谱，充分利用了语音信号的时序信息。倒谱(Ceptrum)是语音信号的又一个特征[48]，有基于线性预测分析(LPC)的倒谱即 LPCC[38]，有基于 Mel 频率弯折的倒谱即 MFCC[38]。基于听觉模型的特征参数提取，如感知线性预测(PLP)分析[49]，试图从不同于声道模型的另一个方面进行研究。

所有这些特征都只包含了语音信号的部分信息。为了充分表征语音信号，人们尝试综合各种特征，并取得了一定的效果。但由于目前语音识别分类器的限制和数学模型描述的局限性，人们尚未充分利用已有的部分信息，于是特征的变换与取舍、特征时序信息的使用等成了重要的研究课题。

现有的语音识别系统中声学特征基本以 MFCC 为主，在 *EasyFlight* 和 *dEar-Attendant* 中也使用 MFCC，并配合上能量参数以及它们的一阶、二阶差分组成一个多维的特征向量来描述语音特征。下面介绍特征的具体内容和提取步骤。

MFCC

从人类听觉系统的研究成果来看，人耳分辨声音频率的过程犹如一种取对数的功能，相当于 Bark 因子的作用[50]，基于此出现了 MFCC，MFCC 比 LPCC 具有更优越的性能。MFCC 的计算流程[38]：1) 对原始语音分帧后进行预加重并乘以哈明窗；2) 离散傅立叶变换；3) 在 Bark 域上使用三角滤波器对频域能量谱进行滤波；4) 离散余弦变换得到 N 维 MFCC 特征。

能量参数

能量参数是一种时域参数。对于同一个人来说，在相同条件下发不同的音节时，所产生的语音流的短时平均能量是不同的，从这个角度来说能量参数是可以用来区分语音的特征的。但能量参数有一个致命的缺点是低鲁棒性，因为不同的时间地点，不同的说话人发音时，能量相差很大，所以简单的能量参数都不宜直接使用。

在本文中，每帧的对数能量被提取出来作为一维特征参数，对数能量的公式如下，其中 N 为分析窗的宽度，$S_t(n)$ 表示第 t 帧中第 n 个点的信号样值。

$$Eng(t) = \log\left(\frac{1}{N}\sqrt{\sum_{n=0}^{N-1} S_t^2(n)}\right) \tag{2-1}$$

特征参数差分

将已有特征参数按时间进行一阶甚至二阶分析得到动态特征，这些动态特征的使用可以提高分类器的性能。如 Wilpon 等人[51]利用了倒谱和对数能量参数的高阶时间派生的一阶和二阶的倒谱派生参数、对数能量派生参数以及第二阶对数能量派生参数，从而有了对时变语音信号的更完整的

两维(时间-频率)表示，实验结果表明错误率可以下降 30%。类似的实验还很多。下面要介绍的是倒谱参数的一阶分析，二阶分析不难从一阶分析得到。

线性回归分析(**LRA**，Linear Regression Analysis)或自回归分析(**ARA**，Auto-Regression Analysis)是一种常用的统计分析方法。其基本思想就是把当前帧左右若干帧的语音综合起来通过 LRA 得到新的参数，与原始参数一起组成组合特征参数。一般地，LRA 具有下面的形式[52]：

$$ARC_t(d) = G \cdot \sum_{n=-n_0}^{n_0} n \cdot CEP_{t+n}(d), \quad 1 \le d \le D \tag{2-2}$$

其中 n_0 为分析窗宽，G 为归一化因子。为保证回归倒谱各维的方差与原始倒谱各维的方差相等，通常取

$$G = 1 / \sqrt{\sum_{n=-n_0}^{n_0} n^2} \text{ 。} \tag{2-3}$$

自回归倒谱也称为 Delta Cepstrum [53][54]。

另外一个常用的动态特征是差分倒谱(differential cepstra)：

$$DC_t(d) = CEP_{t+n_0}(d) - CEP_{t-n_0}(d), \quad 1 \le d \le D \tag{2-4}$$

本论文涉及的两个口语对话系统 *EasyFlight* 和 *dEar-Attendant* 采用同样的特征参数，它们为：

1）12 维 MFCC 参数、对数短时能量和频段能量；

2）上述特征参数的自回归倒谱；

3）自回归倒谱的差分倒谱；

3 部分特征参数组成一个 42 维的特征向量，经过下节中所述的均值归一化过程后，生成归一化特征作为模型训练和识别的特征输入。

2.2 倒谱均值归一化

有关特征研究的另外一个课题是鲁棒性研究。由于语音识别的最终目标是在现实世界中使用，不同人的特点、地方口音的变化、背景噪音的干扰等成了不可忽视的因素，因此必须研究一种方法，使得特征的提取尽可能不受这些因素的影响。

使用倒谱均值归一化(CMN, Cepstral Mean Normalization)[39][40]算法来提高特征的鲁棒性，避免背景噪音的干扰。其中心思想在于通过对计算出的 MFCC 特征进行后处理，MFCC 特征向量的均值在一段时间内归一化为 0。算法认为虽然背景噪音的性质是无法预知的、随时间变化的，但在一段较短的时间内，背景噪音的类型和频谱能量分布都比较稳定、保持不变，这一段时间内频谱能量均值中必然会包含噪音谱，将这一段时间内倒谱均值进行归一化等于将背景噪音谱减掉，消除了背景噪音的影响。

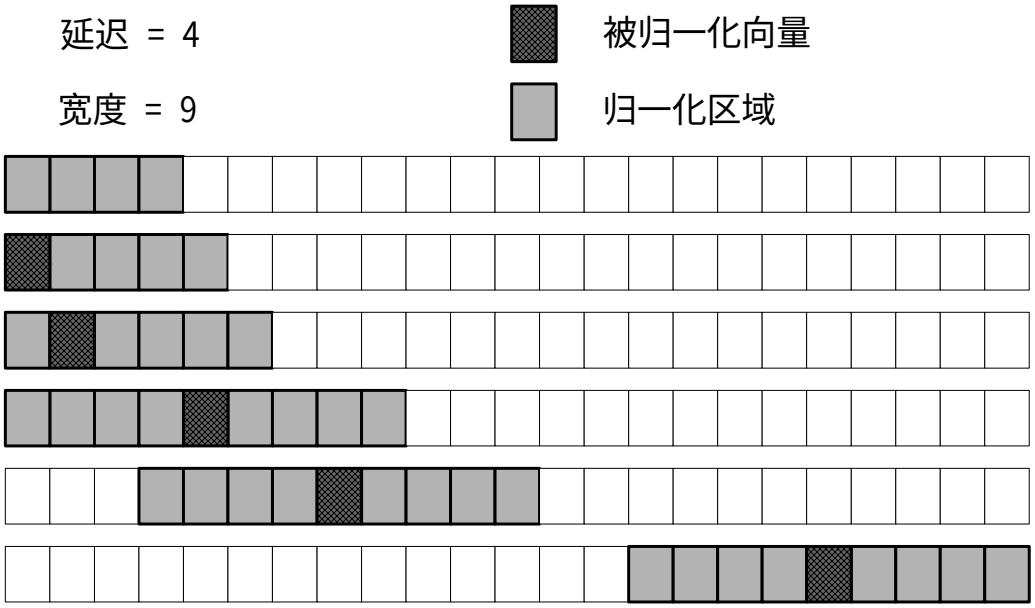


图 2-1 CMN 算法的实现

在具体实现上对某一帧特征向量进行处理时，如图 2-1 所示，先选定一个时间区域（通常包含此特征向量），如图 2-1 中阴影区域，求此时间

区域内的所有原始特征向量的均值向量，用待处理特征向量（阴影区域中的深色方格）减去均值向量即得归一化特征向量。时间区域通常选择 1 秒钟的时间宽度。

CMN 算法的物理意义也不是非常明确，但经过实验证明[39][40]确实可以提高在噪音环境下的语音识别性能。在本文涉及的两个口语对话系统中均使用了 CMN 技术以提高鲁棒性。

2.3 识别基元选择

语音识别基元的选择在语音识别尤其是连续语音识别中是重要的环节。识别基元的选择应该基于如下两个原则：具有灵活性，用它可以组成其他的语音或语法单位；具有稳定性，它应该在不同的语音环境和语言环境中相对稳定。灵活性希望基元尽可能地小，如音素；而稳定性则希望基元尽可能地大，如词甚至词组。

汉语的语音识别和关键词识别必须考虑汉语的特点，常用的基元包括：词(word)、音节(syllable)、声韵母(initial/final)和音素(phone)等。汉语是一个音节性强的语言，一个汉字对应一个音节，而且音节的结构是比较典型的声韵结构。声母和韵母又可以分为更细的音素。

汉语约有 400 个无调音节和 1300 多个有调音节[55]。在进行上下文无关的声学建模时，选用音节作为基元可以取得比较好的性能。但在连续语音识别中，音节间的协同发音现象比较严重，选用音节基元来描述这种现象是十分困难的。汉语有大约 35 个音素。音素基元在英语连续语音识别中得到了广泛的应用，并取得了很好的识别性能[56][57]。对于汉语，音素也是一个很好的选择。但音素并没有反映出汉语语音的特点，而且，相对于声韵母，音素显得更加不稳定，这就给标注带来了困难，进而影响声学建模。而声韵结构是汉语音节特有的结构，基元的大小比较合适，而且上下文关系比较确定，与音节和音素相比，更适合作汉语语音识别的模型基元。

为了使声韵结构更加结构化，提出扩展声韵母集(XIF, Extended Initial/Final)[58]，表 2-1 给出了扩展的声韵母基元定义，一共有 27 个声母

和 38 个韵母。与标准的声韵母定义相比，增加了 6 个零声母{_a, _o, _e, _I, _u, _v }，这样，每个音节都是由两部分组成的，分别对应其声母部分和韵母部分。

表 2-1 扩展声韵母基元列表[58]

声母基元 (27)	韵母基元 (38)
b, p, m, f, d, t, n, l, g, k, h, j, q, x, zh, ch, sh, z, c, s, r, _a, _o, _e, _I, _u, _v	a, ai, an, ang, ao, e, ei, en, eng, er, o, ong, ou, i, i1, i2, ia, ian, iang, iao, ie, in, ing, iong, iou, u, ua, uai, uan, uang, uei, uen, ueng, uo, v, van, ve, vn

当使用标准的声韵母基元集合时，有一些音节只有韵母部分，而没有声母部分。所以，考虑上下文相关信息时，这些韵母既可以搭配声母，又可以搭配韵母，因此，上下文相关声韵母基元数目会很大，超过 10 万个。而使用扩展的声韵母基元集合时，上下文关系比较确定，基元数目减少为约 3 万多个，有效地缓解了数据稀疏问题。同时，使用扩展的声韵母基元也有效地减少识别中的插入错误，其性能也优于标准声韵母基元。因此，在本文涉及的两个口语对话系统中均采用扩展声韵母作为识别基元，类似于英文中的 triphone，将这些基元形象地称为 tri-XIF。

2.4 上下文相关模型建模

在连续语音识别中，协同发音现象非常普遍，为此，当训练数据比较充分时，人们一般都会采用上下文相关（CD，Context-Dependent）的方法来为协同发音建模，并取得了良好的识别效果。对采用 XIF 作为基元训练出的上下文相关声学模型称为 CD-XIF。

在本文中识别基元全部采用三个状态、自左向右无跳跃的 HMM 模型[59]拓扑结构 ;各个基元状态的声学输出概率采用 4 个混合的高斯分布描述，见图 2-2 的表示。其中，最前和最后的两个是虚状态，不产生输出，仅仅用来在搜索解码时连接相邻的 HMM。

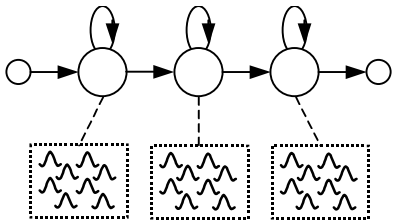


图 2-2 HMM 基元模型拓扑结构

对 CD-XIF 的训练方法如下[60]：

(1) 训练上下文无关的单混合的扩展声韵模型；

(2) 对原始数据中标注的上下文无关的扩展声韵母进行上下文扩展，从而得到上下文相关的声韵标注信息。通过对上下文无关模型的简单复制得到上下文相关的扩展声韵模型。然后通过 Baum-Welch（也称为 Forward-Backward）算法[59][61]的迭代训练来重估参数。

(3) 使用基于决策树的状态共享方法进行状态的合并，生成一个新的基于状态共享的上下文相关的声韵模型，然后通过迭代训练来重估参数。

(4) 使用分裂的方法来增加状态中的混合数目，从而细化模型。

其中，基于决策树（Decision Tree）[62]的状态共享（State Tying）[57]方法是建立上下文相关模型的关键。这个方法已被广泛地应用于大词表连续语音识别系统中[63][64][65][66]，其目的是通过状态共享解决训练数据相对稀疏的问题，并在模型中合成训练集中从未出现过的上下文相关基元。

决策树的基础是根据先验知识定义的一个问题集合。问题集的选取对于基于决策树的声学建模是很重要的。一些研究者的工作就集中在怎样选取更好的问题集[67][68]。在本文的实验中，问题集是从语音学角度选取的[69]。一般根据需要，为每个基本的识别基元（例如所有可能的 tri-XIF 中心基元）的每个状态构造一棵决策树。决策树从结构上讲是完全的二分搜索树，每个结点上携带一个合适的问题，根据数据驱动的原则对它们进行确定，即每个结点所选择的问题应从整体上使训练集数据划分出的类别间

具有尽可能大的声学差异。

决策树方法对模型间的状态共享一般是通过离线的方式对模型进行修正得到的。对每个训练集中出现和未出现（但可能在将来的测试集中出现）的 tri-XIF 中心基元的各个状态，从它所对应的决策树根结点出发，将这个 tri-XIF 的上下文相关部分的基元与当前结点携带的问题进行匹配，根据匹配与否，分别沿着结点的左分枝或右分枝向下推进，直至到达决策树的叶结点。最后，对所有到达相同叶结点的 tri-XIF 状态内部的参数（例如观测概率分布和状态转移概率等）进行完全的共享。决策树方法是一种结合了基于数据驱动方法和基于知识方法的方法。与基于数据驱动方法相比，它能够对训练数据稀少的基元和没有训练样本的基元给出适当的参数估计。与基于知识的方法相比，它能够弥补专家知识不足等缺陷。

本文中所使用的决策树分裂准则是最大似然准则，即选择结点分裂后似然分增加最大的问题作为本结点绑定的问题。决策树的停止分裂采用阈值进行控制。当分裂后的结点中训练样本数目少于一定数量时，或者，当本结点分裂后对数似然分数的增加小于一定的阈值时，停止分裂。

基于决策树的状态参数共享方法具有如下两个突出的优点：首先，声学上相近状态的参数共享起来，可以起到一种平滑作用，使得每个模型都可得到比较充分的训练，从而减小数据稀疏的影响；其次，对于训练集中没有出现、但将来可能在测试集或其它场合出现的 tri-XIF，也可以通过决策树，让它们的每个状态与某个声学特性相近的、已经充分训练好的状态进行参数共享，从而保证识别性能。

通过决策树方法完成模型状态参数共享后，与上下文无关模型的训练方法相似，利用 Baum-Welch 算法，对模型进行若干次迭代，直到模型集的参数收敛，然后逐步增加状态内的高斯混合数（提高模型的复杂度），并重复利用 Baum-Welch 算法进行迭代训练。如此反复，直到达到期望的高斯混合数且模型参数收敛。图 2-3 给出了进行状态共享后 tri-XIF 的 HMM 模型结构示意图。

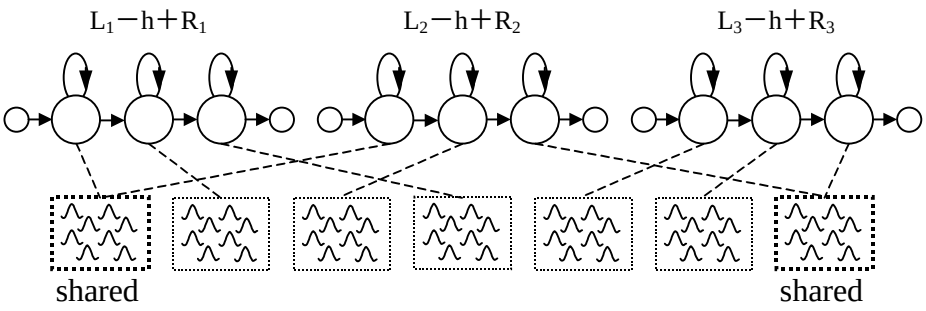


图 2-3 上下文相关且状态共享的 HMM 模型结构

第三章　关键词动态确认

关键词识别（KWR，Keyword Recognition），又叫关键词检出（KWS, Keyword Spotting），就是在连续的、无限制的自然语音流中识别出一组给定的词——关键词，它是语音识别的一个研究方向，与连续语音识别之间存在着密切的关系，也可以说关键词识别是连续语音识别的一个分支[70]。

与连续语音识别相比，关键词识别的特点在于可以忽略关键词以外的语音部分，很多口语现象可以被忽略掉而不影响关键词的识别性能，因此，现有的口语对话系统大多采用关键词识别策略。当关键词表较小时，容易取得令人满意的识别性能；但随着关键词表的增大，词之间混淆度的上升，关键词识别的识别性能将会逐渐下降，最终不能满足产品化、实用化的要求，所以关键词识别的识别性能仍亟需提高。本章的主要内容是介绍作者在关键词识别方面的研究工作，包括分析基准关键词识别算法的缺点，提出关键词动态确认概念，在降低误警率的同时有效地提高了识别正确率。

本章的内容安排如下：第一小节简要介绍关键词识别的背景知识，包括现有的几种关键词识别策略，以及关键词识别算法的评测指标；第二小节介绍基准关键词识别算法：基于音节补白的关键词识别，并详细给出其性能指标，然后通过对实验数据的分析提出基准算法存在的问题；第三小节提出关键词动态确认的概念并给出实现方法；第四小节给出实验结果；最后第五小节进行总结。

3.1 关键词识别算法综述

关键词识别算法通常采用补白模型来实现对非关键词语音段的识别，这类方案的主要特点是：词表内的关键词和补白处于平行关系，为了区分，对关键词和补白加以不同的权值。只有当关键词的权重高于补白的权重时，关键词才能被检出。从直观上可以看出：关键词和补白模型的权重相差越大，关键词越容易被检出，识别正确率越高，同时误警率越高。图 3-1

是一种典型的关键词加权的检出框架。这种方法，往往适用于句子中出现关键词比较少的场合。

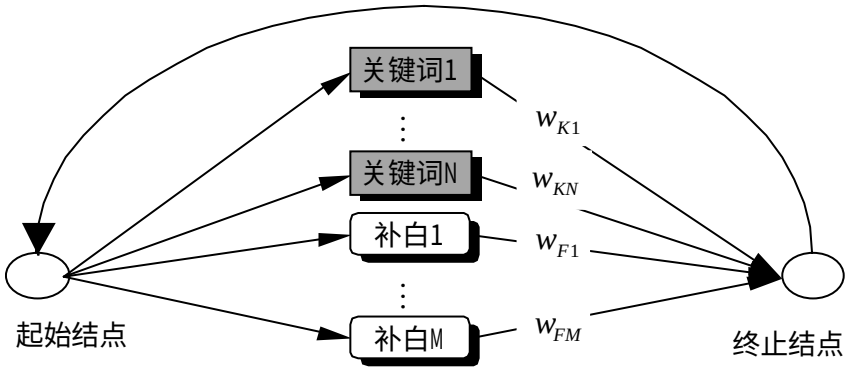


图 3-1 关键词补白加权识别框架[70]

3.1.1 关键词识别算法分类

理想的补白模型应有足够的能力吸收除关键词之外所有剩余的语音信号，同时与关键词模型之间又有足够远的距离，使其不具备竞争关键词所对应的语音段的能力。根据构成补白的声学模型的来源不同，关键词识别算法又分为三类：集外补白模型法[71][72]、子词补白模型法[73]和在线补白模型法[74]。

1）集外补白模型法。集外补白是指专门为补白建立声学模型，完全独立于关键词所对应的声学模型集。在处理口语语音中，集外补白可以只包含一个通用的补白模型，也可以是若干个补白模型形成的集合。在训练补白模型集时，将训练数据中除关键词之外的额外输入分为若干类，其中的每一类对应一个模型，采用 HMM 拓扑结构。集外补白模型法的优点是结构简单、算法复杂度低，尤其在训练数据较少的应用中用途较广，但当训练数据比较充分时，因为其补白模型相对比较简单、数目较少，识别性能不如其它两种方法。

2）子词补白模型法。补白与关键词共享同一套子词单元模型集，补白模型由子词模型直接构成或拼接组合而成。子词一般为比关键词更小的发音单元，如音素、声韵，对汉语来说，还有音节，关键词则是这些小的

发音单元的组合，一个关键词通常是多个子词的串接。可设置一个词典和搜索网络表，在其中定义关键词和补白模型的串接及识别所用的搜索网络，在 Viterbi 搜索框架中可以通过调整关键词和补白模型的惩罚分大小来区别关键词和补白模型。当训练数据充分时，采用上下文相关建模方法的子词模型可以取得最优的识别性能，而且由于子词模型的训练基本和关键词是独立的，所以在实际使用中修改关键词集的定义时无须重新训练子词模型，具有较好的灵活性。缺点则是模型比较复杂、算法复杂度高且需要的训练数据巨大。

3）在线补白模型法。在线补白，最初是 Bourlard 等人在 1994 年提出的一种思想：并不试图从数学上描述补白模型，不专门为补白建立特定的模型，而是在搜索过程中动态地形成一个补白，同关键词进行竞争，对每一帧语音，补白模型的似然度是该帧信号对应的 N 个最优匹配成绩的平均。在这种方式下，补白模型永远不是最佳的匹配者，但也总是排在局部最优匹配的前几位，只有当一段语音同关键词最佳匹配时，关键词的整体成绩才可望在补白的竞争下胜出。这种方式不需要特殊的补白模型构造过程，而且鲁棒性要好于显式的补白模型。因为当在噪音环境下使用时，对于一段语音，所有的模型都匹配不好，打分都低了，最终导致补白模型的得分也相应降低。而在显式补白模型的情况下，所有的模型打分都不准确，最终会造成识别率下降，从而降低系统的鲁棒性。由于在线补白需要计算前 N 个最优匹配成绩的平均，但当关键词个数很少甚至少于 N 时，前 N 个最优匹配成绩的平均就不再是总体较优的路径得分，而可能是总体较差路径得分，这就违背了在线补白模型的初衷。实验证明：在关键词个数很少时，在线补白模型的性能会变得很差。

除了以上三种算法之外，还有一种关键词识别算法：滑动窗算法[75]。这种算法不使用补白模型，而是基于这样的思想：即一开始从语音数据的第一个点开始进行搜索，在得到结果后再选择下一个搜索起点进行下一轮搜索，直至认为没有可能再出现关键词为止。使用滑动窗方法不排除在一遍搜索时，采取关键词加权等策略。由于可以从语音的任意起点开始搜索，这种方法的好处在于最大限度地提高了关键词的检出率，但同时也带来误

警率较高的问题。

3.1.2 评测指标

下面介绍几个评价关键词识别结果的标准。

可接受：若 X 是关键词候选，而 Y 是正确的关键词，若 X 的中间帧落在 Y 的边界之内，则称 X 的时序是可接受的[51]。

识别正确：若 X 的时序是可接受的并且 X 等于 Y，则称 X 是识别正确的。

识别错误：若 X 的时序是可接受的但 X 不等于 Y，则称 X 是识别错误的。

误警：当 X 的时序不可接受时，则称 X 是一个误警(False Alarm)。

识别正确率(Accuracy Rate)：正确识别的关键词占关键词出现总数的百分比，也称检出率。

漏识率：没有正确识别出的关键词占关键词出现总数的百分比，也称漏检率。

误警率：每小时每个关键词的误警数，简记为 fa/kw/hr 或 fa/kw-hr。

然而系统的性能并不是通过上面的参数来孤立地进行评价，在关键词识别时，有时候是允许误警的，或者说，是在某一个误警范围内判断关键词的检出正确率。于是就有了一个评价标准，这就是质量因数(FOM，Figure of Merit)，它的值越高，表示系统的识别性能越好。

质量因数：NIST（国家标准化技术机构）定义了质量因数 FOM，它是每小时每个关键词的十个错误的平均得分，它可用下式来表示。

$$FOM = \frac{(p_1 + p_2 + ... + p_N + ap_{N+1})}{10T} \tag{3-1}$$

这里 p_i 是第 i 个错误发生前的命中百分比，T 是测试的时间(以小时计)，N 是大于 10T-1/2 的最小整数，a=10T-N 是插值因子。这个 FOM 使得没必要

去比较关键词的得分，并产生了每小时每个关键词的从 0 到 10 个错误的平均性能。它同时提供了一个比 0 或 1 个错误的检测率(当这种错误有较大的方差时)更稳定的性能。人们在实际计算的时候，有时把它简化成下面的计算式[76]：

$$FOM = \frac{(p_1 + p_2 + ... + p_{10})}{10} \tag{3-2}$$

ROC(Receiver Operating Characteristic)曲线[76]：这个曲线是关键词的正确检出率与误警率之间的关系曲线，显然它是一条单调非减曲线。曲线下面的面积越大，则说明关键词识别的性能越好。

在关键词识别中，调整关键词和补白模型权重的大小可以改变算法的检出率和误警率，权重相差越大，检出率越高同时误警率越大。两个关键词识别算法进行性能比较时，仅仅比较识别率的大小是说明不了问题的，同时也比较误警率的大小。为了避免关键词和补白权重的干扰，一般采用相同误警率下的检出率多少作为比较标准，或通过设置不同的权重，分别计算出两者的 ROC 曲线来比较两个算法的整体识别性能。

3.2 基准关键词识别算法

本文中采用音节补白关键词识别算法[70]作为基准的关键词识别算法，该算法采用上下文相关的子词模型拼接出汉语中 418 个无调音节作为补白模型。由于补白模型集中包括了汉语中可能出现的所有无调音节，所以补白模型可以描述正常发音的任何语句。关键词也和声学模型无关，同样由子词模型拼接而成，可以修改关键词集而不用重新训练模型。算法中的子词模型为通过大批量数据训练得到的上下文相关的扩展声韵母模型 tri-XIF，搜索过程中无论关键词还是补白均采用跨词(Cross-Word)搜索[77]的 Viterbi 算法，充分发挥了上下文相关模型的性能优势。算法整体具有相对较好的识别性能和良好的灵活性，是现有的几种关键词识别算法中整体性能最佳的算法之一。

3.2.1 拓扑结构

搜索网络的拓扑结构如图 3-2 所示，识别框架和图 3-1 所示的关键词补白加权识别框架几乎完全一样，只是补白模型具体定义为汉语中出现的 418 个音节。在实际的搜索过程中，搜索网络被组织成树状结构。

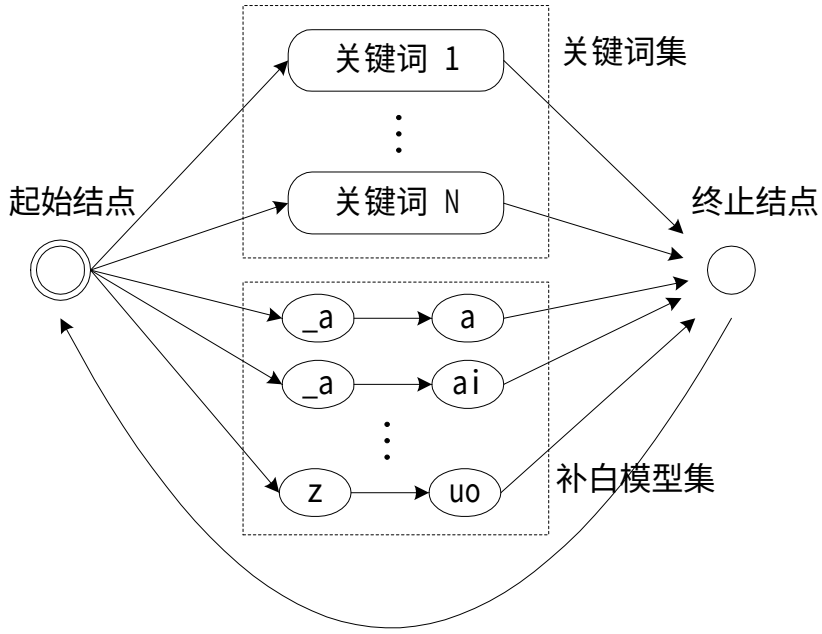


图 3-2 音节补白关键词识别算法搜索空间的拓扑结构

3.2.2 搜索网络结构

基准算法的搜索网络采用词法树(Lexical Tree)[77]的形式组织，它将词表内的词从前至后地尽可能共享在一起的，形成一种前端结点小，后端结点大的树状网络，树状网络中每个节点代表一个上下文相关子词模型，这样可以在搜索过程中极大地节省搜索空间。在基准关键词识别算法中，将关键词和补白模型都用同一个词法树组织起来后会出现一个问题，就是只有在到达词法树叶子节点时才知道刚才识别的内容是关键词还是补白模型，也就是说只有等搜索到达词法树叶子节点时才能加上关键词权重或补白模型权重。为了弥补这一缺点，我们使用了一种折扣技术。首先将关键

词或补白模型都认为是同等地位的词，区别只在于权重不同，通常情况下关键词的权重应大于补白模型权重。然后将权重平均分配到每个父节点上，因为词法树中肯定具有不同权重的词有共同的祖先，这样分配到这个祖先节点的权重就不止一个，这时对折扣分进行比较，取最大值作为这个祖先节点的权重。按照此方法生成的搜索网络如图 3-3 所示

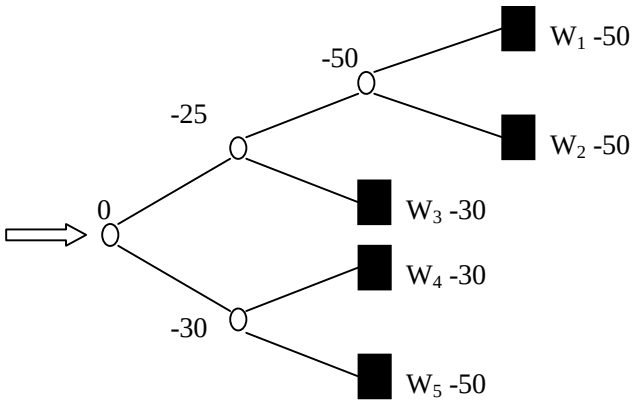


图 3-3 权重的折扣

在搜索过程中，t 时刻搜索到节点 s (父节点为 r)的路径总得分 $Q(t,s)$ 等于声学似然得分 $P(t,s)$ 加上当前节点的权重值 $W(s)$，整个搜索过程可以用如下公式表示，其中 $q(x_t \mid s)$ 表示 s 节点对时刻 t 的特征向量的似然得分：

$$P(t,s) = P(t-1,r)\cdot q(x_t \mid s) \tag{3-3}$$

$$Q(t,s) = P(t,s)\cdot W(s) \tag{3-4}$$

3.2.3 识别性能

在和口语对话系统 *EasyFlight* 和 *dEar-Attendant* 类似的电话信道上测试音节补白关键词检出算法的识别性能，实验环境如下：

(1) 关键词表：采用国内 A 股和 B 股中所有上市股票的名称，所有多音字都作为不同的关键词，共 1389 个关键词；

(2) 测试数据库：包含 10 男声 10 女声数据，每人 100 句共 2000 句。

采用 8k 采样率，通过电话采集。

描述音节补白关键词识别算法识别性能的 ROC 曲线如图 3-4 所示，图中曲线的不同点代表在不同的关键词和补白模型权重的条件下，算法的检出率和误警率。图中，误警率越高的点代表关键词和补白模型权重相差的越大，误警率越小的点代表关键词和补白模型的权重相差越小。

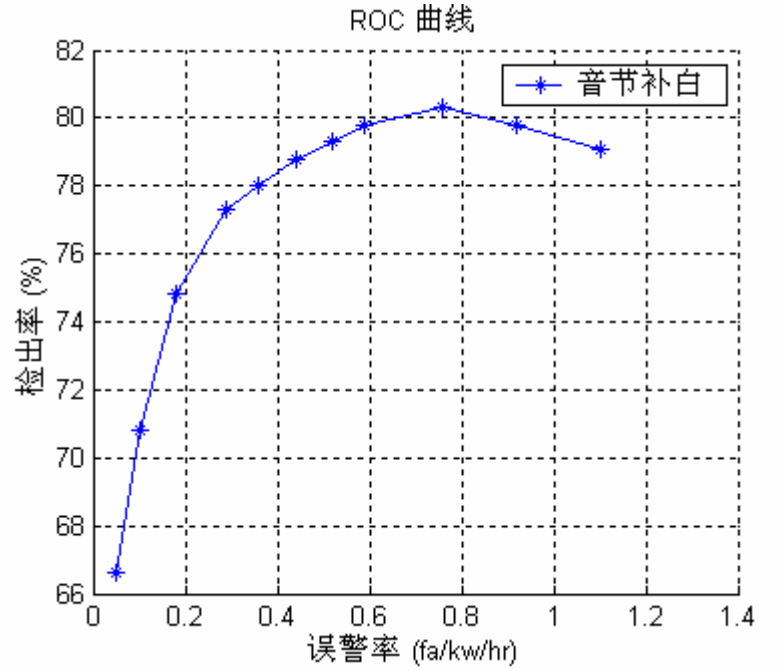


图 3-4 音节补白关键词检出算法的识别性能

3.2.4 实验结果分析

根据对测试语句的手工标注信息，对音节补白关键词检出算法的识别结果进行分析，试图找到关键词无法被正确检出的原因，以期今后可以对症下药有针对性地进行算法改进。在不影响可推广性的前提下，我们选取 ROC 曲线上检出率为 80.3%，误警率为 0.76 这一点，对其识别结果中没有被正确检出的关键词分析原因。

首先定义了四种错误类型：

(1) I 类错误：标注关键词区域内没有任何关键词候选被识别出来，识

别结果中只有补白模型和静音；

(2) II 类错误：标注关键词被识别为另一个关键词候选；

(3) III 类错误：标注关键词区域前面部分被另一个识别出的关键词候选所覆盖，后面部分的识别结果仅有补白模型和静音；

(4) IV 类错误：标注关键词区域前面部分被识别为补白模型，中间或后面被另一个识别出的关键词候选所覆盖；

图 3-5 中举例描述了这四类错误的定义并给出了统计结果。

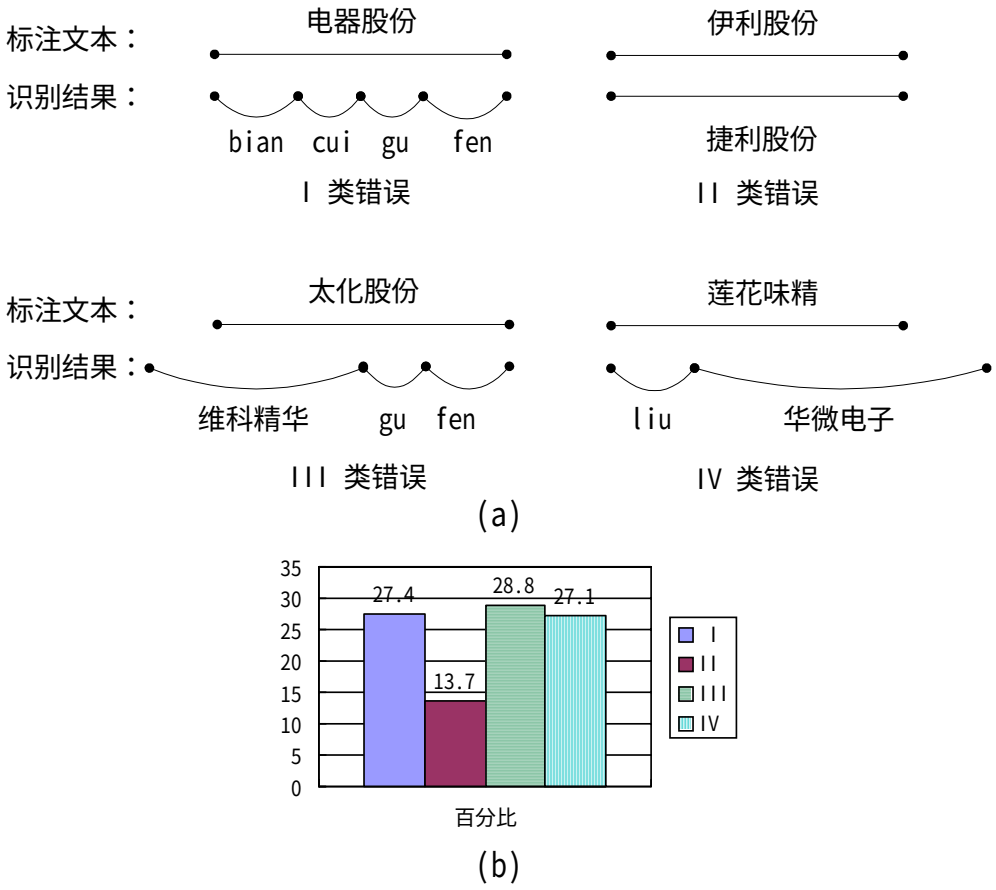


图 3-5 音节补白关键词识别算法错误分析

(a 是 4 类错误举例，b 是 4 类错误在识别结果错误样本中出现的百分比)

从图 3-5 中可以看出，II 类、III 类和 IV 类错误都与另一个被错误识

别出的关键词候选有关。虽然这个关键词候选并不正确，但是经过它的路径的总体似然得分却比经过正确关键词的路径要高，这主要是由于声学模型的不正确匹配所导致。最终，根据统计结果 II 类、III 类、IV 类错误出现的百分比总和为 72.6%，换句话说，72.6%的识别错误都和另一个关键词候选的误警有关。

为了进一步说明错误关键词候选对正确关键词候选的干扰影响，根据手工标注信息，将所有没被正确识别的关键词所对应的语音部分，分别手工从语音文件中取出，组成一组新的测试样本，每个测试样本中只包括关键词的语音没有任何多余语音，共有 394 句。新测试集的识别率为 51.0%，即排除了前后语音的干扰后，有一半左右的关键词可以被正确识别出来。

3.2.5 问题提出

从上面的两个实验数据可以看出关键词错误识别的发生和其它关键词的误警有很大关系。在图 3-4 的音节补白模型识别性能 ROC 曲线中也可以看到：当误警率由小变大时，检出率先是同步增长，到一定程度后趋于稳定不发生明显波动，当误警率继续上升时，检出率开始出现下降趋势。这也说明误警关键词候选的增多会干扰正确关键词候选的识别。

这给我们提出一个问题：如何才能避免关键词误警的发生，降低其干扰作用？毫无疑问，提升声学模型的识别性能可以降低误警关键词候选对正确关键词候选的干扰影响，并改善关键词识别效果。但在现有声学模型等其它条件均不变化的情况下，在搜索过程中准确地避免关键词误警同样可以降低干扰作用，最终提高识别率。这就需要在搜索过程中引入新的方法、新的知识源来达到此目的。

下面将从搜索过程如何避免关键词误警的角度出发，提出关键词动态确认的概念。

3.3 关键词动态确认

误警的关键词候选会干扰正确关键词候选的检出，为了尽可能地排除

误解关键词候选的影响，提升识别性能，本文提出关键词动态确认技术来达到这一目标。

3.3.1 定义

关键词确认技术是语音确认技术的一个方面，能够有效地分辨出关键词识别算法得到的关键词候选正确与否，是正确检出还是误警，目标是在几乎不影响正确关键词的检出率的前提下，尽可能地降低误警率。虽然关键词识别算法中也可以通过调整关键词和补白模型的权重来实现误警率的下降，但相应的要以牺牲检出率为代价。[73][78][79]指出：关键词识别结合关键词确认可以取得更优的识别性能。从降低误警率这一目标出发，关键词确认和我们寻找的方法相一致，值得关注。

在实现方法上，通常的关键词确认和关键词识别分为两阶段完成，为了表述方式，在后文中将这种两阶段的关键词确认称为后处理关键词确认。在关键词识别阶段，识别器给出概率似然值得分最高的待确认关键词候选；关键词确认阶段则根据从关键词候选中提取的置信度特征，与事先训练好的置信度模型，最终给每个待确认关键词候选一个置信度得分，得分高于事先统计出的阈值则保留，否则删除。两阶段关键词确认方法的优点在于在关键词识别结束后可以使用更多更复杂的置信度特征和置信度模型，关键词确认的性能比较好；但是两阶段关键词确认方法有一个无法避免的缺点即如果在关键词识别阶段，输入样本中的关键词部分被识别成为另一个错误的关键词候选，即使在确认阶段能够将识别错误的关键词排除，但是正确的关键词再也不可能被识别出来。

为了达到提升关键词检出的识别性能，本文提出关键词动态确认技术。与传统后处理关键词确认的思路不同，在关键词识别过程中即加入关键词确认步骤，在搜索过程中就可以分辨出一部分误警关键词候选，可以及早地避免了这些误警关键词候选对原本应正确识别出的关键词所带来的负面影响，保障这些原本受到干扰的正确候选最终可以被识别出来，从而有效地提高关键词识别的识别率。

3.3.2 后验概率计算

虽然关键词动态确认和后处理关键词确认不同，但仍然能借鉴后处理关键词确认算法的思路。在后处理关键词确认中，后验概率非常重要，得到广泛使用。根据贝叶斯公式，一段语音 X 是词 W 的后验概率 $P(W|X)$的定义如下：

$$P(W \mid X) = \frac{P(W)P(X \mid W)}{P(X)} = \frac{P(W)P(X \mid W)}{\sum_{W} P(W)P(X \mid W)} \tag{3-5}$$

但是在实际的语音识别系统中，一般都忽略 $P(X)$，因为对给定的 X 来说，$P(X)$是一个常量。因此必须采用一些正规化的方法来估计 $P(X)$，通常可以通过音素层和状态层的解码计算 $P(X)$。在状态层的解码方法中，使用一个由所有状态组成的网络，由 X 对所有状态的先验概率和构成 $P(X)$，在不失代表性的前提下，将语音段 X 表示为多个特征向量的集合：$X = \{x_1, x_2, \cdots, x_T\}$，$P(X)$的计算公式如下：

$$P(X) = \sum_{W} P(W)P(X \mid W) \approx \prod_{t \in T} \sum_{S} P(S)P(x_t \mid S) \approx \prod_{t \in T} \sum_{S} P(x_t \mid S) \tag{3-6}$$

其中，S 代表状态。所由公式(3-5)计算得出后验概率 $P(W|X)$本身就是词 W 的确认分数。在语音确认中，计算出 $P(W|X)$后，将其按时间 T 进行归一化后再与一个固定阈值相比，大于阈值则认为通过确认，反之则认为没有通过确认。阈值通过对数据库的统计得到，很显然阈值的取值范围为(0,1)。根据以上分析，通过确认的关键词候选其似然得分必然如下性质：

$$\frac{P(X \mid W)}{\prod_{t \in T} \sum_{S} P(x_t \mid S)} > (C^V)^T$$

即：$$\log P(X \mid W) > T \log C^V + \sum_{t \in T} \log \left[N_S \cdot \frac{1}{N_S} \cdot \sum_{S} P(x_t \mid S) \right] \tag{3-7}$$

$$\log P(X \mid W) > T\left(\log\left(C^V \cdot N_S\right)\right) + \sum_{t \in T} \log\left(\frac{1}{N_S} \cdot \sum_{S} P(x_t \mid S) \right)$$

在通常语音确认的具体实现上，$\sum_S P(x_t \mid S)$的计算非常耗时间，尤其当采用上下文相关子词模型构成的声学模型时，状态数可能成千上万，为此必须进行简化。在保证不损失确认性能的前提下，可以进行如下简化：

$$\log\left(\sum_S P(x_t \mid S)\right) \approx \log\left(\max_S P(x_t \mid S)\right) \approx \log\left(\max_{S^{active}} P(x_t \mid S)\right) \tag{3-8}$$

在(3-8)中，通过实验可以验证简单用最大值代替和不明显影响确认性能，而且取最大值的范围也可以限制在 t 时刻中所有活动的状态，这样就不需要每一时刻都对所有状态进行打分。这种大幅度的简化可以在不明显影响性能的前提下最大可能地节省了计算量。

3.3.3 后验概率计算公式的简化

后验概率的计算方法，同样可以应用在关键词动态确认中。不过我们希望能够采取和(3-8)简化公式不同的简化方式，这样可以保证关键词动态确认采用的知识源和后处理方式的语音确认不同，否则两者在功能上就会发生重叠。为此，在关键词动态确认中提出另一种简化方式如下：

$$\begin{aligned}\log\left(N_S \cdot \frac{1}{N_S} \cdot \sum_S P(x_t \mid S)\right) &\approx \log\left(N_S \cdot \sqrt[N_S]{\prod_S P(x_t \mid S)}\right) \\ &\approx \log N_S + \frac{1}{N_S}\sum_S \log P(x_t \mid S)\end{aligned} \tag{3-9}$$

则(3-7)式将变为：

$$\log P(X \mid W) > T\left(\log\left(C^V \cdot N_S\right)\right) + \sum_{t \in T}\left[\frac{1}{N_S}\sum_S \log P(x_t \mid S)\right] \tag{3-10}$$

那么式(3-9)的简化是否合理呢，可以用实验来进行验证。定义 100 帧的数据，每帧随机产生 10000 个状态的概率得分，这些概率得分都在[0,1]区间内，观察以下两个函数的取值走向：

$$f_1(t)=\log\left(\frac{1}{N}\sum_{n\in N}x_t^n\right)\qquad f_2(t)=\log\left(\sqrt[N]{\prod_{n\in N}x_t^n}\right)\tag{3-11}$$

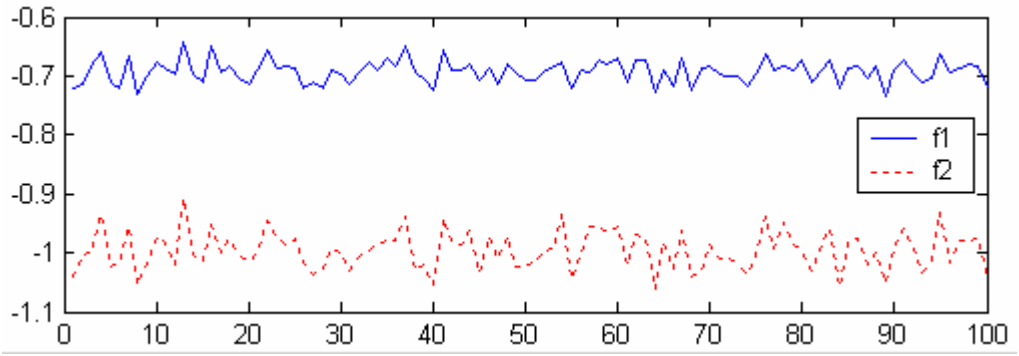


图 3-6 近似函数的比较

从图 3-8 中的比较结果可以看出，两个函数的走向和变化几乎完全相同，两者之间只有一个很小的偏差。因为这个结果是通过对随机数据的分析得到，应该说得到的结论有普遍的适用性，式(3-9)的简化过程是完全合理的。由于在实际的搜索过程中，专门对计算各状态的似然得分进行了处理，可以直接计算出 $\log P(x|S)$，所以式(3-9)的简化可以大量的节省计算量。但由于仍然需要对所有状态计算似然得分，计算量还是大的无法接受，所以还需要进一步的简化。式(3-10)可以进一步简化为：

$$\begin{aligned}\log P(X|W)&>T\left(\log\left(C^V\cdot N_S\right)\right)+\sum_{t\in T}\left[\frac{1}{N_S}\sum_{S}\log P(x_t|S)\right]\\&>T\left(\log\left(C^V\cdot N_S\right)\right)+\sum_{t\in T}\left[\frac{1}{N}\sum_{n=1}^{N}\mathrm{Top}\left(\{\log P(x_t|S)\},n\right)-C^N\right]\\&>T\left(\log\left(C^V\cdot N_S\right)-C^N\right)+\sum_{t\in T}\left[\frac{1}{N}\sum_{n=1}^{N}\mathrm{Top}\left(\{\log P(x_t|S)\},n\right)\right]\end{aligned}\tag{3-12}$$

在式(3-12)中，函数 $\mathrm{Top}(\{\cdots\},n)$ 指取出集合 $\{\cdots\}$ 内第 n 个最优值，则函式(3-12)的简化过程实质上就是并不对全部的状态计算似然得分，而只是计算出前 N 名最优的状态似然得分就可以了。现在来分析一下式(3-12)中不等式右边的第一项，C^N 为正数，根据数据的分布和 N 的取值进行变化，

N 的取值将通过实验来决定，其取值范围可以变化很广。而 C^V 通常也由实验得到，在这里不妨假设可以通过实验选取一个合适的 N 值，保证式(3-12)中不等式右边的第一项等于零，式(3-12)又可以进一步的简化为：

$$\log P(X|W) > \sum_{t\in T}\left[\frac{1}{N}\sum_{n=1}^{N}\mathrm{Top}\left(\{\log P(x_t|S)\},n\right)\right] \tag{3-13}$$

通过这次简化，虽然只需要计算 N 个状态似然得分就可以，但为了保证计算出的 N 个似然得分是最优的前 N 名，仍然需要计算出所有状态的似然得分再进行排序，并没有节省计算量。为了解决这个问题，我们引入一个假设：在 t 时刻时，通过正常的搜索过程所扩展出的状态似然得分中，必然包括最优的 N 个状态似然得分。这个假设的依据就在于，搜索过程本身就是寻找最优解的过程，在任意时刻能够被扩展出到状态应该就是似然得分比较高的状态，似然得分比较低的状态会在剪枝过程中被剪掉，从这个角度来分析，假设具有一定的合理性，由于在整个的过程中已经运用了多处简化工作，可以认为假设所带来的误差可以忽略不计，我们认为这个假设成立。根据这个假设又对式(3-13)进行简化：

$$\log P(X|W) > \sum_{t\in T}\left[\frac{1}{N}\sum_{n=1}^{N}\mathrm{Top}\left(\{\log P(x_t|S^{active})\},n\right)\right] \tag{3-14}$$

根据多步简化而最终得到的式(3-14)，其计算量已经非常小。从中可以看出，只需要在 t 时刻所有活动的状态中，选取前 N 名最优的对数似然得分，将其平均值作为这一帧的 $P(X)$的得分，用这个得分来计算最终的后验概率得分。

3.3.4 虚拟 OOV 模型

根据以上思想来实现关键词动态确认工作。定义一个虚拟 OOV(Out of Vocabulary)模型来完成关键词动态确认工作。在关键词识别过程中，进行时间同步的搜索时，与关键词模型和补白模型一样同步地对这个虚拟 OOV 模型进行打分，当待确认关键词候选的声学得分高于虚拟 OOV 模型在对应语音段的得分时，则认为该关键词候选通过确认并予以保留候选，反之

则认为没有通过确认并将其删除。

虚拟的 OOV 模型并不试图用复杂的数学公式来定义 OOV 模型，而是在任何时刻，将所有活动的声学模型状态对当前特征向量的似然得分的前 N 名进行平均，将平均值作为 OOV 模型对当前特征向量的似然得分。将接连几帧 OOV 模型的似然得分累加起来即是 OOV 模型在一段语音的似然得分和。如下所示，公式(3-15)定义了虚拟 OOV 模型在某一时刻 t 时的似然得分 S_t^{oov} 的计算方式。公式(3-16)定义了虚拟 OOV 模型在一段时间 $t_i \sim t_j$ 内的似然得分 $Q^{oov}(t_i, t_j)$ 的计算方式。

$$S_t^{oov} = \frac{1}{N}\sum_{n=1}^{N} \mathrm{Top}\left(\left\{S_t^{all}\right\}, n\right) \tag{3-15}$$

$$Q^{oov}(t_i, t_j) = \sum_{t=t_i}^{t_j} S_t^{oov} \tag{3-16}$$

其中，$\left\{S_t^{all}\right\}$ 是 t 时刻所有活动状态的似然得分的集合，$\mathrm{Top}\left(\left\{S_t^{all}\right\}, n\right)$ 是取似然得分集合 $\left\{S_t^{all}\right\}$ 中第 n 个最优值的函数。

在进行关键词动态确认时，并不考虑关键词候选路径总体的路径似然得分，仅仅关注关键词候选的似然得分，即搜索路径在关键词候选对应的语音段内累计的似然得分和，关键词候选的历史信息和权重并不考虑。用关键词候选的似然得分与虚拟 OOV 模型在相应语音段内似然得分相比较，关键词候选似然得分较大则认为通过确认保留之，否则认为没有通过确认将其删除。

公式(3-15)中的 N 代表取前多少名似然得分来求平均值。显而易见，当 N 变大 S_t^{oov} 则变小；N 变小 S_t^{oov} 则变大，N 与 S_t^{oov} 为反比例关系。对于特殊情况 $N=1$ 时，$S_t^{oov} = \max\left(\left\{S_t^{all}\right\}\right)$，即虚拟 OOV 模型的似然得分等于每帧所有状态似然得分中的最大值，此时的关键词动态确认应该为最严，不可能有一个关键词候选通过确认。在实际的应用中，调整 N 的值就可以调整关键词动态确认的宽严程度，N 越大越宽。

3.3.5 物理意义

根据上节的分析，可以看出，虚拟 OOV 模型实际上就是采用似然比的方法来对模型的后验概率估值进行近似。我们还可以从另一个角度来理解虚拟 OOV 模型的物理意义。因为 OOV 模型针对每一帧特征向量的似然得分都是其它各状态似然得分中最优 N 个的平均值，所以每一帧的 OOV 模型似然得分都在整个的似然得分空间中处于中上的位置，如图 3-7 所示。在这种方式下，对一段语音 OOV 模型的似然得分和不会是最佳的匹配者，但也总是排在局部最优匹配的前几位，只有当一段语音同关键词候选最佳匹配时，关键词候选才可望在 OOV 模型的竞争下胜出。

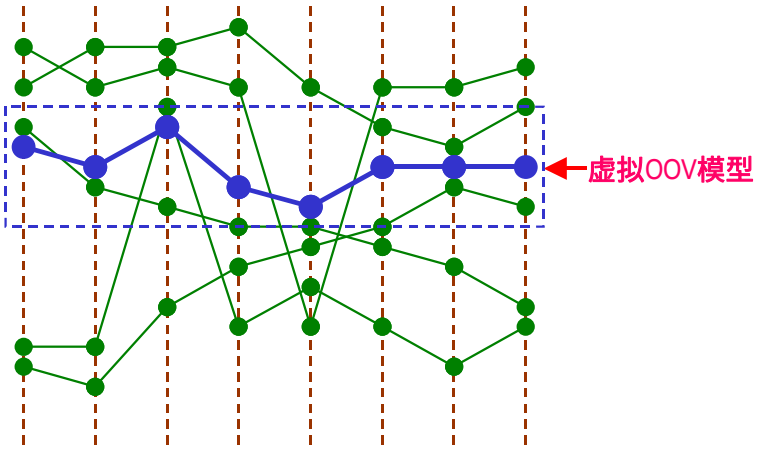


图 3-7　虚拟 OOV 模型在得分空间中的位置

3.3.5 进一步的分析

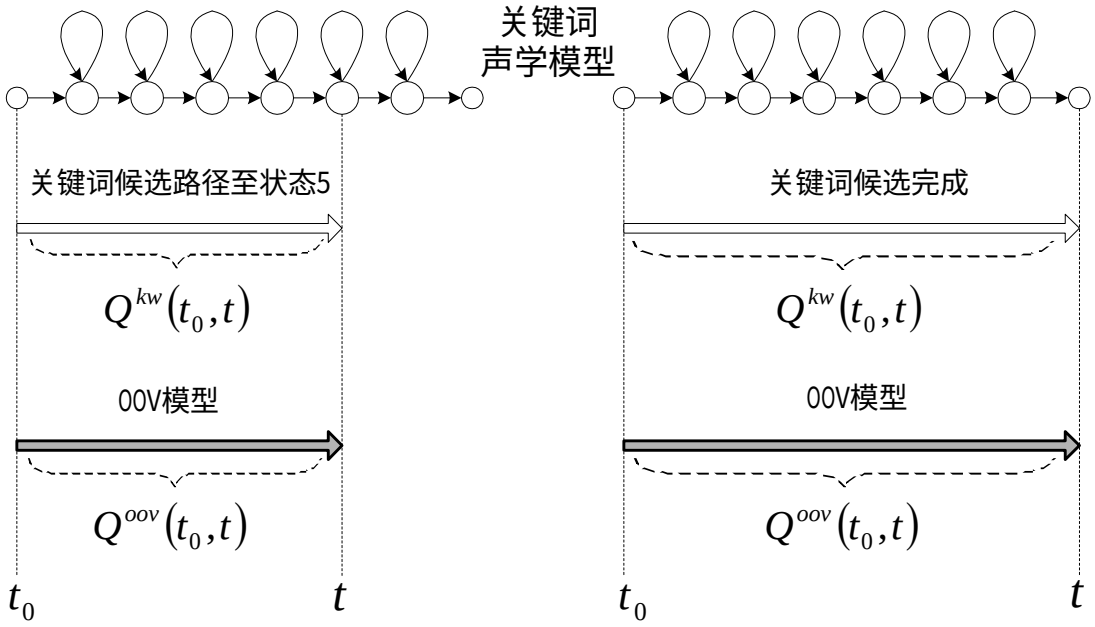
虚拟 OOV 模型也象其它关键词和补白一样放在搜索空间中进行搜索。在和关键词候选进行竞争这点上，虚拟 OOV 模型与补白模型非常类似，但它与补白模型的一点主要不同之处在于：补白模型可以出现在识别结果中，即吸收语音段，而虚拟 OOV 模型只能用作关键词动态确认，不能在最后的识别结果中出现，换句话说，只能由补白模型来吸收语音段而虚拟 OOV 模型不行。

虽然关键词动态确认技术与传统的关键词两阶段确认方法目的一致且思路相近，但效果并不相同，因为两者的算法理论和应用方法都有很大不同。关键词动态确认技术采用一个虚拟 OOV 模型，通过在搜索过程中和关键词候选的声学似然得分相比较来确定关键词候选是否通过确认；而传统的关键词两阶段方法则是在搜索过程结束后，针对关键词候选又提取出多个置信度特征，与训练得到的置信度模型一起判断出是否通过确认。两者功能并不重叠，所以在应用关键词动态确认技术后，同样可以在识别结束后继续使用传统的关键词确认算法，不会影响确认功能的发挥。

基准关键词识别算法中,使用子词拼接出的 418 音节作为补白模型集，在环境噪音比较低的实验室环境和采用朗读式发音时，关键词外的语音段可以准确地被匹配为实际发音音节对应的音节补白模型。但在有一定环境噪音下或说话人采用自然发音时，情况就大不相同了，虽然在特征提取模块加入了去噪的功能，而且也使用了相对自然发音的训练集来训练声学模型，但训练集和测试集之间的差异必然存在。而经过对实验结果的分析，误警往往容易在差异较大的语音段发生，所以差异的存在必将导致关键词识别算法的识别性能下降。

误警容易在差异较大的语音段发生，主要原因为：当待识别样本中有一部分语音段因为各种原因与训练环境不匹配时，所有状态对这个语音段的似然得分就都失去了准确性。对一个识别特征向量来讲，因为特征中包含一些因素在所有的训练样本都找不到，例如外部环境噪音，很可能大部分状态对这个特征向量的似然得分都很小，而且相差不大；而 OOV 模型的似然得分因为是计算最优的几个似然得分的平均值，所以仍然会相对较高。对一段语音来讲，补白模型的似然得分就都会比较小，这时补白模型往往就不能正常工作，因为关键词候选相对补白模型有较高的加权值，所以这时误警会容易发生；而 OOV 模型对这段语音的似然得分和是一组相对较高的似然得分求和得到，肯定是一个相对较高的分数，由其来对关键词候选进行确认就可以保证误警不会轻易发生。这就是虚拟 OOV 模型能够完成不影响识别率的情况下降低误警率的主要原因。

虽然上文主要针对基准算法——音节补白模型法进行分析，但分析的结论可以推广到其它使用集外补白或子词补白的关键词算法中，在这些算法中都存在着和音节补白模型法中一样的问题：如果测试样本与训练集的差异较大，补白模型往往就不能正常工作。只要是使用集外补白或子词补白的关键词算法，采用虚拟 OOV 模型的关键词动态确认技术就都可以取得良好的效果。



(a) 关键词候选未结束时　　(b) 关键词候选刚结束时

图 3-8 关键词候选和 OOV 模型

3.4 关键词动态确认的具体实现

加入关键词动态确认的关键词识别算法的整体流程分为以下几步：

1) 在 t 时刻，让所有活动状态对特征向量打分；

2) 用公式(3-15)计算虚拟 OOV 模型对当前特征向量的似然得分；

3) 完成关键词和补白模型词内部和词间的路径扩展；

4) 应用关键词动态确认，对未结束和刚结束的关键词路径进行 OOV 模型剪枝；

5) $t=t+1$如果 t 为最后一帧则输出识别结果，否则转 1）

从上可以看出，在关键词动态确认时，将虚拟 OOV 模型放在搜索空间中和其它的关键词和补白模型一起进行识别打分。在执行确认时并不考虑关键词候选路径总体的路径似然得分，仅仅关注关键词候选的似然得分，即搜索路径在关键词候选对应的语音段内累计的似然得分和，关键词候选的历史信息和权重并不考虑。如图 3-8 所示，关键词动态确认的具体实现分为以下两步来完成：

1) 关键词候选未结束时

关键词候选路径即使还没有将整个关键词声学模型走完，就已经可以对其它的关键词候选产生影响，如果可以尽早地将置信度比较低的关键词候选路径排除掉，则可以避免它对其它关键词路径的影响。采用关键词候选现有的似然得分与 OOV 模型在相应语音段内似然得分相比较来进行确认。因为关键词候选路径还没有走到最终一个状态，所以它最终的关键词候选得分还不得而知。很可能出现一个关键词候选在前几状态的匹配的不好，但在最后几个状态又匹配的很好，最终整体的得分还是很高，为了避免这种关键词候选路径在最初匹配不好时就被排除掉，在关键词候选未结束时，采用束剪枝的策略：

$$\text{当}\qquad Q^{kw}(t_0,t) < Q^{oov}(t_0,t) - f_{beam}^{oov} \tag{3-17}$$

将关键词候选路径剪枝掉，其中 f_{beam}^{oov} 为 OOV 模型束剪枝的剪枝宽度。

2) 关键词刚结束时

当关键词候选刚结束时，立即对其进行关键词动态确认，用关键词候选的似然得分与 OOV 模型在相应语音段内似然得分相比较，关键词候选似然得分较大则认为通过确认保留之，否则认为没有通过确认将其删除。如下公式所示：

当　　$Q^{kw}(t_0,t) < Q^{oov}(t_0,t)$　　(3-18)

将关键词候选路径剪枝掉。

采用以上两个步骤完成关键词动态确认。

关键词动态确认的拓扑结构如图 3-9，在图 3-2 音节补白模型法的拓扑结构上增加了虚拟 OOV 模型，连接虚拟 OOV 的两条虚线弧表示虚拟 OOV 模型的作用是只进行关键词动态确认而不是象其它补白模型一样在识别结果中体现。

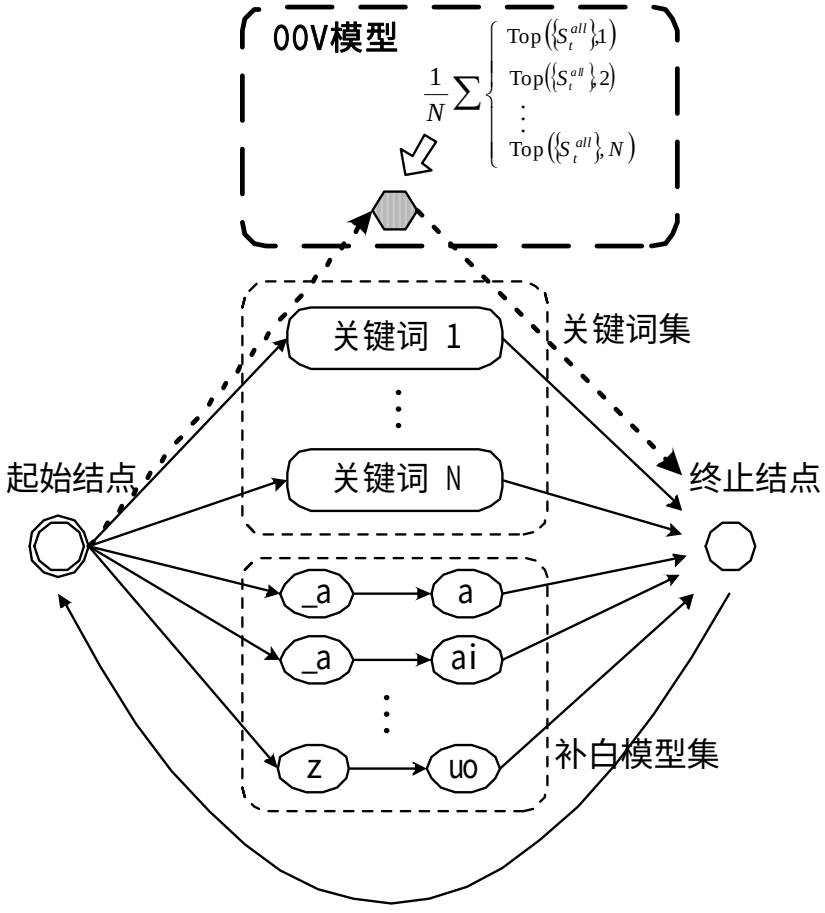


图 3-9 关键词动态确认的拓扑结构

3.5 实验结果与分析

从上文的分析可以看出，虚拟 OOV 模型推导过程中为了减少计算量，使用了大量的简化，这些简化是否能够达到预期的效果，必须通过实验来进行验证。语音识别是一门实验科学，实验一直是研究中至关重要的一个环节，有很多成熟的技术都是通过实验来寻找规律最终总结出来的，例如：MFCC 特征的提取，可以说只要能够经受的住实验的考验就是成功的理论。本节就要采用大量的实验来验证关键词动态确认在实际应用中的效果。

3.5.1 实验设计

在和口语对话系统 *EasyFlight* 和 *dEar-Attendant* 类似的电话信道上测试关键词动态确认技术的识别性能，实验设计如下：

关键词表：采用国内 A 股和 B 股中所有上市股票的名称，所有多音字都作为不同的关键词，共 1389 个关键词；

测试数据库：包含 10 男声 10 女声数据，每人 100 句共 2000 句。采用 8k 采样率，通过电话采集；每句测试样本中必包含一至三个关键词；

实验步骤：分为以下两个

(1) 比较基准算法在加入关键词动态确认前后的识别性能，以测试关键词动态确认技术是否有效；

(2) 比较在其它关键词算法中加入关键词动态确认前后的识别性能，以测试关键词动态确认技术是否有较好的可推广性和适应性；

评测指标：以 ROC 曲线为主要的评测指标，在 ROC 曲线的坐标内，曲线越靠上方、下方的面积越大，说明曲线代表的关键词识别算法的识别性能越好。

3.5.2 实验一：关键词动态确认技术的性能

图 3-10 是基准算法——音节补白关键词识别算法在加入关键词动态确认前后的 ROC 曲线图。在不失代表性的前提下，每种算法各取了十个

不同的关键词和补白模型权值测试了检出率和误警率并绘出了各自的ROC曲线，图中两条曲线上具有相同标号的点所取的权值相同。

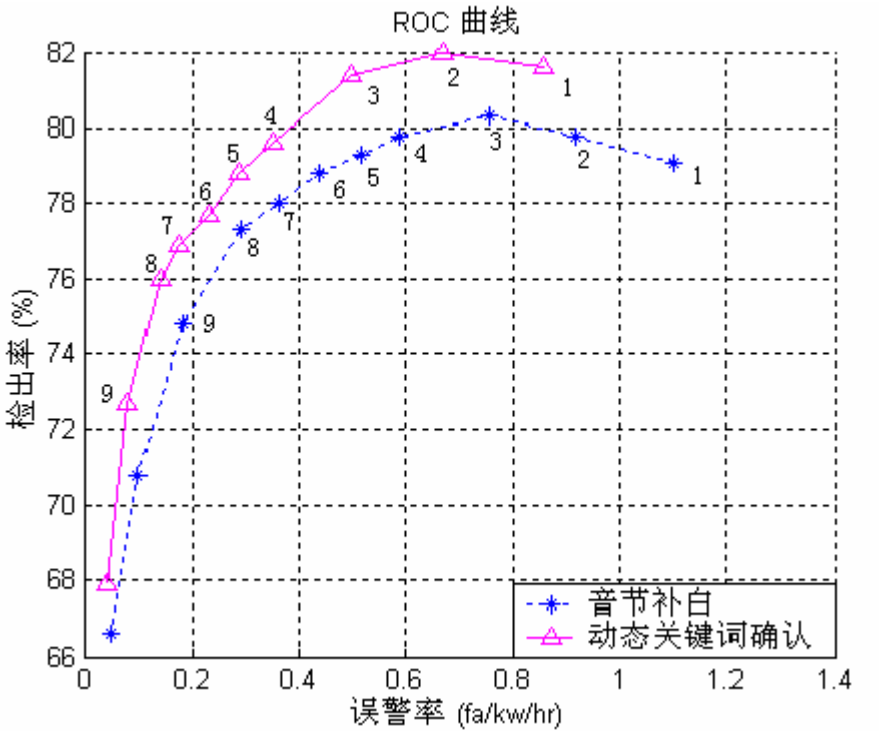


图 3-10 音节补白模型法上关键词动态确认技术的性能比较

从图 3-10 中可以看出：不同的权值范围中关键词动态确认的效果不同。1) 在基准算法误警率比较高时，即权值取 1、2、3 时，使用关键词动态确认技术的新算法不仅明显降低了误警率，而且使得检出率也有所上升，漏检率下降了 10%左右；2) 在基准算法误警率适中时，即权值取 4、5 时，使用关键词动态确认后误警率明显下降的同时检出率变化不大，基本持平；3) 在基准算法误警率较低时，即权值取 6、7、8、9 时，使用关键词动态确认后误警率明显下降的同时也伴随着检出率的下降。这主要因为：在误警率高时，关键词动态确认可以去除大量误警关键词候选并避免它们对正确关键词候选的影响，从而使得被正确检出的关键词增多，虽然关键词动态确认也会使得一小部分正确候选被去除，但最终仍然提高了检出率；在误警率低时，本身误警关键词候选对正确关键词候选的影响就比

较少，在关键词动态确认后，最终的检出率会有所下降，但从 ROC 曲线来分析，识别性能仍然优于基准算法。

从以上的分析可以得出：加入关键词动态确认技术的算法在整体识别性能上要明显优于基准的音节补白模型法，而且在误警率较高时，采用关键词动态确认在使误警率大大下降的同时，使漏检率下降了约 10%。从实验结果上验证了关键词动态确认技术确实具有明显的效果。

3.5.3 实验二：关键词动态确认技术的可扩展性

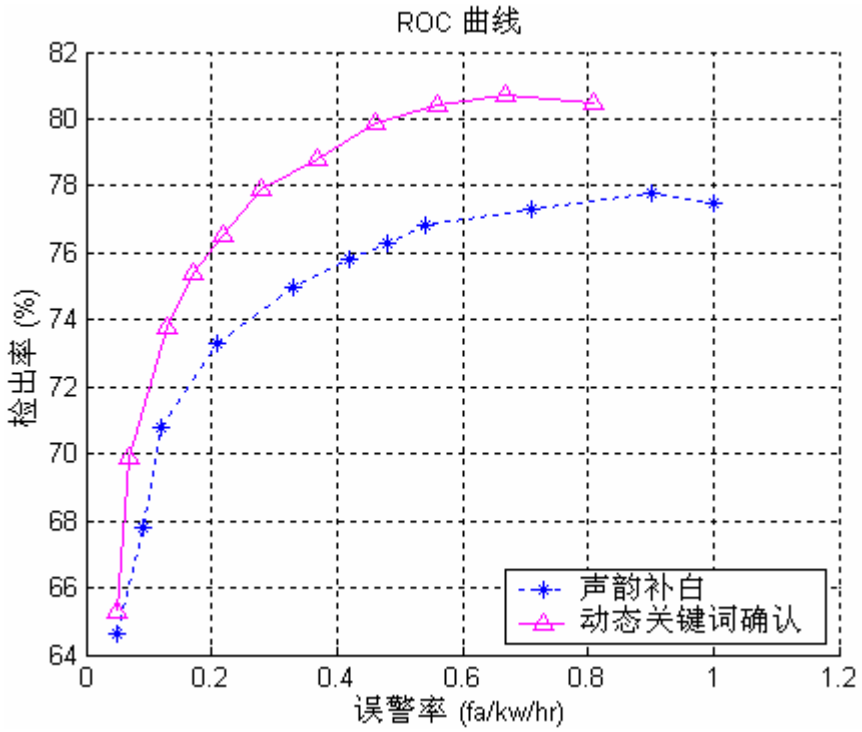


图 3-11 声韵补白模型法上关键词动态确认技术的性能比较

音节补白模型法是子词关键词识别策略的一种，它使用声学模型中的子词模型来拼接出补白模型，上节中的实验结果说明关键词动态确认技术在子词关键词识别策略中是有效的。由于本文中关键词动态确认的提出就是以音节补白模型法为分析基础的，在其基础上取得明显效果不足为奇，但关键词动态确认在其它关键词算法上是否仍然有效呢，即关键词动态确

认的可扩展性如何，还需要进行进一步的实验进行验证。

为了测试关键词动态确认技术的可扩展性，选取了属于集外关键词识别策略中的声韵补白模型法进行实验。和音节补白模型法中一样声韵补白模型法仍然采用上下文相关子词模型 tri-XIF 作为声学模型拼接出关键词模型。但不同之处在于声韵补白模型法专门为补白模型集训练了一个上下文无关的扩展声韵母模型，共有 66 个补白模型在搜索网络中和关键词竞争。实验结果如图 3-11 所示。

从实验结果中可以看出，结合关键词动态确认技术的声韵补白模型法的 ROC 曲线明显高于声韵补白模型法的 ROC 曲线，说明动态确认技术在声韵补白模型法中确实是有效的。比较图 3-10 与图 3-11，可以看出声韵补白模型法比音节补白模型法的识别性能要差，两者均加入关键词动态确认技术后，声韵补白模型法的识别性能仍然略差，这是因为组成音节补白的 tri-XIF 训练的更精细，识别能力更强。但图 3-11 中两条 ROC 曲线的走向与图 3-10 中 ROC 曲线的走向非常类似，这可以说明动态确认技术在声韵补白模型法中起到的效果与在音节补白模型法中起到的效果一样。由此我们可以推断关键词动态确认在子词关键词识别策略和集外关键词识别策略中均有效。

3.5.4 分析与讨论

从上节的实验结果与分析中可以看出，关键词动态确认技术可以应用在属于集外关键词识别策略和子词关键词识别策略的几种典型的关键词识别算法，但关键词动态确认技术不能应用在在线补白模型法中。这是因为在关键词动态确认中起核心作用的虚拟 OOV 模型的构建方法根本上是从在线补白模型法中的在线补白模型的构建方法上借鉴而来的。在线补白模型的功能和作用在很大程度上已经涵盖了虚拟 OOV 模型的功能，所以在线模型法中不能使用以虚拟 OOV 模型为算法核心的关键词动态确认技术。但在将来的工作中，如果研究出与在线补白模型的知识背景不相同的 OOV 模型，还是可以将关键词动态确认的思路应用在在线补白模型法中，而且一定可以使算法的整体性能得到明显提高。

3.6 小结

本章通过详细分析关键词识别算法的具体实验结果，发现关键词的漏检错误大多数都与另一错误关键词的误警错误有关这一现象，并提出关键词动态确认的概念，它在搜索过程中就引入虚拟 OOV 模型来对产生的关键词候选进行确认，及早的将不正确的关键词候选剪除，从而避免其对正确关键词候选的干扰影响。实验表明：采用关键词动态确认技术后关键词检出算法的识别性能有明显提高，在误警率相同的条件下，误识率下降了约 10%。

总之，将关键词动态确认技术应用于集外关键词识别算法和子词关键词识别算法，有效地在降低误警率的同时提高了关键词识别正确率，使整体性能有明显提高，具有理论价值和实际意义。

关键词动态确认技术虽然可以明显改善关键词识别算法的整体识别性能，但并没有使得关键词识别算法的识别性能有一个质的飞跃，改进后的关键词识别算法性能仍没有达到口语对话系统中对语音识别模块的要求。如果想使提高关键词识别的性能达到口语对话系统的要求，必须引入新的知识源来指导识别。在下章中将详细介绍作者在这方面的研究工作，将口语对话系统中的语境知识提取出来，使用语境知识来对关键词识别进行指导。

第四章　语境知识指导的关键词识别算法

现有的口语对话系统的语音识别算法主要采取三种识别策略：有限状态网络指导的语音识别、基于小领域 N-Grams 语言模型的连续语音识别和关键词识别。有限状态网络和 N-Grams 语言模型都用来描述语言知识信息，可以取得比较好的识别性能，但关键词识别算法以其良好的鲁棒性而获得更多关注。虽然关键词算法在众多语音研究工作者的共同努力下已经取得很大的进步，但现有的关键词识别性能仍不能满足口语对话系统的高识别率、高鲁棒性的需要。

本章的主要内容在于介绍作者在如何使用语境知识来提高关键词识别性能的方面所作的研究成果，提出了语境知识指导的关键词识别算法，利用识别自动机来表述对话管理器给出的期待焦点下的语境知识，并用其来指导关键词识别。实验表明：语境知识指导下的关键词检出算法具有很高的识别性能和鲁棒性，基本能够满足口语对话系统的需要。

本章的内容安排如下：第一小节简要介绍口语对话系统中几种识别策略，以及各自的优缺点；第二小节介绍对话语境知识对识别的指导作用；第三小节提出语境知识指导下的关键词识别策略；第四小节则以 *EasyFlight* 和 *dEar-Attendant* 两个口语对话系统为例，详述对话语境知识对关键词识别的实现；第五小节给出实验结果和分析；最后第六小节进行本章总结。

4.1 几种语音识别框架及其在对话系统中的应用

使用计算机为工具进行自动语音识别的历史可追溯到 20 世纪 60 年代，之后的十几年内，小词汇量、特定人、孤立词的语音识别技术得到了实质性的进展；70 年代后期，语音识别开始了从特定人到非特定人、从孤立词到连接词、从小词汇量到大词汇量的扩展；自从 80 年代开始 HMM 模型框架被引入语音识别领域以来，语音识别技术得以向大词汇量连续语音识别方向发展，出现了大量识别率比较高的系统，并使语音识别技术逐渐

走向实际应用。近年来，口语对话系统的研究越来越受到重视，传统的语音识别技术面临新的挑战，由于口语对话中的自然语音现象十分复杂，导致现有的语音识别算法的识别性能大大下降[60]。

在口语对话系统中采用三种主要识别策略：有限状态网络指导下的语音识别、基于小领域 N-Grams 语言模型的连续语音识别和关键词识别。这三种策略各有优缺点和应用背景。

有限状态网络指导下的语音识别：

有限状态网络指导下的语音识别也叫模板匹配(Template Matching)，特点是使用有限状态网络来描述语言知识信息。专家经过对训练语料库的分析，总结出一个上下文相关文法或上下文无关文法来描述在口语对话系统中所有可能出现的语言现象，并生成一个对应的有限状态网络，在这个网络的指导下进行语音识别，语音识别结果完全符合有限状态网络。

有限状态网络具有高度的预测性，在其指导下的语音识别策略具有识别率高的优点，当输入的语音样本完全符合有限状态网络，算法可以取得很高的识别率；但有限状态网络也具有比较明显的缺点：低鲁棒性，因为语音识别结果完全符合有限状态网络，所以当总结出的有限状态网络不完善或输入的语音样本中包含一些无法预知的口语现象时，识别结果中必然会发生一定错误，会大大降低识别率。虽然可以采用一种从逐步添加的语料库中抽取规则并组成有限状态网络的思路[75]，这种思路试图通过逐步分析在应用中遇到的有限状态网络无法描述的语言现象，并在应用中逐步完善有限状态网络。但低鲁棒性仍然是该法难以得到广泛应用的瓶颈。

有限状态网络指导下的语音识别策略主要应用在比较简单的口语对话系统中，这是因为其描述能力比较低，无法准确概括出复杂口语对话系统中可能出现的语言现象，而且在复杂系统下的网络结构难以得到，随着网络规模的扩大性能也难以保证。但由于有限状态网络的生成比较简单快速，不需要大批量的语料库，在比较简单的口语对话系统中仍然有一定的应用价值。

基于小领域 N-Grams 语言模型的连续语音识别：

基于小领域语言模型的连续语音识别策略和有限状态网络指导下的语音识别策略主要的区别就在于：采用 N-Grams 语言模型来描述可能出现的语言现象，而不是使用有限状态网络。另外，与有限状态网络需要由专家来总结不同，N-Grams 语言模型需要从大批量语料库中经训练得到。许多系统使用基于 N-Gram 的连续语音识别策略，它们的差异往往体现在词表和 N-Gram 的规模上。在 August[43]中，Gustafson 使用了规模为 500 的词表，以及基于 70 个词类和 229 个词类对的词类 Bi-Gram。旅游信息系统 LOADSTAR[17]使用了词和词类的混合模型，词类的规模为 733。相对基于词的模型来说，词类语言模型的物理(句法)意义更明显，而且能够避免出现模型稀疏问题。

与有限状态网络类似，小领域 N-Grams 语言模型也具有较强的预测性，在其指导下的连续语音识别策略也具有识别率高的优点。而且因为 N-Grams 语言模型在统计时对数据稀疏问题采取了一定的平滑手段[80][81][82][83]，使得其鲁棒性要好于有限状态网络，但对于 OOV 问题和口语现象，N-Grams 语言模型也无法很好的处理。除了鲁棒性较低这一缺点外，N-Grams 语言模型的另一个限制条件就是需要大量的数据库来训练得到，这需要花费大量的人力、物力来制作，在很多口语对话系统中都不容易得到规模足够、覆盖比较全面的数据库。

小领域 N-Grams 语言模型指导下的连续语音识别策略主要应用在规模比较大，比较复杂的口语对话系统中。这主要是因为 N-Grams 语言模型所需的训练数据库比较大，代价比较高，在比较简单的口语对话系统中用有限状态网络就可以比较有效地完成识别。

关键词识别：

关键词识别策略在上文中介绍的比较详细，其特点在于可以忽略关键词以外的语音部分，口语现象可以被忽略掉而不影响关键词的识别性能，具有很高的鲁棒性，而且和应用背景无关，可推广性强，这也是目前关键

词识别策略在口语对话系统中得到广泛应用的主要原因。但如上文所述，关键词识别策略的最大缺点在于识别率相对较低，在三种主要识别策略中识别性能最差。如何提高关键词识别策略在口语对话系统中的识别性能，成为当前新的研究热点。

表 4-1 列出了三种识别策略各自的优缺点，可供比较。综上所述，口语对话系统中的语音识别存在多种研究思路，取得了一些成果，但效果并不理想，均没有满足口语对话系统对语音识别的要求，另外缺乏应用于特定领域下的针对性。

表 4-1 口语对话系统中三种语音识别策略的对比

比较内容	有限状态网络	小领域 N-Grams 语言模型	关键词识别
识别率	高	高	相对较低
鲁棒性	差	中等	好
对口语现象的解决	差	差	好
是否容易实现	容易	困难	容易

4.2 对话语境知识对识别的指导作用

在口语对话系统中，语境知识对语音识别的指导主要分为两方面：对话主导策略和高层背景知识。

4.2.1 对话主导策略对语音识别的指导

口语对话系统中的对话管理器存在三种管理策略，一是用户主导(User Initiative)，是指以用户的意图来主导对话的内容及进程，以用户发问系统回答的形式出现；二是系统主导(System Initiative)，是指主要由系统来确定下一步的对话内容，以系统发问用户回答的形式出现；三是混合主导(Mixed Initiative)，是指用户和系统双方根据需要均可以采取主动的对话方式，以平等交谈的形式出现。

系统主导的特点是用户只能根据系统的发问来回答，对话进程完全由

系统控制，用户几乎没有摆脱预先设定模式的自由。这在一些简单的场合，或者对交谈内容的可靠性要求很高的场合是可以接受的，但是不适合用于要求具有一定友好度的公共服务场合。用户主导则是另一个极端，用户可以在领域限定的唯一约束下，自由地选择自己感兴趣的话题，这时系统被动地跟随。应该说这是一种理想化的模式，但是现有的语音识别和语言理解技术不足以支持这种模式。

有鉴于此，大多数口语对话系统均采用混合主导的对话策略。在这样的对话过程中，用户可以在需要获取某项信息时主动提出查询请求和查询条件，而系统在不能给出结果时会主动反馈用户，比如说提示查询条件不全，提示用户放宽收紧查询条件，或者给出另外的建议等。混合主导的对话策略还基于这样一个重要的前提，那就是用户参与对话有明确的获取信息的目的，因此，他们在使用口语对话系统时会采取主动配合的态度，以尽快实现对话目的。这个假设不光用于对话管理器的设计，对于语音识别和语言理解的设计也有指导意义。

在每回合识别用户的输入之前，口语对话系统可以预测出用户可能谈论的所有话题。当采用用户主导策略时，这个预测话题集就包括所有用户可能说的内容；但当采用系统主导策略时，则可以将预测话题集缩小到相当小的范围内；即使采用混合主导策略时，随着对话的逐渐深入，系统也能够将预测的话题范围逐步缩小至一定程度。所以，当系统采用系统主导策略或混合主导策略时，如果语音识别器可以和对话管理器相互配合，根据预测话题集可以缩小当前识别所用的关键词集和规则集的大小，一定可以明显提高语音识别算法的识别效果。在下文中，我们将每回合的预测话题集称为每回合的**期待焦点**(expected focus)。

4.2.2 高层知识对语音识别的指导

微软的研究人员曾做过一个实验，用人耳与机器同时来听一批不成句子的词，当使用上下文相关音素模型时，发现机器识别的效果比人耳还要好点。但是在实用中，机器的语音识别远不及人耳，原因是人的大脑在加工处理信息的过程中，把语义的知识及语法的知识与听来的语音信号一块

处理，把事实上并不清晰的信号处理成了合乎语义并合乎语法的句子。这个例子很生动地说明了语境知识对语音识别的巨大作用。

从人类的语言感知过程也可以发现高层知识对识别的影响。主要有以下三种[84]：

1) 句法语义知识：对于一种先前没有学习过的语言，听话者不具备任何知识，那么听到这样的语音之后会觉得难以捉摸，语句中的音素也无法分辨清楚；而如果是一门系统学习过的外语，那么不光可以辩明组成句子的词串，还能分辩其中的音素。

2) 方言知识：平时生活于某一方言音区的人，如果不加准备地听到另一个其所熟悉的方言时，往往会觉得不知所云；但如果事先知道将要听到这种方言，则会跟正常情况那样能够听得很自如。

3) 场景语境知识：从一个专业场合转到另一个专业场合，为能够更好地参加讨论，人们也往往需要调整思维以适应变换了的讨论主题。

上述这些人类语言感知的现象，对自动语音识别是一个启发：如果其它层面的知识能够在识别中得以应用，则对语音识别的准确率肯定会有所贡献。

从上文的分析可以看出，在口语对话系统中对话主导策略和高层背景知识都可以为语音识别模型提供有力的支持，只要能够将这些语境知识充分的利用起来，和语音识别器有效的结合在一起，一定可以大大提高语音识别的性能。

从表 4-1 的对比中可以看出，在口语对话系统所采用的三种识别策略中，利用有限状态网络或 N-Grams 语言模型来指导的连续语音识别策略即是这种思路的具体体现，它们试图采用有限状态网络或统计语言模型来描述整个系统的背景知识，在识别过程中即发挥其指导作用，取得了良好的识别性能。但利用语境知识来指导语音识别也有其局限性，这就是不可能将所有的语境知识完全表示出来。如果用有限的语境知识来指导语音识别，当发生用户的输入不在已有的知识背景之内时，语音识别器很可能仍

然按照已有的语境知识来识别，最终发生识别错误。

口语对话系统所采用的另一种识别策略——关键词识别策略则可以很好地解决鲁棒性的问题，但由于其缺乏语境知识的支持，识别性能不理想。从直观上来感觉，如果将语境知识引入关键词识别策略，新的算法必然可以取得高识别率和高鲁棒性均佳的识别性能，图 4-1 中的小例子则很有代表性的说明了这点。

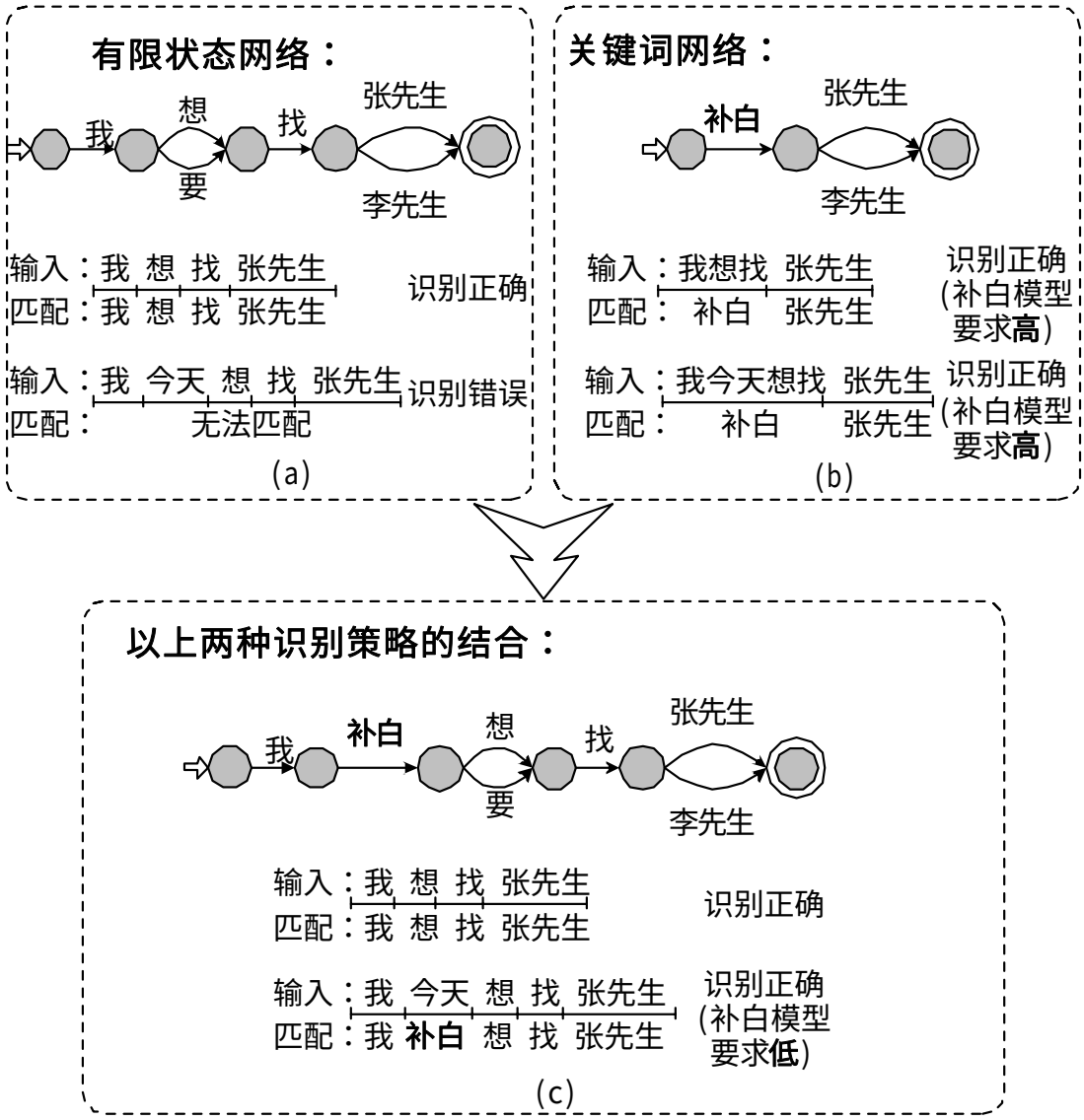


图 4-1 语境知识的指导作用举例

图 4-1 中给出了一个非常简单的人名查询的对话识别。图 4-1(a)描述的是有限状态网络指导下的语音识别策略的识别结果。当待识别语句完全符合有限状态网络时可以得到正确的识别结果，但当待识别语句中出现集外词，包含一段有限状态网络之外的语音段时，就很有可能得不到正确的识别结果，说明了有限状态网络具有识别率高、鲁棒性差的特点。图 4-1(b)描述的是关键词识别策略的识别结果，无论待输入语句中是否有集外词都不会因为拓扑结构的问题而得不到正确识别结果。但这对关键词识别算法的能力提出了挑战，因为需要补白模型来精确匹配除关键词以外的语音段，所以对补白模型的能力要求很高，这也是在实际应用中关键词识别策略鲁棒性好而识别率偏低的主要原因。而图 4-1(c)中描述的第三种识别策略，将关键词识别策略中的补白模型结合在有限状态网络中。当待识别语句符合有限状态网络，有限状态网络发挥作用，可以取得很好的识别结果；当有集外词或口语现象发生时，补白模型发挥作用吸收有限状态网络以外的语音段，同样可以取得很好的识别结果。相对于关键词识别策略，新识别策略的补白模型需要匹配的语音段很少，所以对补白模型的要求不高。在吸收了有限状态网络和关键词识别两种识别策略的优点后，新算法可以达到识别率高鲁棒性好的最佳识别性能。这个例子充分说明：使用语境知识来指导关键词识别，可以取得更好的整体识别性能。

4.3 语境知识指导下的关键词识别策略

基于上节的分析，提出对话语境知识指导下的关键词识别策略。它的中心思想是，在混合主导的对话管理策略下假设用户对交互过程主动配合，利用每一对话状态下的语境知识指导关键词识别算法，以提高每一对话状态下感兴趣单元的识别率。对话语境知识对语音识别的指导机制可如此表示：根据当前的期待焦点，提炼出期待焦点下的语境知识，总结出适合当前对话的自动机模板，并与关键词识别算法有机的结合在一起，在其框架下进行语音识别。

4.3.1 语境知识的表示形式

采用何种形式来表示语境知识成为新的识别策略中一个关键所在。选择方法的主要原则有两点：一、将语境知识最有效地表示出来并且能够充分发挥其效力；二、可以和关键词识别策略简单易行地结合在一起。通常有两种背景知识表示形式：有限状态网络和 N-Grams 语言模型，表 4-2 列出在表示语境知识能力上它们各自的优缺点。

表 4-2 两种语境知识表示策略的对比

比较内容	有限状态网络	N-Grams 语言模型
获取方式	专家总结或数据统计	数据统计
表示能力	强	强
对训练数据需求	相对较小	相对较大
是否容易训练得到	相对容易	相对困难

从表 4-2 的分析可以看出：有限状态网络的灵活性比较好，需求的训练数据要求比较小。这点在实际应用中非常重要，因为这意味着构建一个有限状态网络比构建一个统计语言模型在工作量上容易很多。有限状态网络可以由专家在对语料库进行分析的基础上很快的总结出来，而统计语言模型则需要比较充分的训练数据才能得到。在很多比较小的口语对话系统中，例如 *dEar-Attendant*，根本不需要也不值得花费大量精力来训练统计语言模型。

另外，语境指导下的关键词识别策略的中心思想还在于提炼出不同期待焦点下的语境知识来指导关键词识别，这就要求有多个期待焦点的语境知识需要被表示出来。如果采用统计语言模型，即使具备了规模较大的训练语料库，也无法将训练语料库中的各个训练样本按照实际发生的概率分配到不同的期待焦点下，因而得不到不同期待焦点下的统计语言模型。但是采用有限状态网络，则可以由专家针对语料库进行分析，对每个期待焦点都可以总结出各自的有限状态网络。

综上所述，语境指导下的关键词识别策略中采用有限状态网络来表示

不同期待焦点下的语境知识是可行的。

4.3.2 语境知识指导机制的细化

对于特定领域下的对话系统，可对语境知识指导机制进行细化以明确处理流程。

1) 期待焦点

对话管理器一般包含对对话历史的管理和数据库的查询，对话历史和查询结果共同构成当前对话状态，对话状态将决定对话管理器的应答。在每回合识别用户的输入之前，对话管理器可以得到当前的期待焦点。尤其在混合主导的对话策略中，系统应答在必要时会主动向用户给出提示或提问，这时对下一回合用户的语句内容将有明显的期望，期待焦点的内容可以限制在一定范围内。对话管理器可以以自动机（或规则）的形式在对话状态间进行切换。

2) 活动规则集

基于规则的语言理解方法用规则来描述语句，特定期待焦点下的语句可以用特定的规则集来描述，这些规则集即是语境知识的体现。对规则做标注，以确定其在不同期待焦点下的活动标志，当给定期待焦点时，由语言理解模块形成相应的活动规则集。当对话刚刚开始，期待焦点包含所有的话题，此时的活动规则集即是包含所有规则的全集。

在口语对话系统所使用的整个规则集中，在某一期待焦点下，活动规则集以外的规则形成非活动规则集。虽然活动规则集描述的是当前焦点所期待的内容，但由于鲁棒性的原因，非活动规则集也不能完全不考虑。活动规则集和非活动规则集共同构成期待焦点下的识别自动机。

3) 识别自动机

得到活动规则集之后，将活动规则集中所涉及到的关键词选择出来形成活动关键词表，并与非活动规则集中出现的关键词表一起，转换为有限状态自动机形式，称之为识别自动机。识别自动机即是期待焦点下的语境

知识的表示形式。

4) 搜索网络

将补白模型加入到识别自动机的适当位置，形成一种具有补白模型的扩展有限状态网络，这个扩展有限状态网络即是最终的搜索网络，识别时以此搜索网络指定的路径进行搜索。

至此，对话语境对识别的指导机制可以明确化成图 4-2 所示的构架，其中对话管理器负责产生当前焦点，语言理解器生成识别自动机和确定活动词表及非活动词表，语音识别器在该自动机约束下进行识别。

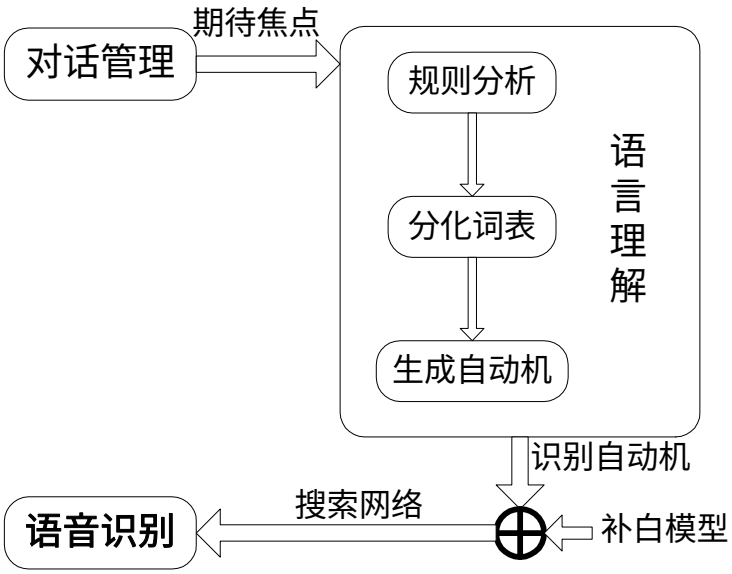


图 4-2 语境知识指导机制

应该说，图 4-2 只是语境知识指导机制的一种实现模式，并不排除可以有其它的实现形式，因为对话语境可以不只是期待焦点，而对话语境知识也不一定用活动规则集的方式体现。

4.4 语境知识指导下关键词识别策略的实现

本节将给出对话语境知识指导下关键词识别策略的具体实现。部分内容将以 *EasyFlight* 为例来说明问题。

4.4.1 识别自动机的生成

1) 期待焦点的演进

口语对话系统采用基于主题树/森林[84]知识表示结构的对话管理方法。主题树是对话管理的基础，它用于表示和维护对话涉及的多个主题，对于一个特定主题能够完备地表示该主题的所有信息，并可用于表示对话过程中该主题的当前状态，领域内所有主题对应的主题树构成主题森林。主题树中节点信息来源于填充过程，包括来自用户表达的信息，以及数据库查询结果。

初始的期待焦点可以包括系统所包含的全部主题，因为此时系统不知道用户需求什么服务。随着对话的进行，根据对话管理器的状态转移，期待焦点随之进行变化，称之为期待焦点的演进。

期待焦点的演进包括两部分，其一是主题树内部状态演进过程，也就是随着主题树内部信息的变化而进行的状态转移。这又可以分三种情况：一、当主题树中某项必要信息缺乏时，系统会主动提出询问；二、当某项信息多于一种选择时，系统会提示用户明确其选择；三、一般情况下系统会期待用户的下一语句仍属于当前主题。期待焦点的第二种演进过程，是主题树之间的状态演进过程。这是在当前主题的各种条件均已确定之后，即将跳转到下一主题时发生的。根据口语对话系统的不同，一个主题完全结束后有可能还可以预测出用户下一语句所属的主题范围，也可能完全无法预测出，重新回到初始的期待焦点。

2) 焦点下的活动规则集

期待焦点可以根据语义上的差别分成几类，而通过使用语义文法可以用允许交叉的不同规则子集来表示不同期待焦点下的语义类。期待焦点下的活动规则集本质上是在当前对话状态下，用户最可能说的内容的抽象表示，这个抽象表示将在指导关键词识别时起到重要作用。

在 *EasyFlight* 中，期待焦点分成以下几类：地点、航班号、日期、时间、机型、航空公司、票数、身份证号码、“是/否”等。通过对规则进行

类别标注，将不同焦点下的规则区分开来，便形成焦点下的活动规则集。

规则类别标注的几个准则：

可区分性。必须使语义上不具有包含关系的语义范畴（文法符号）用不同的类别进行标注。而对语义上具有包含关系的语义范畴，则可用几种类别同时进行标注。

完整性。表示一个完整的语义类的规则要尽量分为一类。主要是为了保证所有该项语义类能够完全地被表示，而不会出现某些语句组成模式由于不在类别标注之内而不能被识别器所识别。

简单性。考虑到要应用于识别器，活动规则集如果过于复杂，则不太容易使用比较简单的识别框架。

3) 活动关键词表的得到

得到活动规则集之后，将活动规则集中所涉及到的关键词选择出来形成活动关键词表。活动规则集是在当前对话状态下用户最可能说的内容的抽象表示，则活动关键词表即是当前对话状态下，用户最可能说的关键词集合。活动关键词表中的关键词是当前对话状态下出现概率最高的关键词。例如在 *EasyFlight* 中，期待焦点处在查询航班的时间信息时，用来描述“打折”、“价格”等其它内容的关键词则不太可能出现在用户的语句中。但不能排除用户在当前话题没有结束时突然跳转话题，所以为了鲁棒性要求，所有的关键词都应该在最终的搜索网络中体现出来，只是活动关键词表中的关键词应该得到更多的重视。

4) 识别自动机的生成

有限状态自动机与正则文法的描述等价，对上下文无关文法的规则作类别标注时如果遵循简单性的原则，就可以保证选取的活动规则集满足有限状态自动机的条件。将期待焦点下的活动规则集用有限状态自动机来表示，就形成了期待焦点下的识别自动机。在这种自动机中，弧代表语义内容而结点代表连接关系。为了对规则作一些区别，可以根据训练数据或者由专家来决定，对规则加入概率，形成包含概率的有限状态自动机，概率

值将会在关键词识别时得到体现。

4.4.2 识别自动机与补白模型的集成

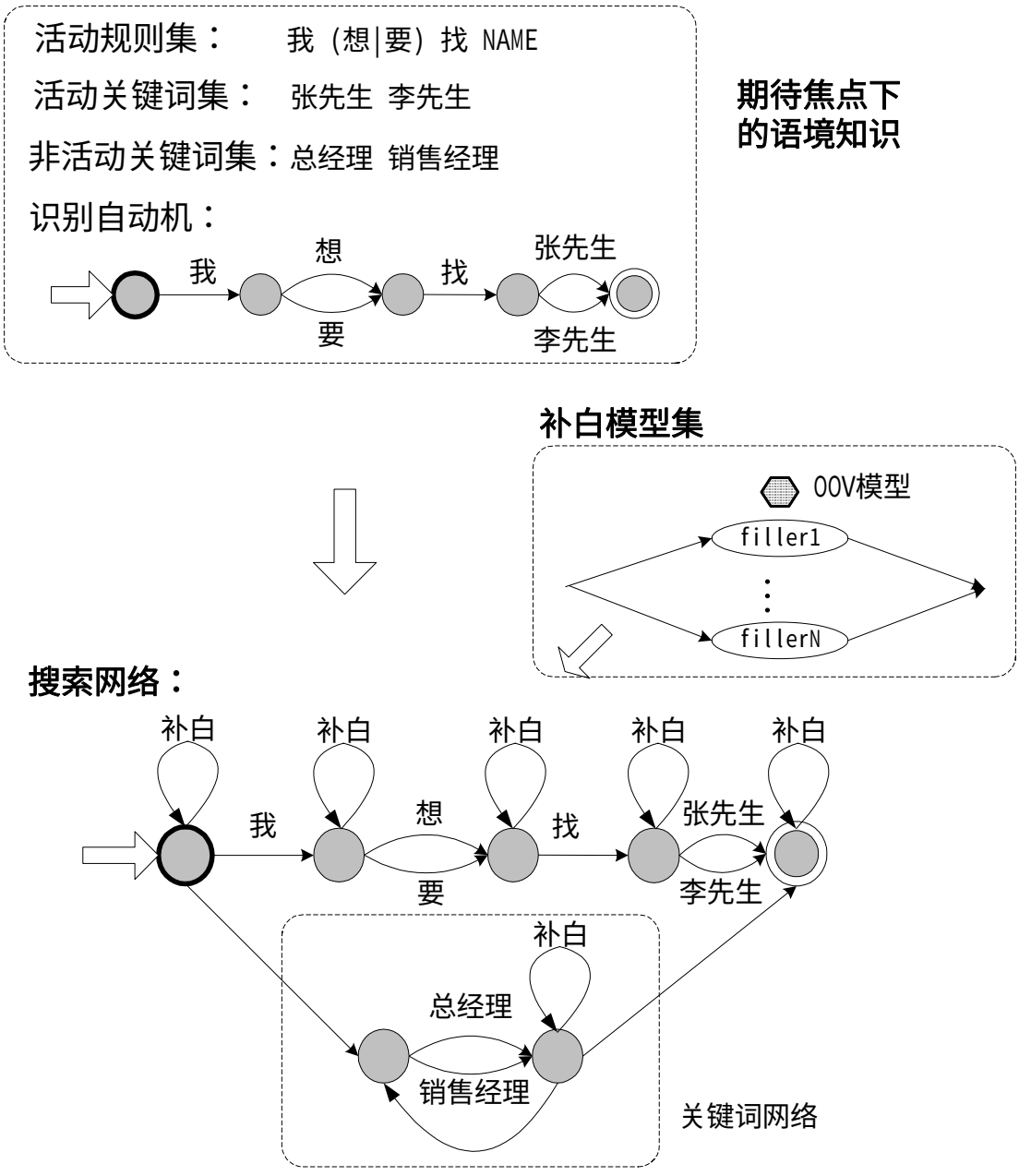


图 4-3 识别自动机与补白模型的集成

根据期待焦点下的语境知识生成识别自动机后，用其来指导关键词识

别，只需要将关键词识别策略中的补白模型与识别自动机结合后形成最终的搜索网络，识别在搜索网络下进行即可。最简单的结合方式是直接在识别自动机的任何结点上都加入补白模型。在识别自动机的任何结点上都加入一个代表补白模型的弧，弧指向自己，这表示在识别时，在每个结点上都可以按照补白模型扩展路径，也可以直接按照其它语义弧扩展路径。如果可以预测出在识别自动机的哪些结点可能出现口语现象，而另一些结点不可能出现口语现象时，则可以有选择的在可能出现口语现象的结点上加入补白模型，其它结点可以不加入补白模型。

为了保持关键词识别策略原有的鲁棒性，除了加入补白模型的识别自动机部分之外，再将原有的关键词识别网络也引入搜索网络，识别自动机与关键词识别网络之间是并列的关系。当用户输入符合当前的语境知识即符合识别自动机时，识别自动机在搜索中起主导作用；而当用户输入完全不符合当前的语境知识时，关键词识别网络起主导作用。在图 4-3 中举例说明识别自动机与补白模型的集成。

在得到整个搜索网络后，为了降低计算复杂度，可以运用有限状态网络的确定化技术减少网络规模，进而提高识别速度。确定化的中心思想是，从起始结点出发，将出弧符号相同的后继结点归成一个结点集，然后从该结点集出发，继续此归类过程，直至再没有新结点集出现为止。以所有结点集为新结点，生成新的自动机，它就是确定有限状态网络。通常情况下，有限状态网络都可以进行确定化，而确定有限状态网络的大小要比原有网络小的多。

4.4.3 性能分析

加入补白模型并不会给原有识别自动机指导的语音识别带来影响。类似于补白模型在关键词识别算法中的使用，在补白模型上会加一个比较低的权值。这就保证了当匹配语音中的识别自动机包括的部分时，补白模型匹配的概率得分会比较低，只有当匹配的语音段不在识别自动机中时补白模型的概率得分才会比较有优势。所以补白模型的加入并不会干扰原有识别自动机指导下的语音识别策略具有高识别率的优点。

在构建搜索网络时，必须保证在网络中有一条路径从起点到终点全部由补白模型组成。这样即使输入语句完全不符合识别自动机，算法等于退化到关键词识别算法，也能保证关键词的正确检出，具有较高的鲁棒性。

当用户输入符合当前的语境知识即符合识别自动机时，识别自动机在搜索中起主导作用，体现出其识别率高的特点；当用户输入部分符合语境知识，即输入语句中包含口语现象时，识别自动机负责指导识别使口语现象被识别自动机中的补白模型所吸收，同样可以取得很高的识别率；即使当用户输入完全不符合当前的语境知识时，关键词识别网络起主导作用，保持了原有鲁棒性好的优点，仍然可以取得一定的识别率。最终的算法可以说包含了有限状态网络指导下的语音识别策略和关键词识别策略两种算法的各自优点，互相弥补了各自的缺点，形成一种具有高识别率和高鲁棒性的算法。

综上所述，语境知识指导下的关键词识别策略基本能够满足口语对话系统中对识别器的要求

4.5 实验结果与分析

4.5.1 实验设计

在已经产品化的口语对话系统 *dEar-Attendant* 上进行实验，来测试语境知识指导下的关键词识别策略的识别性能。*dEar-Attendant* 是语音自动转接系统，系统认为用户的输入语句与现实生活中用户与电话接线员之间的对话方式一样，每句均包含一个被拨叫人的姓名。由于系统比较简单，期待焦点只有一个即为初始状态下的期待焦点，在此期待焦点下，通过分析现实生活中可能出现的实例，总结出一个识别自动机来表示期待焦点下的语境知识。下面实验中用来进行比较的有限状态网络指导下语音识别策略也采用此语境知识来指导语音识别。为了评测算法本身的识别性能，在下面的测试中，都没有使用语音确认技术进行后处理。

关键词表：*dEar-Attendant* 中包含的被拨叫人名，共 110 个；

测试数据库：测试数据均通过 8k 采样率的电话信道采集得到；四个测试集，每个 500 句均为自然发音，其特点分别如下所示：

1) 测试集 I：完全符合识别自动机，被呼叫人名在词表内；

2) 测试集 II：完全符合识别自动机，被呼叫人名不在词表内；

3) 测试集 III：不完全符合识别自动机，被呼叫人名在词表内；

4) 测试集 IV：不完全符合识别自动机，被呼叫人名也不在词表内；

它们之间的主要区别为：测试集 I 和测试集 II 是完全符合语境知识的测试数据，而测试集 III 和测试集 IV 的每个句子特意在识别自动机的基础上增加了一至两种口语现象，使其不完全符合识别自动机。测试集 I 和测试集 III 的被呼叫人名在关键词表内，而另两个测试集的被呼叫人名都在关键词表外，在识别结果中应该被拒识掉。

测试集 III 和测试集 IV 中出现的口语现象主要包括以下几种：咳嗽、吹气、OOV、重复、片断等等。都和实验的背景口语对话系统没有关系，是所有口语对话系统中均可能出现的口语现象。

实验步骤：分为以下两阶段

1) 不考虑拒识情况下的识别性能测试；

2) 考虑拒识情况下的识别性能测试。

评测指标：因为已知每句待识别样本中必然包含一个被呼叫人名，所以可将关键词识别算法的两个评测指标：识别正确率和误警率的定义做一点修改，使其物理意义更加明确，在本章下面的两个实验中所采用的识别正确率和误警率均是采用新的定义。新定义如下：

$$\text{识别正确率}=\frac{\text{包含关键词且被正确识别的句子数}}{\text{包含关键词的句子总数}}\times 100\% \qquad (4\text{-}1)$$

$$\text{误警率}=\frac{\text{不包含关键词但被识别出关键词的句子总数}}{\text{不包含关键词的句子总数}}\times 100\% \qquad (4\text{-}2)$$

4.5.2 实验一：不考虑拒识情况下的识别性能测试

采用有限状态网络指导下的语音识别算法和关键词识别算法作为对比算法。由于拒识的实现必然以降低一定识别正确率为代价，所以在不考虑拒识情况下的识别率可以说是算法识别能力的上限值，本实验的目的就在于测试三种算法的识别能力上限。在实验中强制限制每句测试样本中必然会包含一个关键词表中存在的被呼叫人名。实验结果如表 4-3 所示：

表 4-3 不考虑拒识情况下的识别性能比较

测试集	有限状态网络 /%	关键词识别 /%	语境知识指导下的关键词识别/%
测试集 I	96.6	88.0	**97.2**
测试集 III	27.2	81.6	**90.2**

从上表的测试结果可以看出，当测试数据符合识别自动机时，有限状态网络指导下的语音识别算法的识别率为 96.6%，要远远优于关键词识别算法；而当测试数据不符合识别自动机时，其识别正确率仅仅只有 27.2%，下降这么大充分说明了有限状态网络指导下的语音识别算法鲁棒性不高。但是无论在哪个测试集上，语境知识指导下的关键词识别算法均取得最佳的识别性能，识别正确率均超过了 90%，平均比关键词识别算法提高了 9 个百分点。这说明语境知识的引入不但没有影响关键词识别算法原有的高鲁棒性，且又大大提高了其识别性能，新算法具有高识别率和高鲁棒性的优点。

4.5.3 实验二：考虑拒识情况下的识别性能测试

在考虑拒识情况下的识别性能主要测试的是算法的综合能力。由于有限状态网络指导的语音识别策略鲁棒性较差且本身不具备拒识能力，只使用关键词识别策略和语境知识指导下的关键词识别策略进行对比。将四个测试集合并为一个测试集，在新的测试集中符合识别自动机的测试样本和不符合识别自动机的测试样本各占一半，被呼叫人名为集内关键词和集外关键词的也各占一半。采用 ROC 曲线来评测两种算法的整体性能，在坐

标系中 ROC 曲线越靠上的识别性能越优。实验结果如图 4-4 所示：

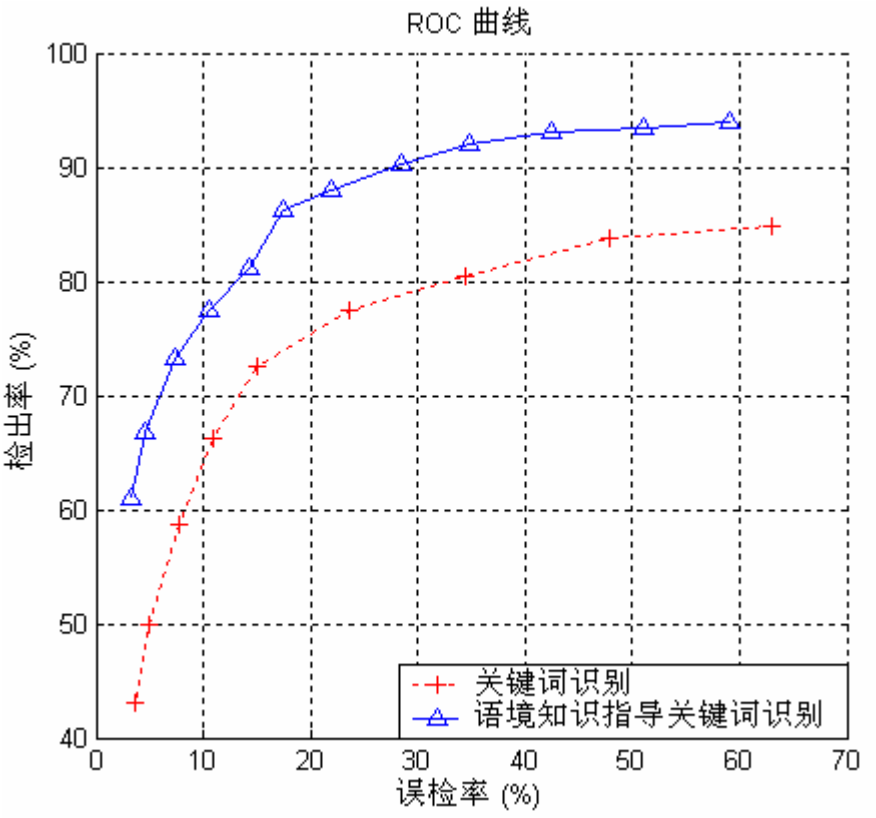


图 4-4 考虑拒识情况下的识别性能测试

从图 4-4 的识别结果看，两个算法的 ROC 曲线的走势基本一致。当比较误警率相同时的识别正确率时，有语境知识指导的关键词识别算法比没有语境知识指导的关键词识别算法平均高出了 12 个百分点，这再一次说明新算法的整体性能明显好于传统的关键词识别算法。主要的原因在于加入了语境知识指导后，新算法对整个句子每个部分的识别都更准确，最终使得关键词部分的识别定位也更准确。

4.5.4 分析与讨论

从图 4-4 中的 ROC 曲线显示出，语境知识指导下的关键词识别算法在误警率约 30%时，识别正确率超过 90%，如果再考虑上语音确认的后处理，

新算法可以在更低的误警率时达到 90%的识别正确率，这样的识别性能基本可以达到口语对话系统的产品化要求。另外，本章的实验虽然是在假设用户配合的前提下，但测试集中出现各种口语现象的测试样本的比例比较高，达到 50%。事实上，在实际的系统运行过程中，因为用户知道与他对话的是计算机而不是真正的人，用户在输入语句时会非常认真，口语现象出现的比例相当小。结合表 4-3 和图 4-4 可以看出，当输入测试样本不包含口语现象时，语境知识指导下的关键词识别算法可以达到 95%以上的识别正确率，可以满足口语对话系统对识别器的要求。

虽然用来测试的背景系统比较简单，但本实验目的是评测对话系统中语音识别的性能而不牵扯语义理解，所以实验结果是有说服力的。另外，在测试集中引入的口语现象却与领域无关，包括咳嗽、吹气、OOV、重复、片断等等，基本上是各种领域的口语对话系统中均会发生的口语现象。所以，实验结果和结论具有比较广泛的适用性和可推广性，在其它的对话系统中应该可以取得类似的实验结论。

语境知识指导下的关键词识别算法必然有其局限性，就是要求用户主动配合口语对话系统。因为在加入语境知识后，算法会对用户的输入有一个预测，如果用户故意不配合或语境知识总结的不完善，造成用户输入与语境知识偏差非常巨大时，语境知识反而会带来一定的副作用，当然这种类似现象在语音识别的其它领域也经常发生。因而要求语境知识的总结应该尽可能完善，要求用户也能基本配合口语对话系统，在这样的前提下，语境知识指导下的关键词识别算法可以取得高识别率和高鲁棒性的最佳识别性能。

4.6 小结

现有的口语对话系统的语音识别算法以关键词识别、有限状态网络指导下的语音识别和基于小领域统计语言模型的连续语音识别为主，关键词识别算法以其良好的鲁棒性而获得更多关注，但其性能仍不能满足口语对话系统的需要。本章试图从使用语境知识来指导关键词识别这一思路出发

来提高关键词识别算法的识别性能，提出语境知识指导下的关键词识别策略，具有如下特点：

1) 避开或弥补了特定小领域下语言模型不易获得的缺陷；

2) 显著提高了每个期待焦点下的关键词识别正确率，具有高识别率的特性；

3) 仍然保持关键词识别算法的高鲁棒性的优点，当输入语句中包含口语现象时仍具有相当高的识别正确率。

总之，语境知识指导下的关键词识别算法，具有高识别率和高鲁棒性的特性，基本满足了口语对话系统对识别器的要求，具有理论价值和实际意义。

在下章中将详细介绍作者在语音确认方面的研究工作，通过分析识别结果中上下文对待确认词的影响，将上下文的置信度特征也作为待确认词置信度特征的一部分，从而明显提高语音确认的性能。加入上下文语音确认方法后，语境知识指导下的关键词识别策略可以明显降低误警率，提高整体识别性能。

第五章　上下文相关的语音确认策略

语音确认也称为置信度测量(CM, Confidence Measure)，是语音识别中一个相对较新的领域，主要是对一段语音内容的处理过程进行确认。语音确认是建立在语音识别的基础上的，尤其与关键词识别有密切的关系，在语音识别结束后，针对待确认词候选计算出确认得分并决定是否通过确认。使用语音确认技术后可以有效地降低关键词识别的误警率[73][78][79]，即使在口语对话系统中使用语境知识指导关键词识别时已经大幅度提高了关键词的识别性能，应用语音确认技术后仍然可以有效降低误警率，提高总体性能。语音确认步骤在口语对话系统中是不可或缺的。

提高语音确认的性能，可以从两方面着手研究：提出新的具有区分度的确认特征和提出区分能力更强的确认模型。本章主要详细介绍作者在提出新确认特征方面的研究工作，在语境知识指导下的关键词识别策略基础上，提出了上下文相关的语音确认算法。算法中将待确认词候选的前后序词候选的确认特征也作为待确认词候选确认特征的一部分，体现出识别结果的上下文对当前待确认词的影响。实验表明：该算法的拒识性能明显好于传统的语音确认算法。

本章的内容安排如下：第一小节简要介绍语音确认的基本原理以及现今口语对话系统中采用的语音确认方法；第二小节分析上下文知识对识别结果的影响；第三小节提出上下文相关的语音确认策略；并详细描述了上下文相关置信度特征的定义以及实现方法；第五小节给出实验结果和分析；最后第六小节进行本章总结。

5.1 语音确认的研究现状

在语音识别领域，“确认”这个词首先用在说话人确认上。语音确认的研究是在 80 年代中期开始的，到了 90 年代初期，关键词的确认有了很大的发展，在电话查询系统中得到了实际的应用[78]。到了 90 年代中期以后，语音确认得到了长足的发展，确认的方法也呈现了多样化。有最大似

然值和基于概率假设检验的两种方法，而且把广义概率下降方法(Generalized Probabilistic Descent)应用到语音确认中，使系统性能有了很大的提高[73]。

语音确认的研究内容主要是以下几个方面：

1) 提取确认用的各种特征；

2) 联合多路特征给出确信度分数；

3) 评价单个特征有效性和整个系统的性能；

下面分几个小节逐一对其进行简要介绍。

5.1.1 确认特征

确认特征根据其使用的知识源的不同，大致可以分为三类：声学层面特征(Acoustic Features)、语言模型特征(Language Model Features)、词图特征(也称为词网络特征)(Word Graph Features or Word Lattice Features)。

声学层面特征：

理论上，给定的语音特征序列 X，后验概率 $P(W|X)$本身就是词 W 的确认分数。根据贝叶斯公式，$P(W|X)$可以由公式(5-1)计算得出。

$$P(W \mid X) = \frac{P(W)P(X \mid W)}{P(X)} = \frac{P(W)P(X \mid W)}{\sum_{W} P(W)P(X \mid W)} \tag{5-1}$$

但是在实际的语音识别系统中，一般都忽略 $P(X)$，因为对给定的 X 来说，$P(X)$是一个常量，因此必须采用一些特殊方法来估计 $P(X)$。目前有多种方法估计 $P(X)$的计算方法。

采用一个全词网络(all-word network)[79][85][86]来计算 $P(X)$。将搜索空间的语法限制去掉，任何一个词都可以连接其他所有的词，任意一个词序列都可以被识别出来。通过普通识别器在这个无语法限制的搜索空间上得到的最佳路径的似然值即可近似为 $P(X)$。这样，用相同的识别器在正常的搜索空间中得到的最佳路径的似然值与用上述方法得到的 $P(X)$相比就可以

近似得到 $P(W|X)$。如果这个近似值低于预先设定的阈值，那么我们拒识 W，否则接受。为了计算简单，通常并不直接使用全词网络，而可以使用一个全音素网络(all-phone network)[87][88][89][90]。通过普通的识别器在其上得到最佳路径，其得分即可近似为 $P(X)$。

这种确认特征还可以分为三个层次：帧级(frame level)，音素级(phone level)和词级(word level)，通常词级的特征有两种方法得到：1) 由词所在各帧的帧级特征直接求平均得到；2) 先对词中的每个音素求其所在各帧的帧级特征的平均值，得到每个音素的音素级特征，然后在对词内所有音素的音素级特征求平均最终得到词级特征。从文献中给出的实现结果可看出，第二种方法要优于第一种方法，因为避免了搜索过程中待确认词的某些音素所停留的时间过短这种不合理情况[91][92]。

还有多种计算后验概率的方法都可以作为确认特征，例如后验词假设概率(Posterior Word Hypothesis Probability) [93][94]、N-Best 列表后验概率(N-Best List Posterior Probability) [95][96]等等。

语言层面特征：

[97]中对识别结果词串计算语言模型分数，分数必须大于某一阈值才接受该识别结果。针对结果词串序列在语言模型训练数据库中是否出现或部分出现，给出一个得分，用此作为语言层面特征。

词网格特征或词图特征：

[95][96]提出在识别结果词图或词网格中，计算与待确认词在时间处于并列竞争位置的其它词候选的个数作为确认特征。显而易见，竞争词候选越多说明待确认词候选越不可靠。

其它的词网格特征还有邻接相关数目[98]等确认特征。

从文献中看，各个确认特征的区分侧重点是不相同的，根据共同的实验表明：多种特征共同运用时一般都使语音确认的性能有升无减。

5.1.2 确认模型

提取出多路确认特征以后，必须联合多路特征，给出最后的确认结果。语音确认要解决的实际上是一个分类问题，在第一步识别阶段给出待确定的识别结果以后，第二步中的确认过程对待确认的识别结果进行分类，一般来说只要分成“对”或“错”两类即可。大多数的分类技术和方法在语音确认中都是适用的，例如线性分类方法[99]、决策树[100]、神经元网络方法[99][101]以及支持向量机法[102]。本文的实验采用应用非常广泛的 Fisher 线性分类方法[103]作为确认模型，在此简单介绍其原理。

$$g(x) = w^T \cdot x + w_0 \tag{5-2}$$

在线性分类法中通过公式(5-2)把 d 维空间的样本投影到一条直线上，形成一维空间，即把维数压缩到一维，式中 $x=[x_1, x_2, \ldots, x_d]^T$ 是 d 维特征向量，即样本向量，$w=[w_1, w_2, \ldots, w_d]$为权向量。$w_0$ 是个常数，称为阈值权。然而即使样本在 d 维空间里形成若干紧凑的互相分得开的集群，若把它们投影到一条任意的直线上，也可能使几类样本混在一起而变得无法识别。但在一般情况下，总可以找到某个方向，使在这个方向的直线上样本的投影能分开得最好。问题是如何根据实际情况找到这条最好得到并最易于分类的投影线。

Fisher 线性分类法认为最优的线性投影应该尽可能达到两个标准：1) 投影后在一维空间里各类样本尽可能分得开一些，两类均值之差越大越好；2) 各类样本内部尽量密集，即类内离散度越小越好。因此 Fisher 准则函数定义为上述两项的商。只要找到使 Fisher 准则函数取极大值的投影方向即找到了最优的投影方向，即将 d 维分类问题转化为一维分类问题。具体实现步骤请参考文献[103]。

5.1.3 评测指标

首先给出一些定义：

✧ N ：待确认词串中所含的词数（并不一定是用户输入的词数）；

✧ N_c ：待确认词串中正确识别的词数；

✧ N_e ：待确认词串中错误识别的词数；

✧ $N_{c\text{-}E}$：待确认词串中识别正确但标记错误的词数（错误拒绝）；

✧ $N_{e\text{-}C}$：待确认词串中识别错误但标记正确的词数（错误接受）；

针对语音确认的评测，常见的有以下几种评测指标。

确认错误率(Confidence Error Rate) [104]：

$$CER = \frac{\text{标记错误的词数}}{\text{待确认的总词数}} = \frac{N_{c-E} + N_{e-C}}{N} \tag{5-3}$$

互熵(Cross Entropy)[91]：

$$CREP = \frac{1}{N}\sum_{w}\left[\delta_w \log(c_w) + (1-\delta_w)\log(1-c_w)\right] \tag{5-4}$$

上式中 c_w 是假设词 w 的概率似然值，当词 w 确认为正确时，δ_w 取 1，否则 δ_w 取 0，N 是待确认的所有此数。如果经过确认阶段以后，$CREP$ 变大了，则说明确认方法是有效的。

错误拒绝率(FRR, False Rejection Rate)：错误拒绝比率，如下所示：

$$FRR = \frac{N_{c-E}}{N_c} \tag{5-5}$$

错误接受率(FAR, False Acceptance Rate)：错误接受比率，如下所示：

$$FAR = \frac{N_{e-C}}{N_e} \tag{5-6}$$

ROC(Receiver Operating Characteristic)曲线：错误拒绝率和错误接受率两者之间是有关联的，错误拒绝率越高错误接受率越低，反之亦然。用 ROC 曲线可以很好地描述两者之间的关系，并且可以找到一个最佳操作点。在坐标系中，ROC 曲线越靠近坐标轴表示语音确认的性能越好。

等错误率(EER, Equal Error Rate)：在坐标系中 ROC 曲线与从左下角

到右上角的对角线的交点，可以认为是错误拒绝率和错误接受率的最佳折中方案。等错误率越小代表语音确认的性能越好。

在多数语音确认的应用中，对 FRR 要求比较严格，不能过大，否则正确候选的损失较大，会带来很坏的负面影响。所以也将 FRR 等于 5%或 2.5%时 FAR 的值作为语音确认的评测标准。

5.2 上下文对识别结果的影响

近年来语音确认的研究工作集中在两个方面：1) 提出新的有区分能力的确认特征；2) 提出区分能力更强的确认模型。尤其是前一个方面发展更快，对语音确认性能改善的贡献更大，本章的研究工作也集中在如何提出新的确认特征上。

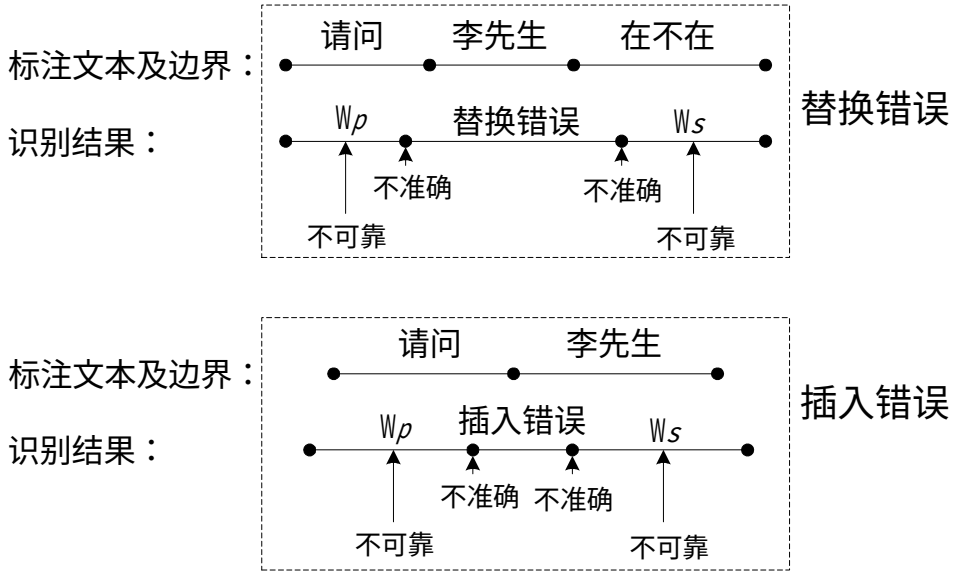


图 5-1 语音识别中识别结果互相干扰现象示例

在语音识别中，往往会出现识别结果互相干扰的现象：如图 5-1 所示，在一句输入语音样本的识别中，如果某一个词的识别发生替代和插入错误，很可能会发生识别结果的前后时间边界与正确的边界不一致的现象；中间词识别结果的边界定位不准确时，必然会导致识别结果中前后接序词

的边界定位也会不准确；从而会影响前后接序词的识别准确度。根据以上的推理过程，可以总结出如果某一个词的识别发生错误，将会在一定程度上影响前后接序词的识别准确度。

将上面的推理过程反过来分析，可以得出，如果某识别结果的前后接序词的识别结果发生错误，则很可能会导致中间词识别结果边界定位不准确，从而影响中间词识别结果的识别准确度。

5.2.1 问题提出

这种识别结果互相干扰的现象在语音识别的多个领域中均有发生。在连接词识别、关键词识别和大词表连续语音识别等只要是属于连续语音识别的算法中都会发生这种现象，尤其在大词表连续语音识别中这种现象还表现为错误识别结果对后面识别结果语言模型概率得分的干扰。在口语对话系统中所采用的识别策略，无论是关键词识别、有限状态网络指导的连续识别、应用小领域语言模型的连续识别还是语境知识指导下的关键词识别，都属于连续语音识别的范畴，所以都会发生识别结果互相干扰的现象。

虽然这种互相干扰的负面影响到底有多大并不好估计，但我们可以确信这种互相干扰必然存在。这就提出了一个问题：如果在识别过程中可以知道某识别结果的前后接序词是错误的，则中间识别结果的正确性就很值得怀疑。如果可以提出新的确认特征专门描述前后接序词的特性，以改善语音确认的性能，则无论口语对话系统所采用的识别策略为哪一种，都可以有效地提高整个识别模型的识别性能。

在识别过程中不可能知道识别结果最终是否正确，最多只能在语音确认结束后知道识别结果的确认得分高低。根据上面的分析，试图考虑待确认词候选的前后接序识别候选的确认度，希望其能够对当前待确认词候选的确认带来帮助。直观上可以感觉出：如果前后接序词识别的不稳定、确认度不高，那中间词的确认度也值得怀疑。

5.3 上下文相关的语音确认

为了尽可能地提高语音确认的性能，本章提出上下文相关的语音确认策略，与传统语音确认策略相比，其创新点在于从识别结果互相干扰现象入手，考虑待确认词前后接序词的确认度对待确认词确认度的影响，在语音确认中使用上下文相关确认特征方式。

5.3.1 上下文相关确认特征

如何最好的将上下文信息引入语音确认算法中是新算法的关键之处。在识别结束时，唯一能够利用的可以表述前后接序词确认度的信息只有前后接序词的确认得分。在此使用统计学方法，用一个 Fisher 线性分类器将前后接序词和待确认词的确认得分组合成为一个置信度得分，称之为待确认词的上下文相关确认得分。公式(5-2)可以展开为式(5-7)，其中上标为 *pre* 的代表前序词的参数；上标为 *cur* 的代表待确认词的参数；上标为 *sub* 的代表后序词的参数。

$$g(x) = w^{pre} \cdot x^{pre} + w^{cur} \cdot x^{cur} + w^{sub} \cdot x^{sub} + w_0 \tag{5-7}$$

如果前后接序词和待确认词的确认得分也通过一个 Fisher 线性分类器计算得到，则式(5-7)中的三个确认得分就由另一个线性公式(5-8)计算出。其中 $f^{[pre,cur,sub]}$ 分别代表前序词、待确认词、后序词的确认特征向量。

$$x^{[pre,cur,sub]} = w^T \cdot f^{[pre,cur,sub]} + w_0' \tag{5-8}$$

将公式(5-8)代入(5-7)中，可以得到公式(5-9)：

$$\begin{aligned} g(x) &= w^{pre} \cdot \left(w^T \cdot f^{pre} + w_0'\right) + w^{cur} \cdot \left(w^T \cdot f^{cur} + w_0'\right) \\ &\quad + w^{sub} \cdot \left(w^T \cdot f^{sub} + w_0'\right) + w_0 \\ &= w^T \cdot \left(w^{pre} \cdot f^{pre} + w^{cur} \cdot f^{cur} + w^{sub} \cdot f^{sub}\right) \\ &\quad + \left[w_0' \cdot \left(w^{pre} + w^{cur} + w^{sub}\right) + w_0\right] \\ &= \tilde{w}^T \cdot \tilde{f} + \tilde{w}_0 \end{aligned} \tag{5-9}$$

其中，

$$\tilde{w}=\begin{bmatrix} w^{pre}\cdot w \\ w^{cur}\cdot w \\ w^{sub}\cdot w \end{bmatrix},\quad \tilde{f}=\begin{bmatrix} f^{pre} \\ f^{cur} \\ f^{sub} \end{bmatrix},\quad \tilde{w}_0=w_0'\cdot\left(w^{pre}+w^{cur}+w^{sub}\right)+w_0 \qquad (5\text{-}10)$$

从公式(5-9)中可以看出，对确认得分再使用 Fisher 线性分类器的效果等同于将前后序词和待确认词的确认特征向量合并为一个大向量，对这个新向量应用 Fisher 线性分类器。而且由 Fisher 线性分类器直接训练参数，比分两步训练更灵活。为此提出上下文相关确认策略，直接将前后接序词的传统确认特征也作为待确认词确认特征集的一部分，形成一个规模是原有确认特征三倍的大确认特征向量。然后使用统计方法从训练数据中训练出确认模型，完全由确认模型来解决上下文信息的引入。为了表述方便，将传统的仅仅用来描述当前待确认词候选的确认度信息的确认特征称为局部确认特征，而考虑了上下文信息的称为上下文相关确认特征。上下文相关确认特征的组成如图 5-2 所示：

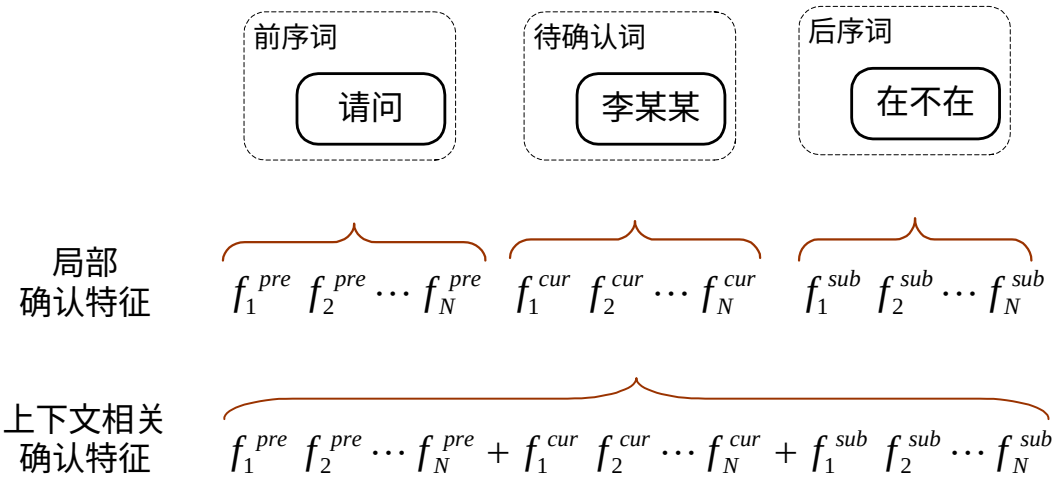


图 5-2 上下文相关确认特征的构成

无论原来采用何种置信度特征，上下文方式都可以起作用。因为，上下文相关置信度特征的本质是引入上下文对当前识别结果的影响，与普通的置信度特征希望从待确认词本身的各种特点出发完全不同，也就是说所使用的知识源完全不同。因此，无论哪种置信度特征都可以用来构建上下文相关的置信度特征。

上下文相关确认特征是通过 Fisher 线性分类器推导出来的，但并不限于 Fisher 线性分类器。上下文相关语音确认策略的中心思想是上下文相关确认特征，即将前后接序词的局部确认特征也作为待确认词确认特征集的一部分，形成上下文相关确认特征。无论采用什么确认模型，只要其有能力将多维特征组合在一起就可以起作用。

5.3.2 处理补白模型和静音模型

当口语对话系统中采用语境知识指导下的关键词识别策略时，待确认关键词候选的前后序词共有四种情况：其它关键词、识别自动机中除关键词以外的词、补白模型和静音模型。在前两种情况下，前后序词的局部确认特征都有明确的物理意义，但当前后序词是补白模型或静音模型时，局部确认特征的物理意义就有了一定的问题。因为，补白模型和静音模型识别结果的持续时间可能非常短也可能非常长，而补白模型和静音模型的状态数又远远小于关键词，在这种情况下词图确认特征和声学确认特征都会发生一定的偏差。为了弥补这种偏差，仍然采用统计方法对其处理，在上下文相关确认特征集中再增加四维新特征，如下：

- ✧ 前序词是否是补白模型；
- ✧ 前序词是否是静音模型；
- ✧ 后序词是否是补白模型；
- ✧ 后序词是否是静音模型；

这四维特征专门表示前后序词是否属于补白模型或静音模型，和其它的确认特征放在一起进行训练，由确认模型训练出它们之间的相互联系。这种方式是否能够奏效，将在后文中进行验证。

5.4 实验结果与分析

上下文相关语音确认策略是否合理，上下文相关确认特征的构建方法是否有效，即是否能够将上下文知识有效地引入语音确认算法中来，都需要详尽的实验进行验证。本章的实验是在第四章的实验基础上进行的。

5.4.1 实验设计

实验仍以口语对话系统 *dEar-Attendant* 为实验背景，在语境知识指导下的关键词识别策略的基础上测试上下文相关语音确认策略的整体性能。整体实验可以说是在节 4.5.2 的基础上进行的，在不失代表性的前提下，在图 4-4 中语境知识指导下的关键词识别算法的 ROC 曲线上，选取识别正确率为 92%、误警率为 35%时为操作点，以其作为本章两个实验的基础。

关键词表：*dEar-Attendant* 中包含的被拨叫人名，共 110 个；

测试数据库：将节 4.5.1 中的四个测试集合并，共 2000 句，一分为二，1000 句作为确认训练集，另 1000 句作为确认测试集。确认训练集和确认测试集完全独立，保证实验结果的正确性和可扩展性。

确认特征：选用以下几种确认特征作为局部确认特征，这些确认特征也将被用来构建上下文相关确认特征集。这些确认特征中既包括声学层面特征，也包括词图特征，因为语境知识指导下的关键词算法中没有使用语言模型，所以语言模型特征也不在局部确认特征中。局部确认特征如下：

- ✧ 词级后验概率得分；
- ✧ 词中最小的音素级后验概率得分；
- ✧ 词的持续时间；
- ✧ 词结束时搜索空间中的路径条数；
- ✧ 结果词图中，与词候选处于并列位置的其它词候选个数；

确认模型：Fisher 线性分类器。

实验步骤：实验步骤分为以下几步：

(1) 实验一：测试上下文相关语音确认策略的整体性能；

(2) 实验二：测试语音确认与关键词识别结合后的整体性能；

5.4.2 实验一：测试上下文相关语音确认策略的整体性能

首先测试上下文相关语音确认策略的整体性能，测试在语境知识指导下的关键词识别的识别结果上进行。将确认训练集的识别结果与标注文本进行比较，人工将对应的确认特征分为正确识别确认特征集与错误识别确认特征集，用统计方法训练出 Fisher 线性分类器。再根据训练得到的 Fisher 线性分类器参数，对确认测试集的识别结果计算确认得分，并与标注文本相对照得到实验结果。实验结果如下所示：

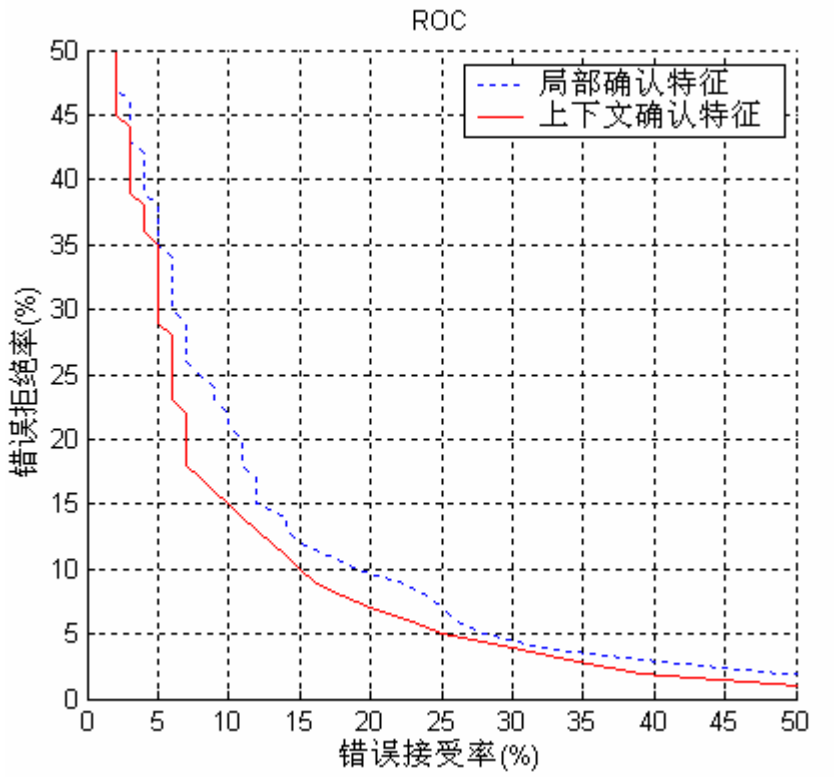


图 5-3 上下文相关语音确认的整体性能

首先选用 ROC 曲线作为评测指标，在坐标中 ROC 越接近坐标轴越好。图 5-3 中是局部确认特征与上下文相关确认特征的 ROC 曲线，从图中可看出两种确认语音确认性能的总体走向。明显可以看出上下文相关确认特征的 ROC 曲线比局部确认特征的要更接近坐标轴，说明在语音确认能力上，

上下文相关确认特征要明显好于传统的局部确认特征。为了进一步的比较，再使用其它评测指标进行比较，如表 5-1 所示：

表 5-1 中采用等错误率、错误拒绝率为 2.5%和 5%时的错误接受率三个评测指标来比较局部确认特征和上下文相关确认特征区分能力的差别。从三个指标中都可以看出，上下文相关确认特征的区分能力强于局部确认特征，每个指标都下降了 15%左右，充分说明了上下文相关语音确认策略的合理性，和上下文相关确认特征的有效性。

表 5-1 上下文相关语音确认的性能评测

算法	等错误率/%	错误接受率/%(错误拒绝率＝5%)	错误接受率/%(错误拒绝率＝2.5%)
局部确认特征	14.0	28.0	43.0
上下文相关确认特征	12.0	24.0	35.0
↓	14.3	14.3	18.6

5.4.3 实验二：测试语音确认与关键词识别结合后的整体性能

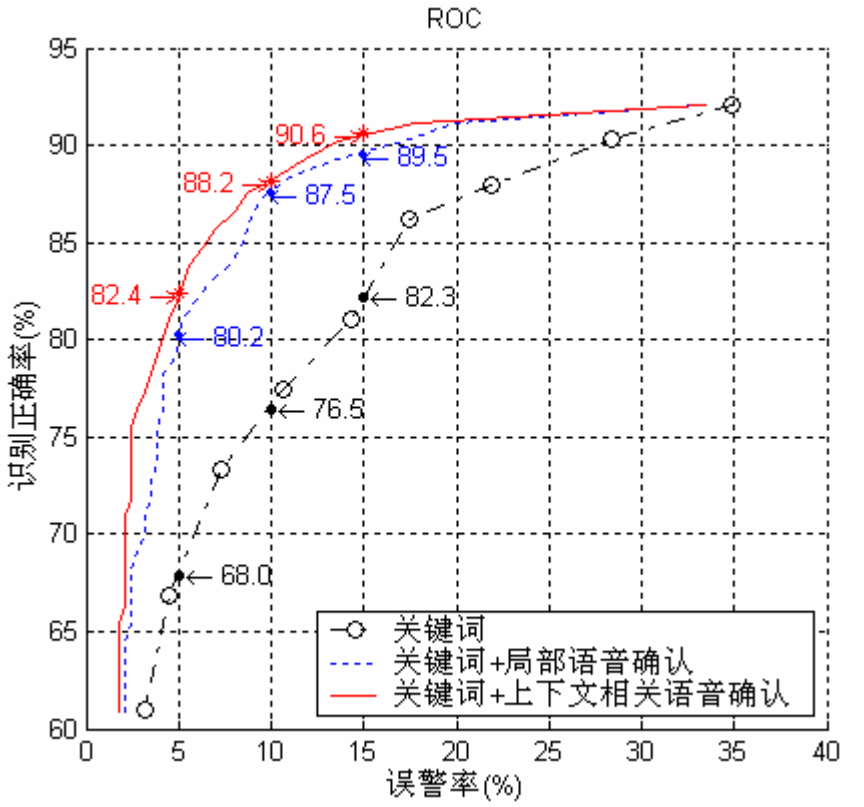


图 5-4 关键词识别结合语音确认的整体性能

接下来测试关键词识别算法和语音确认技术结合后的整体性能。分别将局部语音确认和上下文相关语音确认应用在关键词识别结果上，画出两条新的 ROC 曲线，与原关键词识别算法的 ROC 一起比较，如图 5-4 所示。

可以看出，加入语音确认后关键词识别算法的 ROC 曲线在坐标轴上要远远高于原关键词识别算法的 ROC 曲线，证明语音确认与关键词识别结合后的整体性能要优于关键词识别算法本身。使用局部语音确认策略和使用上下文相关语音确认策略的两条 ROC 曲线相比，后者明显要好于前者。参照图中专门标出的误警率为 5%、10%、15%时，三条 ROC 曲线上对应的识别正确率的值，可以看出加入上下文相关语音确认的关键词识别

算法对应的三个值是最高的。实验充分表明了上下文相关语音确认策略与关键词识别相结合可以取得最优的整体识别性能，又一次证明了上下文相关语音确认策略的合理性，和上下文相关确认特征的有效性。

5.4.4 分析与讨论

本章的创新点是引入待确认词前后序词的确认信息以提高对中间词语音确认的准确度，实验均是在语境知识指导下的关键词识别算法的基础上进行的，实验结果证明上下文相关语音确认策略是有效的。但存在一个问题就是是否在所有的语音识别算法中，实验结果前后序词的确认信息都能够得到并且都有用？显然，孤立词识别算法中识别结果就只有一个，肯定不能使用上下文相关语音确认策略，不过在关键词识别和连续语音识别中都可以使用上下文相关语音确认策略。

从实验结果图 5-3 中显示，本章提出的上下文相关语音确认策略与传统语音确认策略相比，虽然有明显性能提高但并不是非常巨大。这主要是因为并不是所有测试样本都存在识别结果互相干扰的现象，而且即使存在识别结果互相干扰现象也不一定就导致确认出错，也就是说因为互相干扰原因导致确认出错的测试样本占的比例并不高，所以即使上下文相关语音确认策略能够使用上下文信息把这些样本都纠正过来，表现在整个测试集上确认性能的提高也会很有限。

实验证明上下文相关语音确认策略是有效的，但并不能证明通过使用上下文相关确认特征来引入上下文信息的方式是最好的。使用上下文相关确认特征的方式是一种统计学方法，可以说是没有办法的办法，也许还有更好的方式来引入上下文信息，这有待于今后的研究工作来发现。

5.5 小结

语音确认策略在口语对话系统的语音识别中很重要，能够明显降低关键词识别的误警率，一直以来倍受关注。本章试图从引入上下文知识来指导语音确认这一思路出发来提高语音确认算法的性能，提出上下文相关语

音确认策略。将待确认词前后序词的确认特征也作为待确认词确认特征的一部分，构成上下文相关确认特征，将其引入上下文信息，具有如下特点：

1) 上下文相关确认特征在原有确认特征的基础上引入上下文信息；

2) 上下文相关确认特征构建与原有的确认特征相互独立，不冲突；

3) 上下文相关确认特征不限制使用何种确认模型；

4) 上下文相关语音确认策略的拒识性能明显好于传统的语音确认策略；

总之，上下文相关语音确认策略具有良好的拒识性能，可以提高口语对话系统的整体性能，具有理论价值和实际意义。

第六章　总结与展望

6.1 论文工作总结

本文针对口语对话系统中语音识别任务中的若干难点，以关键词识别策略和语音确认策略为研究对象，在提升关键词识别算法的识别性能、对话语境知识对关键词识别的指导及提出新的语音确认特征诸方面进行了初步探索研究，提出了若干新方法、新策略，并通过实验证明了其有效性，同时也为口语对话系统语音识别领域的深入研究奠定了一定基础。

概括来说，本文的工作重点与贡献主要体现在如下几个方面：

1) 提出关键词动态确认

通过详细分析关键词识别算法的具体实验结果，发现关键词的漏检错误大多数都与关键词的误警错误有关，针对这种现象，提出关键词动态确认的概念。在搜索过程中引入虚拟 OOV 模型来对产生的关键词候选进行确认，及早地将不正确的关键词候选剪除，从而避免其对正确关键词候选的干扰影响。实验表明：采用关键词动态确认技术后关键词识别算法的识别性能有明显提高，在误警率相同的条件下，误识率下降了约 10%。

2) 提出对话语境知识指导下的关键词识别策略

目前的各种语音识别方法，或由于能力太弱而不适应对话系统的复杂性，或由于要求大批数据的支持而可行性较差，在实际应用任务中的性能均不太令人满意。为改善对话系统中语音识别的性能，提出了对话语境知识指导下的关键词识别策略，利用对话管理器给出期待焦点信息，确定对应焦点下的活动词表、活动规则集，生成相应的识别自动机，并用其来指导关键词识别。该识别框架具有良好的可扩展性，将诸如规则、统计和经验等知识融为一体。实验表明：语境知识指导下的关键词识别算法具有很高的识别性能和鲁棒性，基本能够满足口语对话系统的需要。

3) 提出上下文相关的语音确认策略

语音确认策略在口语对话系统的语音识别中很重要，能够明显降低关键词别的误警率。本文通过分析识别结果中互相干扰现象的发生，发现待确认结果前后序词的确认度会对待确认词本身确认度有一定的指导作用，据此引入上下文知识来指导语音确认，提出上下文相关语音确认策略。利用待确认词与其前后序词的确认特征组成待确认词的上下文相关确认特征。实验表明：该算法的拒识性能明显好于传统的语音确认算法。

虽然本文的研究是基于口语对话系统中语音识别任务，但其中的一些关键思路，例如利用语境知识指导语音识别、上下文相关的语音确认策略等等，能够推广到其它背景下的语音识别任务中。

6.2 下一步研究的展望

本文虽然在对话系统的语音识别方面进行了一些初步研究，提出了一些新方法和新思路，取得了一定的成果，但同时也发现了一些不足之处。下面将指出这些不足点，以及计划进一步深入开展研究的若干方向。

1) 知识统计方法在关键词识别中的应用

本文主要从基于规则的方法进行语境知识的表示，利用其对关键词识别进行指导。但实际上，规则方法和统计方法各具优势。规则方法在领域小或简单时比较方便，但它的弱点在于需要比较多的人工干预，这表现在：1). 需要设计人员对领域特点的了解相当深入，以便总结出简单、高效而又具有良好可读性的文法；2). 语义解释函数（或过程）需要针对语义规则进行逐条设计，语义解释函数必须随语义规则的改变而同步更新。虽然统计方法也有其缺点，但也有其明显优点：1). 弥补人工经验知识的不足。经验性的知识可以认为总是不够的，而统计方法描述平均规律的特性往往能覆盖经验性知识所忽略之处。2). 模型或规律的（半）自动获得性。给定数据，统计方法可以自动获得模型，数据越多模型也越精确（当然有上限）。因此可以考虑统计方法在语境知识的表示方面的使用，将其与规则方法进行结合。

统计和规则的结合可以从两个方面来进行：1). 统计和规则统一的语

言模型，比如说可以落实到 N-Gram 和 CFG 的统一。2). 统计方法进入语义分析，在分析得到的中层概念的基础上，基于统计规律“猜测”用户的意图。

2) 提出更复杂的语音确认模型

第五章的研究中，均采用 Fisher 线性分类器作为语音确认模型，在现有的几种语音确认模型中，Fisher 线性分类器是最简单的确认模型。上下文相关确认特征因为包含三个词的确认特征，通常维数会比较多，采用相对简单的线性分类器是否能够表示确认特征之间的复杂关系，充分发挥上下文相关确认特征的能力值得怀疑。所以在今后的研究中，应该引入更为有效的语音确认模型。而且从提高语音确认算法性能出发，提出更高级语音确认模型也是一条重要的研究途径。

3) 语音识别中的去噪处理和自然语音发音建模

语音识别经过几十年的发展，已经可以在实验室环境下达到很高的识别率，但在口语对话系统中识别性能却大打折扣，这主要是因为口语对话中的噪音、信道差异以及自然发音方式都大大影响了语音识别的识别性能。虽然本文提出的语境知识指导下的关键词识别策略基本达到了口语对话系统的需要，但如果能从根本上解决噪音、信道差异以及自然发音的影响，则语音识别性能将更上一级台阶。

参考文献

[1] Allen J F, Schubert L K, Ferguson G M et al. The TRAINS project: A case study in building a conversational planning agent. Technical Report 532, Dept. of Computer Science, University of Rochester, Rochester, NY 14627-0226, 1994

[2] Goldschen A, Loehr D. The role of the DARPA communicator architecture as a human computer interface for distributed simulations. Technical report, MITRE Corporation, 1999

[3] Os D E, Boves L, Lamel L, et al. Overview of the ARISE project. In: Proceedings of the 6th European Conference on Speech Communication and Technology (EuroSpeech). Budapest, Hungary, 1999. 1527-1530

[4] Failenschmid K, Thornton J H S. End-user driven dialogue system design: The REWARD experience. In: Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP). Sydney, Australia, 1998. 37-40

[5] Wahlster W. VERBMOBIL: Translation of face-to-face dialogs. In: Proceedings of the 3rd European Conference on Speech Communication and Technology (EuroSpeech). Berlin, Germany, 1993. 29-38

[6] Zue V. Conversational interfaces: Advances and challenges. In: Proceedings of the 5th European Conference on Speech Communication and Technology (EuroSpeech). Rhodes, Greece, 1997. KN: 9-18

[7] Rudnicky A. Creating natural dialogs in the Carnegie Mellon communicator system. In: Proceedings of the 6th European Conference on Speech Communication and Technology (EuroSpeech). Budapest, Hungary, 1999. 1531-1534

[8] Lee C H. Spoken dialogue processing towards telecommunication applications. In: Proceedings of International Symposium on Spoken Dialogue. Sydney, Australia, 1998. 43-50

[9] Potamianos A, Kuo H K, Lee C H, et al. Design principles and tools for multimodal dialog systems. ESCA Workshop: Interactive Dialogue in Multi-Modal Systems, 1999

[10] Sutton S, Cole R, et al. Universal speech tools: the CSLU toolkit. In: Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP). Sydney, Australia, 1998. 3221-3224

[11] Lamel L, Rosset S, Gauvin J L, et al. The LIMSI ARISE system. In: Proceedings of IEEE 4th Workshop on Interactive Voice Technology for Telecommunications Applications, Torino, Italy, 1998. 209-214

[12] Lamel L, Rosset S, Gauvain J L, et al. The LIMSI ARISE system for train travel information. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). Phoenix, USA, 1999. 501-504

[13] Gallwitz F, Aretoulaki M, Boros M, et al. The Erlangen spoken dialogue system EVAR: A state-of-the-art information retrieval system. In: Proceedings of International Symposium on Spoken Dialogue. Sydney, Australia, 1998. 19-26

[14] Takezawa T, Yamamoto S. Dialogue processing in a speech-to-speech translation system between Japanese and English. In Proceedings of International Symposium on Spoken Dialogue. Sydney, Australia, 1998. 35-42

[15] Aust H, Schroer O. Application development with the Philips dialog system. In: Proceedings of International Symposium on Spoken Dialogue. Sydney, Australia, 1998. 27-34

[16] Huang Y F, Zheng F, Xu M X, et al. Language understanding component in Chinese dialogue system. In: Proceedings the 6th International Conference on Spoken Language Processing (ICSLP). Beijing, China, 2000. 1053-1056

[17] Huang C, Xu P, Zhang X et al. LODESTAR: A mandarin spoken dialogue system for travel information retrieval. In: Proceedings of the 6th European Conference on Speech Communication and Technology (EuroSpeech). Budapest, Hungary, 1999, 1159-1162

[18] Meng H M, Tsui W C. Comprehension across application domains and languages. In: Proceedings of the 2nd International Symposium of Chinese Spoken Language Processing (ISCSLP). Beijing, China, 2000. 69-72

[19] Lin B S, Lee L S. Computer-aided design/analysis for Chinese spoken dialogue systems. In: Proceedings of the 2nd International Symposium of Chinese Spoken Language Processing (ISCSLP). Beijing, China, 2000. 57-60

[20] Hugunin J, Zue V. On the Design of Effective Speech-Based Interfaces for Desktop Applications. In: Proceedings of the 5th European Conference on Speech Communication and Technology (EuroSpeech). Rhodes, Greece, 1997. 1335-1338

[21] Issar S. A Speech Interface for Forms on WWW. In: Proceedings of the 5th European Conference on Speech Communication and Technology (EuroSpeech). Rhodes, Greece, 1997. 1343-1346

[22] Asoh H, Matsui T, Fry J, et al. A Spoken Dialog System for a Mobile Office Robot. In: Proceedings of the 6th European Conference on Speech Communication and Technology (EuroSpeech). Budapest, Hungary, 1999. 1139-1142

[23] Sasajima M, Yano T, Kono Y. EUROPA: A Generic Framework for Developing Spoken Dialogue Systems. In: Proceedings of the 6th European Conference on Speech Communication and Technology (EuroSpeech). Budapest, Hungary, 1999. 1163-1166

[24] Kono Y, Yano T, Sasajima M. BTH: An Efficient Parsing Algorithm for Word-Spotting. In: Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP), Sydney, Australia, 1998. 2067-2070

[25] 杨开城. 一种基于句法语义特征的汉语句法分析器. 中文信息处理学报, 2003, 14(3)

[26] Byron D K. Improving Discourse Management in TRIPS-98. In: Proceedings of the 6th European Conference on Speech Communication and Technology (EuroSpeech), Budapest, Hungary, 1999. 1379-1382

[27] Boros M, Heisterkamp P. Linguistic Phrase Spotting in a Simple Application Spoken Dialogue System. In: Proceedings of the 6th European Conference on Speech Communication and Technology (EuroSpeech). Budapest, Hungary, 1999. 1985-1986

[28] Wang Y Y. A Robust Parser for Spoken Language Understanding. In: Proceedings of the 6th European Conference on Speech Communication and Technology (EuroSpeech). Budapest, Hungary, 1999. 2055-2058

[29] Denecke M, Weibel A. Dialogue Strategies Guiding Users to Their Communicative Goals. In: Proceedings of the 5th European Conference on Speech Communication and Technology (EuroSpeech), Rhodes, Greece, 1997.1339-1342

[30] Zanten G V. User Modelling in Adaptive Dialogue Management. In: Proceedings of the 6th European Conference on Speech Communication and Technology (EuroSpeech), Budapest, Hungary, 1999. 1183-1186

[31] Papineni K A, Roukos S, Ward R T. Free-flow Dialog Management Using Forms. In: Proceedings of the 6th European Conference on Speech Communication and Technology (EuroSpeech). Budapest, Hungary, 1999. 1411-1414

[32] Lin B S, Wang H M, Lee L S. Consistent Dialogue Across Concurrent Topics Based on an Expert System Model. In: Proceedings of the 6th European Conference on Speech Communication and Technology (EuroSpeech). Budapest, Hungary, 1999. 1427-1430

[33] Pargellis A, Kuo J, Lee C H. Automatic Dialogue Generator Creates User Defined Applications. In: Proceedings of the 6th European Conference on Speech Communication and Technology (EuroSpeech). Budapest, Hungary, 1999. 1175-1178

[34] Lussier F E, Morgan N. Effect of speaking rate and word frequency on pronunciations in convertional speech. Speech Communication, 1999, 29: 137-158

[35] Decker A M, Lamel L. Pronunciation variants across system configuration, language and speaking style. Speech Communication, 1999, 29: 83-98

[36] Greenberg S. Speaking in shorthand – a syllable-centric perspective for understanding pronunciation variation. Speech Communication. 1999, 29: 159-176

[37] 林焘, 王理嘉. 语音学教程. 北京: 北京大学出版社. 1992

[38] Davis, S B, Mermelstein P. Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. on Acoustic, Speech and Signal Processing, 1980, 28(4): 357-366

[39] Viikki O, Laurila K. Noise Robust HMM-based Speech Recognition Using Segmental Cepstral Feature Vector Normalization. In: ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels, 1997. 107-110

[40] Viikki O, Laurila K. Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition. Speech Communication 1998. 25: 133-147

[41] Zheng F, Song Z J, Fung P, et al. Mandarin Pronunciation Modeling Based on CASS Corpus. Sino-French Symposium on Speech and Language Processing, Beijing, China, 2000. 47-53

[42] Gustafson J, Lindberg N, Lundeberg M. The August Spoken Dialogue System. In: Proceedings of the 6th European Conference on Speech Communication and Technology (EuroSpeech), Budapest, Hungary, 1999. 1151-1154

[43] Lopez-Cozar R, Rubio A J, Garcia P, et al. A New Word-Confidence Threshold Technique to Enhance the Performance of Spoken Dialogue Systems. In: Proceedings of the 6th European Conference on Speech Communication and Technology (EuroSpeech). Budapest, Hungary, 1999. 1395-1398

[44] Seide F, Kellner A. Towards an Automated Directory Information System. In: Proceedings of the 5th European Conference on Speech Communication and Technology (EuroSpeech). Rhodes, Greece, 1997. 1327-1330

[45] Hakkani-Tur D, Tur G, Stolcke A, et al. Combining Words and Prosody for Information Extraction from Speech. In: Proceedings of the 6th European Conference on Speech Communication and Technology (EuroSpeech), Budapest, Hungary, 1999. 1991-1994

[46] Yamashita Y. Keyword Spotting Using F0 Contour Matching. In: Proceedings of the 5th European Conference on Speech Communication and Technology (EuroSpeech). Rhodes, Greece, 1997. 271-274

[47] Spilker J, Weber H, Gorz G. Detection and Correction of Speech Repairs in Word Lattices. In: Proceedings of the 6th European Conference on Speech Communication and Technology (EuroSpeech). Budapest, Hungary, 1999. 2031-2034

[48] Rabiner L R, Schafer R W. Digital Processing of Speech Signals. USA: Prentice-Hall, Inc., 1978.

[49] Hermansky H. Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Am., 1990, 87(4):1738-1752

[50] Zwicker E. Subdivision of the audible frequency range into critical bands. J. Acoust. Soc. Am., 1961, 33

[51] Wilpon J G, Miller L G, Modi P. Improvements and applications for key word recognition using hidden Markov modeling techniques. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, Canada, 1991. 309-312

[52] Furui S. Speaker-independent isolated word recognition using dynamic features of speech spectrum. IEEE Trans. on Acoustic, Speech and Signal Processing, 1986, 34(1): 52-59

[53] Soong F K, Rosenberg A E. On the use of instantaneous and transitional spectral information in speaker recognition. IEEE Trans. on Acoustic, Speech and Signal Processing, 1986, 36(6): 871-879

[54] Rabiner L R, Wilpon J G, Soong F K. High performance connected digit recognition using hidden Markov models. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), New York, USA, 1988. 119-122

[55] 郑方，牟晓隆，徐明星，等. 汉语语音听写机技术的研究与实现. 软件学报，1999, 10(4): 436-444.

[56] Lee C H, Rabiner L, Pieraccini R, et al. Acoustic modeling for large vocabulary speech recognition. Computer Speech and Language, 1990, 4(2): 127-165

[57] Young S J, Odell J J, Woodland P C. Tree-based state tying for high accuracy acoustic modeling. In: Proceedings of ARPA Workshop on Human Language Technology. 1994. 307-312

[58] Zhang J Y, Zheng F, Li J, et al. Improved Context-Dependent Acoustic Modeling for Continuous Chinese Speech Recognition. In: Proceedings of the 7th European Conference on Speech Communication and Technology (EuroSpeech), Alborg, Denmark, 2001, 1617-1620

[59] Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition. IEEE Trans on Acoustic, Speech and Signal Processing, 1989, 77(2): 257-285

[60] 宋战江. 汉语自然语音识别中发音建模的研究: [博士学位论文], 北京: 清华大学, 2001

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

[62] Breiman L, Friedman J H, Olshen R A, et al. Classification and regression trees. Wadsworth, Belmont, CA, 1984

[63] Bahl L R, Souza P V, Gopalakrishnan P S, et al. Decision trees for phonological rules in continuous speech. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto Canada, 1991. 185-188

[64] Hwang MY, Huang X, Alleva F. Predicting unseen triphones with senones. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). Minneapolis, USA, 1993. 311-314

[65] Reichl W, Chou W. Decision tree state tying based on segmental clustering for acoustic modeling. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seattle, USA, 1998. 801-804

[66] Young S J. The general use of tying in phoneme-based HMM speech recognizers. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). San Francisco, USA, 1992. 569-572

[67] Willett D, Neukirchen C, Rottland J, et al. Refining tree-based state clustering by means of formal concept analysis, balanced decision treesand automatically generated model-sets. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). Phoenix, USA, 1999. 565-568

[68] Beulen K, Ney H. Automatic question generation for decision tree based state tying. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seattle, USA, 1998. 805~808

[69] 曹剑芬. 现代语音基础知识. 北京: 人民教育出版社, 1990

[70] 郑方, 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, 1997

[71] Renals S, Morgan N, Bourlard H M, et al. Connectionist probability estimators in HMM speech recognition. IEEE Trans on Speech and Audio Processing, 1994, 2(1): 161-174

[72] 冯俊兰. 口语语音识别的声学建模改进和解码方案研究: [博士学位论文],北京: 中国科学院声学研究所, 2001.

[73] Rose R, Paul D. A hidden Markov model based keyword recognition system. In: Proceedings International Conference on Acoustics, Speech and Signal Processing (ICASSP), Albuquerque, USA, 1990. 129-132

[74] Bourlard H, Hoore B D, Boite J M. Optimizing Recognition and Rejection Performance in Word- Spotting System. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). Adelaide, Australia, 1994. 373-376

[75] 陆正中. 口语对话系统中的语音识别研究: [硕士学位论文], 北京: 清华大学, 2002

[76] Rohlicek J R, Russel W, Roukos S, et al. Continuous Hidden Markov Modeling for Speaker-Independent Word Spotting. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). Glasqow, England, 1989. 627-630

[77] 张国亮, 徐明星, 李净, 等. 语音识别中基于两层词法树的跨词搜索算法. 清华学报(自然科学版), 2003, 7: 981-984

[78] Feng M W, Mazor B. Continuous word spotting for applications in telecommunications. In: Proceedings 2nd International Conference on Spoken Language Processing (ICSLP), 1992. 21-24

[79] Cox S, Rose R. Confidence Measures for the Switchboard Database. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1996. 511-514

[80] Katz S M. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer, IEEE Trans on Acoustic, Speech and Signal Processing, 1987, 35(3): 400-401

[81] Jelinek F. Self-organized language modeling for speech recognition. Readings in Speech Recognition, Morgan Kaufmann, San Mateo, CA, 1991. 450-506

[82] Jelinek F. Statistical Methods for Speech Recognition. The MIT Press, 1997.

[83] 牟晓隆. 汉语听写机的研究与实现: [硕士学位论文], 北京: 清华大学, 1998

[84] 燕鹏举. 对话系统中的自然语言理解研究: [博士学位论文], 北京: 清华大学, 2002

[85] Williams G, Renals S. Confidence measures from local posterior probability estimates. Computer Speech and Language, 1999, 13: 395-411

[86] Cox S, Dasmahapatra S. A high-level approach to confidence estimation in speech recognition. In: Proceedings of the 6th European Conference on Speech Communication and Technology (EuroSpeech), Budapest, Hungary, 1999. 41-44

[87] Asadi A, Schwartz R, Makhoul J. Automatic modeling for adding new words to a large-vocabulary continuous speech recognition system. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, Canada，1991，305-308

[88] Young S R. Detecting misrecognition and out-of-vocabulary words. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), Adelaide, Australia, 1994. 21-24

[89] Rivlin Z, Cohen M, A phone-dependent confidence measure for utterance rejection. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). Atlanta, USA , 1996. 515-518

[90] Willett D, Neukirchen C, Rigoll G. Efficient search with posterior probability estimates in HMM-based speech recognition. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seattle, USA, 1998. 821-824

[91] Weintraub M, Beaufays F. Neural-network based measures of confidence for word recognition. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). Munich, Germany, 1997. 21-24

[92] Koo M W, Lee C H, Juang B H. A new decoder based on a generalized confidence score. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seattle, USA, 1998. 213-216

[93] Wessel F, Macherey K, Schluter R. Using word probabilities as confidence measures. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seattle, USA, 1998. 225-228

[94] Evermann G, Woodland P C. Large vocabulary decoding and confidence estimation using word posterior probabilities. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), Istambul, Turkey, 2000. 2366-2369

[95] Gillick L, Ito Y. A Probabilistic approach to confidence estimation and evaluation. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). Munich, Germany, 1997. 879-882

[96] Schaaf T, Kemp T. Confidence measures for spontaneous speech recognition. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). Munich, Germany, 1997. 875-878

[97] Lin C. Word and acoustic confidence annotation for large vocabulary speech recognition. In: Proceedings of the 5th European Conference on Speech Communication and Technology (EuroSpeech). Rhodes, Greece, 1997. 815-818

[98] Bansal D, Ravishankar M K. New features for confidence annotation. In: Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP). Sydney, Australia, 1998.

[99] 熊振宇, 吴文虎, 徐明星. 置信度计算方法的比较和结合. 第六届全国人机语音通讯会议, 深圳, 2001.

[100] Paul C, Chun J, Daniel W, et al. Is this Conversation on Track. In: Proceedings of the 7th European Conference on Speech Communication and Technology (EuroSpeech). Alborg, Denmark, 2001. 2121-2124

[101] San S R, Pellom B, Ward W. Confidence Measures for Dialogue Management in the CU Communicator System. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), Istambul, TurKey, 2000. 1234-1240

[102] Ma C X, Randolph M A, Drish J. A Support Vector Machines-Based Rejection Technique for Speech Recognition. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), Salt Lake City, USA, 2001.

[103] 边肇祺, 张学工等. 模式识别. 北京: 清华大学出版社, 2000

[104] Siu M, Gish H. Evaluation of word confidence for speech recognition systems. Computer Speech and Language, 1999, 13: 299-319

致　谢

衷心感谢导师吴文虎教授和郑方副教授对本人的精心指导。吴文虎教授不光在专业方面以精深造诣给我以很好的学术指引，而且其严谨求实的治学作风和平易近人的待人原则，使我受益匪浅；郑方副教授忘我的工作精神以及严谨求实的科学作风将是我长期学习的榜样。在此，谨向两位恩师致以最诚挚的谢意！

感谢语音技术中心的其它老师，包括方棣棠教授、李树青教授和徐明星老师，以及实验室的全体同窗。他们对我的论文工作也给出了许多帮助，在此一并向他们表示感谢。

最后，衷心感谢我的妻子和家人，他们无私的爱和默默的关怀，一直伴随着我的奋斗过程。

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：_____

附录　博士就读期间完成的其它研究任务

汉语听写机中的搜索策略研究与系统集成

1. 搜索策略的研究背景

这部分研究内容基于清华大学计算机系语音技术中心开发的大词汇量、非特定人、连续汉语语音识别系统（又称听写机）*EasyTalk*。对于输入的连续语音流，*EasyTalk* 的识别算法将声学模型和语言模型结合在一起对其进行识别打分，一次得到最终的识别结果。

在 *EasyTalk* 中，采用时间同步的基于词 Viterbi 搜索算法[1][2]作为核心的识别算法。为进一步提高识别性能，采用了集成搜索策略，摈弃传统的声学层搜索与语言层搜索独立分开的两遍搜索策略，将声学层和语言层搜索合而为一，利用语言模型的预测作用，达到更好的识别性能。下面首先介绍集成搜索策略的体系结构，再对语言模型如何使用进行详细介绍。

2. 集成搜索策略

在大词汇量连续语音识别系统中，实现一个快速有效的搜索算法对整个系统的最终性能起着至关重要的作用。一个好的算法需要在搜索的过程中尽量使用准确的联合概率来衡量候选路径，至少要保证联合概率较大的路径尽可能地保留下来，直至搜索任务全部结束。近年来，国内外研究者们作了大量的研究工作，提出了不少实用的搜索算法。这些搜索算法，可以根据不同的研究角度，分为不同的类别。

从搜索空间角度，可以分为基于词法树的搜索和基于线性词表的搜索，这两种搜索算法的差别在于前者利用了词汇的发音具有大量的共享前缀的特性，从而减少了声学得分的重复计算。

从搜索时各种信息使用时机的角度，可分为一遍搜索（*One-pass*）[1]和多遍搜索（*Multi-pass*）[3]，一遍搜索是将所有可利用的信息按照某种准

则集成为一个合一的路径评价标准，通过一遍扫描语音信号便给出侯选结果的方法；而多遍搜索则是将这些信息根据不同的需要，在搜索的不同阶段，单独作为或部分集成为路径的评价标准，进行多遍扫描给出侯选的搜索算法。

集成搜索策略是一遍搜索策略在大词汇量连续语音识别中的一种具体形式。它将两个最主要的知识源：声学模型和语言模型集成在一起进行识别搜索，对路径打分时既有声学层面得分也有语言层面得分，将两个得分通过某种准则计算出合一的路径得分，合一的路径得分作为搜索过程中的唯一标准。采用集成搜索策略，通过一遍扫描语音信号便可以给出词候选识别结果。集成搜索策略的整体流程，如图 F-1 所示。

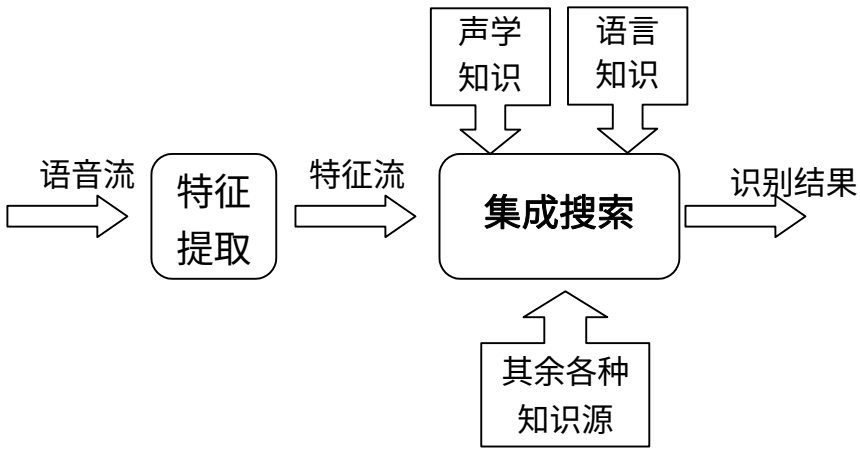


图 F-1 集成搜索策略的整体流程

集成搜索策略最主要的优点是发挥了语言模型知识的预测作用。传统的识别算法采用多遍搜索策略，将声学层和语言层分别独立进行搜索，如果声学层的识别候选中不包含正确的音节，在第二步语言层识别时，无论语言模型的能力再怎么强也不可能得到正确的词识别结果。集成搜索策略发挥了语言模型知识的预测作用，在声学层识别的同时语言层也开始进行识别，用词法树和语言模型对其进行指导，直接得到词识别结果，避免了多遍搜索策略中声学层识别错误时所产生的无法弥补的损失。

3. 语言模型预测

集成搜索策略要求语言模型知识尽可能早地加入都搜索中。Ney[1]提出的基于词的搜索算法是在词法树搜索到叶子结点时，即搜索到词结束时再进行语言模型打分，将语言模型得分加入到搜索路径整体得分中。这种方法基本实现了声学层识别与语言层识别的结合，而且计算复杂度不高，容易达到搜索算法实时的目的；但路径在词法树从根结点到叶子结点的阶段中，整个搜索过程将只由声学模型和简单的词法约束来控制，由于一般词法树从根结点到叶子结点所跨越的声学时间段相对都比较长，在这之间足以发生很大的偏差[4]。

为了在搜索中保证评价函数的合理性及尽早地应用语言模型知识，语言模型折入概率被引入到集成搜索策略中，定义如下：

$$g_{uv}(m)=\begin{cases}\sum\limits_{w\in\Pi(m)} p(w\mid u,v) & \text{如果}m\text{为非叶子结点}\\ p(w\mid u,v) & \text{如果}m\text{为叶子结点}\end{cases} \qquad \text{(F-1)}$$

式中的$(u,\ v)$是路径前导词，$p(w\mid u,v)$是三元文法语言模型(trigram)得分。其中 m 为词法树中的任一结点；$\Pi(m)$表示由结点 m 可以到达的叶子结点（词）的集合，称为叶子集。这个概率是词法树内部某结点所有可能的语言得分和，即可以定义为这个结点的语言模型得分，这一定义被称为概率的折入（*Factorization*）[5]。图 F-1 给出了一个具体例子。

为了简化折入概率的计算，(F-1)式中的求和计算可以用取最大值来代替，就向 *Viterbi* 算法中用最大值来代替求和一样，性能不会下降。即定义：

$$g_{uv}(m)=\begin{cases}\max\limits_{w\in\Pi(m)} p(w\mid u,v) & \text{如果}m\text{为非叶子结点}\\ p(w\mid u,v) & \text{如果}m\text{为叶子结点}\end{cases} \qquad \text{(F-2)}$$

由于在树状结构中，从任一结点可以到达的叶子结点的集合相当于从该结点的所有子结点出发可到达的叶子结点集合的总集，即

$$\Pi(m)=\bigcup_{n\in S(m)}\Pi(n) \qquad n\text{ 为 }m\text{ 的子结点} \qquad \text{(F-3)}$$

其中 *S(m)*表示结点 *m* 的子结点集。因此，在实际计算时，还可以用一个简单递归公式来进一步减少求和或取最大值的运算量，过程如下所示：

$$g_{uv}(m)=\begin{cases}\max\limits_{n\in S(m)} g_{uv}(n) & \text{如果}m\text{为非叶子结点}\\ p(w\mid u,v) & \text{如果}m\text{为叶子结点}\end{cases} \tag{F-4}$$

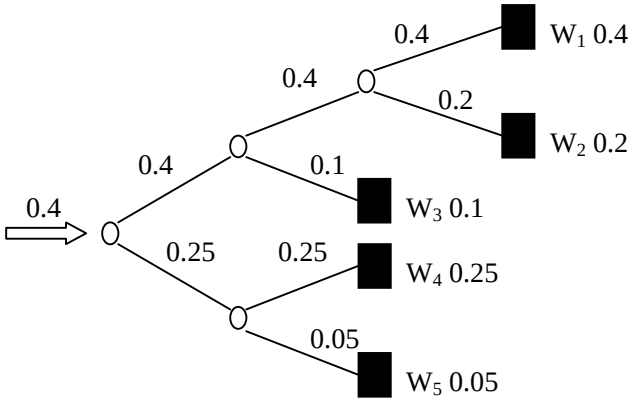


图 F-2 语言模型概率折入

在引入语言模型折入概率之后，不需到达词法树的叶子结点，就可以很方便地根据折入概率就语言模型得分加入到搜索路径总得分中，实现语言模型对搜索的预测，这就是语言模型预测技术[5][6]（*LMLA，Language Model Look-Ahead*）。

为了方便表述，引入定义：

$Q_{uv}(t,s)$ 表示前导词对为*(u,v)*的候选路径在词法树的约束下，在 t 时刻到达状态 s 的最佳得分；

$H_{uv}(t,s)$ 表示前导词对为*(u,v)*的候选路径在词法树的约束下，在 t 时刻到达状态 s 的最佳声学模型得分。

在词法树非叶子结点的任意状态和结点跳转时，搜索的评价函数可以定义为：

$$H_{uv}(t,s)=\max_{\sigma}\{q(x_t,s\mid\sigma)\cdot H_{uv}(t-1,\sigma)\} \tag{F-6}$$

$$Q_{uv}(t,s)=H_{uv}(t,s)\cdot\left[g_{uv}(n(s))\right]^{f} \tag{F-7}$$

其中，$n(s)$表示状态 s 所在的结点，f 是语言模型得分的加权系数。

由于每个结点包括叶子结点的语言模型概率都被语言模型的折入概率考虑在内，因此在到达叶子结点时不需要对 *n-gram* 加以单独考虑。

4. 声学模型预测

在语言模型的预测过程中，经过的词法树每个结点都需要计算语言模型折入概率，而从式(F-4)知道，该概率的计算复杂度与结点的叶子集规模成正比。因此，伴随着语言模型预测技术的除了搜索精度的提高之外，还有大量的求和计算或比较，以及对 *trigram* 概率值的频繁读取。虽然可以只根据具体情况计算出需要的折入概率而无需事先计算出所有的折入概率，甚至降低语言模型的精度只使用二元文法语言模型(*bigram*)，但是对于一个实时系统而言，仍然是很大的负担。

采用声学模型预测（*AMLA，Acoustic Model Look-Ahead*）[6]来解决语言模型预测计算复杂的问题。声学模型预测指的是每次在搜索即将进行一个新的结点时，即在一个新的模型基元的开始，暂时不直接扩展路径，而是先对扩展之后的路径进行在未来 N 帧内是否会被剪枝的预测。如果预测不会被剪枝，则该跳转被允许进行，否则的话则被禁止，从而使得计算复杂度很高的语言模型预测还没有开始时即禁止部分路径的扩展，减少了大量无效的计算。

为了简化计算，预测的区域不能太长，通常为 100ms 以内，这么短的时间内可以认为路径仍然处于一个模型基元内，不会发生模型基元之间的跳转，预测也就用一个 HMM 对其进行打分即可。从声学模型预测的定义看出非常类似于剪枝过程，所以可以根据剪枝过程来实现声学模型预测。一般来说，首先计算出一个尽可能准确的近似声学得分，称作声学预测得分。然后在预测的过程中，声学预测得分与当前路径的历史得分相结合，通过一定的比较准则来完成对路径扩展可能性的检查。具体描述如下：

设 $\tilde{q}(H;t,\Delta t)$ 为声学预测跨越时间从 t 到 $t+\Delta t$，跨越时间内所处的模型

基元为 H。然后计算出需要声学模型预测的所有路径中最大的路径得分 $\tilde{Q}_{uv}(t,s)$，利用剪枝阈值 f^{AMLA}(<1)完成声学模型预测，检查需要声学模型预测的所有路径，并拒绝掉满足

$$Q_{uv}(t,s)\cdot\tilde{q}(H;t,\Delta t)<\tilde{Q}(t)\cdot\max_{H}\tilde{q}(H;t,\Delta t)\cdot f^{AMLA} \tag{F-8}$$

的所有结点的扩展可能。

因为声学模型预测的引入就是为了减少语言模型预测带来的巨大计算量，所以声学模型预测本身的计算复杂度不能太大。在声学模型预测时可以采用非常简单的声学模型。当声学模型采用上下文相关声学模型时，也可以使用上下文无关声学模型进行预测，这时只需要关心识别基元的中心元素即可。另外，为了避免在预测时对 N 帧特征的状态划分问题，Ortmanns[6]提出可以使用单状态的 HMM 进行预测，此时的预测计算复杂度最小，只需要将 N 帧特征对一个状态的打分累计即可，实验证明并明显不影响预测的性能。

为了保证集成算法的实时性，除了声学模型预测外，传统的几种剪枝技术也必须合理应用，如灰度图剪枝等等。只要剪枝技术运用合理，使用越多的剪枝技术可以得到越快的识别算法。

5. 实验结果与分析

声学模型的识别基元定义和训练方法和第二章完全相同。声学模型训练集都从 863 数据库中得到，863 数据库共有 80 人的男声数据。所有的语音都是在低噪音的环境下录制而成，采样率为 16kHz。训练数据库包括 70 个男声数据，约 36400 句语音。词表大小约 50000 词，其中单音节词占 20.81%，双音节词占 65.50%，三音节词占 7.36%，四音节词占 6.33%。语言层采用基于 *bigram* 的统计语言模型，训练语料库包括 1993 和 1994 年《人民日报》的全文，以及《市场报》、《新华社文稿》、《经济日报》的摘编等约 2 亿字的文本，并预先进行了分词和词号标注。在语言模型中，使用一种基于 Turing 概率估计的平滑算法[4]来解决由于数据稀疏而造成的 0 概率问题，从而降低语言模型的困惑度。

测试数据库也从 863 数据库中得到，包括 2 个男声数据 1000 句、约 4500s 的语音样本，在一台 PentiumII 450 MHz 的计算机上进行实验，识别结果下表所示：

表 F-1 集成搜索策略的识别性能

算法	字识别率/%	识别时间/s
集成搜索策略	92.3	4195

从实验结果中看，集成搜索算法可以取得较高的识别性能和基本实时的搜索速度。

6. 参考文献

有关这部分的详细研究内容及相关引用，请参阅下列文献：

[1] Ney H, Haeb-Umbach R, Tran B H, et al. Improvements in Beam Search for 10,000-word Continuous Speech Recognition. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). San Francisco, USA, 1992. 13-16

[2] Ortmanns S, Ney H, Seide F, et al. A Comparison of Time Conditioned and Word Conditioned Search Techniques for Large Vocabulary Speech Recognition. In: Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP). Philadelphia, USA, 1996. 2091-2094

[3] Li Z, Boulianne G. Bi-directional Graph Search Strategies for Speech Recognition. Computer Speech and Language, 1996, 10(4): 295-321.

[4] 武健. 汉语语音识别中统计语言模型的构建及其应用: [硕士学位论文], 北京: 清华大学, 2000

[5] Ortmanns S, Ney H, Eiden A. Language-Model Look-ahead for Large Vocabulary Speech Recognition. In: Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP). Philadelphia, USA, 1996. 2095-2098

[6] Ortmanns S, Eiden A, Ney H, et al. Look-ahead Techniques for Fast Beam Search. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). Munich, Germany, 1997. 1783-1786

个人简历、在学期间的研究成果及发表的论文

个人简历

张国亮，1977 年 01 月 08 日出生于陕西省宝鸡市，1994 年 9 月考入清华大学计算机科学与技术系学习，1999 年 7 月本科毕业并获得工学学士学位，同年 9 月免试保送清华大学计算机科学与技术系攻读计算机科学与技术学科工学博士学位至今。

在国际和国内学术刊物上发表的论文（第一、二作者）

[1] Zhang G L, Zheng F, Wu W H. Tone recognition of Chinese continuous speech. In: Proceedings of the 2nd International Symposium of Chinese Spoken Language Processing (ISCSLP). Beijing, China, 2000. 207-210

[2] Zhang G L, Zheng F, Wu W H. A two-layer lexical tree based beam search in continuous Chinese speech recognition. In: Proceedings of the 7th European Conference on Speech Communication and Technology (EuroSpeech). Alborg, Denmark, 2001. 1801-1804

[3] Zhang G L, Yan P J, Xu M X, et al. An automatic speech recognition strategy directed by the semantic knowledge in dialogue system. In: Proceedings of the 3rd International Symposium of Chinese Spoken Language Processing (ISCSLP). Taipei, 2002. 319-322

[4] 张国亮，徐明星，李净等．语音识别中基于两层词法树的跨词搜索算法．清华学报(自然科学版). 2003, 7: 981-984

[5] Zheng F, Zhang G L. Integrating the energy information into MFCC. In: Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP). Beijing, China, 2000. 389-392

[6] Zheng F, Zhang G L, Song Z J. Comparison of Different Implementations of MFCC, In: Journal of Computer Science and Technology, Allerton Press (New York). 2001, 16(6): 582-589

[7] Sun H, Zhang G L, Zheng F, et al., Using word confidence measure for OOV words detection in a spontaneous spoken dialog system, in Proceeding of the 8th European Conference on Speech Communication and Technology (EuroSpeech). Geneva, Swiss. 2003. 2713-2716