摘 要

汉语语音听写机（CDM，Chinese Dictation Machine）是一种非特定人、大词汇量的连续或连接词汉语语音识别系统，目的是将人的语流自动转化为相应的文本信息。本文对汉语语音听写机涉及的一些关键技术进行了较深入的研究，主要针对两个问题：一、声学模型的建立和改进；二、N-gram 统计语言模型的建立与改进。具体包括：

1. 研究了经典隐式马尔可夫模型（HMM）及其派生模型，提出了一种用于声学识别的高斯混合分段模型（GMSM，Gaussian Mixture Segmentation Model）。该模型采用高斯混合分布描述观察矢量在特征空间中的分布，同时对 HMM 的训练过程和识别中的 Viterbi 解码过程进行简化：训练时使用观察序列的线性或非线性分段确定状态，识别时则用帧同步搜索算法实现状态解码。
2. 研究了 GMSM 实现中的一些具体问题，包括方差矩阵估计的简化；迭代初值的选取；训练数据不足和奇异数据的解决；迭代终点的确定；识别速度的提高等。
3. 对 GMSM 的总体性能进行了实验，测试和分析了训练数据帧移的变化、计分策略的简化，以及训练收敛阈值的变化对 GMSM 性能的影响。
4. 确定了汉语语音听写机中语言模型数据库的层次结构。
5. 分析了基于退化算法（Back-off Algorithm）的 N-gram 语言模型在实现中的一些问题，提出了修正退化频度估计算法。它对原退化频度估计算法中折扣系数和零概率分配系数的计算进行了修正，以便构建实用的 N-gram 统计语言模型。
6. 对 N-gram 语言模型的搜索策略进行了分析，给出了一种实用的搜索方法。
7. 依据包括 1993 年人民日报全文、市场报摘编、以及新华社文稿摘编在内的共约 4 千万词的语料库和 24713 个词的基本词表，实现了基于修正退化频度估计算法的 Tri-gram 语言模型。
8. 对实现的 Tri-gram 语言模型的性能进行了实验和分析。考察了 GMSM 和基于修正退化频度估计的 Tri-gram 语言模型用于连接词汉语语音听写机原型的综合性能。

关键词：汉语语音听写机，高斯混合分段模型（GMSM），修正退化频度估计算法，N-gram 语言模型

ABSTRACT

CDM (Chinese Dictation Machine) is a speaker independent, large vocabulary, continuous speech or connected word based Chinese speech recognition system, used to transform human speech to corresponding text information. In this paper, some key techniques for CDM have been studied. Two main aspects are focused on: the establishment and improvement of the acoustic model; the implementation and refinement of N-gram statistical language model. In detail, the following are included:

1. The classic HMM (Hidden Markov Model) and some derived acoustic models are studied, and a new model named GMSM (Gaussian Mixture Segmentation Model) is proposed for recognition. GMSM uses Gaussian Mixture PDF (Probability Density Function) to describe the distribution of feature vectors in the feature space. It simplifies HMM's training and Viterbi decoding process, using linear or non-linear segmentation of observation sequence to determine the states while training, using a Frame-Synchronous Network Search Algorithm to decode states while recognizing.
2. Some details of GMSM implementation are investigated, including simplification of variance matrix estimation, initial estimates of GMSM parameters, solution of training data insufficiency and oddity, decision of iteration termination, and acceleration of recognition process etc.
3. Experiments are made to evaluate GMSM's overall performance, and how the variations of training frame shift interval, scoring scheme, and convergence threshold influence GMSM's performance are studied.
4. The structure of CDM Language Model Database is settled.
5. Some implementation problems of the Back-off Algorithm based N-gram language model are analyzed, and a modified back-off frequency estimation algorithm is proposed to amend the calculation of discount coefficient and normalizing constant in the origianl algorithm, so that an applicable N-gram statistical language model can be achieved.
6. The searching strategies of N-gram language model are investigated, and a practical searching method is provided.
7. Based on the word list of 24713 words and the training corpus of 40,000,000 words containing the full text of the People's Daily 1993, digestion of the Market News and manuscripts of the XINHUA news agency, a Tri-gram language model using the modified back-off frequency estimation algorithm is implemented.
8. Experiments are carried out to analyze the performance of the implemented Tri-gram language model. The overall performance of a CDM prototype using GMSM and the modified back-off frequency estimation algorithm based Tri-gram language model is also studied.

KEY WORDS: Chinese Dictation Machine, Gaussian Mixture Segmentation Model, Modified Back-off Frequency Estimation Algorithm, N-gram Language Model

致 谢

感谢我的导师郑方博士对我论文工作的悉心指导。郑方老师兢兢业业的工作作风，严谨的科学态度和不断创新的精神是我在今后学习和工作中的榜样。感谢吴文虎教授和胡起秀教授带领我进入语音识别这个领域，多年来，吴老师为人师表的治学作风，对我的谆谆教诲和严格要求是我成长的动力。在我的研究工作尤其是语言模型方面的工作中，方棣棠教授给予了我很多详细的指导，帮助我把握了工作的方向。同时，方老师渊博的知识和对待科学问题的完全认真的态度给我留下了深刻的印象，这些都是我的宝贵财富。

感谢一同工作的詹津明、柴海新、徐明星、宋战江、郭庆、王霞、武健、翁武斌等同学，他们的帮助和友谊是我工作不断进步的条件。特别感谢詹津明同学在语言模型方面与我的有益讨论，武健同学在声学模型性能评测方面对我的帮助。

感谢所有关心、支持和帮助我的老师，同学和朋友们。

目 录

图 表 目 录

第一章 综述

语音识别（SR，Speech Recognition）的目的是让计算机“听懂”人说话，它是发展人机语音通信和新一代智能计算机的主要组成部分，也是目前业界普遍认为很有前途的一项技术。随着计算机处理能力和存储能力的不断增强，一些非常复杂的语音识别算法能够得以实现，语音识别的性能得以不断提高，这种人与计算机的自然交互方法也日益引起人们的重视。

1.1 语音识别的分类与汉语语音听写机

语音识别按照被识别人的范围可以分为特定人（Speaker Dependent）和非特定人（Speaker Independent）语音识别；按照词汇量的大小可以分为小词汇量（Small Vocabulary）和大词汇量（Large Vocabulary）语音识别；按照说话方式可以分为孤立词（Isolated Word）、连接词（Connected Word）和连续语音（Continuous Speech）的语音识别。

非特定人语音识别要求语音识别系统能够识别具有不同年龄、性别、口音的人的语音。由于不同人发音的速度，习惯，声学特性等各方面均不相同，这就要求计算机能够提取寻找、归纳其中的相似性，达到说话人不经过训练就能够使用的目的，这是非常困难的。

在语音识别中，一般把规模在几百个这样数量级的词表称为中小词汇量，几千个以上这样数量级的词表称为大词汇量。能够识别的词表的大小也是评价语音识别系统性能的重要方面。对于中小词汇量的识别器，识别算法可以作较大幅度的简化来争取效率，同时基本不降低识别性能。然而，这样的系统并不能够通过简单地增加词表容量构成大词汇量的识别系统。在大词汇量识别系统中，不同词汇在识别特征空间的区分度相对减小，识别难度随词表容量的加大而上升。

连续或连接词语音识别是使用基本语音单元的有限个模型对输入的连续语流或部分连续的语流进行识别，得到待识别语流对应的语音单元的序列。这种方式显然比孤立词识别更加自然，但是连续发音时语流中相邻语音单元界限模糊，还会因相互影响产生吃音、协同发音等各种变化，给识别造成很大困难。

汉语语音听写机（CDM，Chinese Dictation Machine）是非特定人、大词汇量的连续或连接词识别系统，目的是由计算机将人的语流转化为相应的文本信息。在当今人与计算机交互日益频繁的条件下，探索高效而自然的交互方式是人们不断努力的目标。汉语语音听写机正是这样一种十分有潜力的人机交互系统，它可望把人从不自然的信息输入方式中解放出来，从而推进计算机的应用和发展。

1.2 汉语语音听写机中关键技术

特定人、小词汇量的孤立词识别系统在 70 年代后期就达到了令人满意的结果，而非特定人、大词汇量、识别连续语音的听写机在 80 年代有了初步进展，进入 90 年代，某些方面进一步有了实质性的发展。总体来讲，汉语语音听写机的关键技术包括声学特征抽取，声学建模和语言建模等几个方面。

语音的声学特征主要包括时域特征、频域特征等类型。声学信号的时域参数在语流切分等方面有重要的作用，例如短时平均能量、过零率、基音周期等。对于声学建模来说，目前实用的特征基本是倒谱和倒谱的派生参数。倒谱系数主要包括 LPC 导出倒谱和 mel 刻度的倒谱两种。Wilpon[Wilpon 89]等把加权的倒谱和差分倒谱串接起来形成一个大的矢量作为声学特征矢量，取得了好的效果。此外，也有人使用两维的“ 时-频谱 ”(Time-Frequency Spectrum)表示语音信号的特征[Wilpon 91]，它考虑了语音信号的时变特征，是频谱的一种高阶时间派生参数。

在 70 年代，声学建模的主要技术是以动态规划（DP，Dynamic Programming）为基础的动态时间弯折[Vintsjuk 68]（Dynamic Time Warping）。进入 70 年代后期，人们寻找了各种新的声学建模方法。具有重要意义的是矢量量化技术[Linde 80]（VQ，Vector Quantization），它具有很好的数据压缩能力和比较理想的聚类功能。到 80 年代中期，隐式马尔可夫模型的广泛研究与应用，使声学建模有了实质性的进展。HMM 能描述不同层次的语音单元，由 Viterbi 解码可以得到与语音序列对应的最佳状态序列，便于解决连续语音识别的问题。但是，经典 HMM 在实用时也存在着一些缺陷，例如 HMM 的时空复杂度很高，导致训练和识别的效率低，难以直接应用；语音帧间的独立性假设；马尔可夫链的假设等等。针对这些缺陷，尤其是复杂度过高这一点，90 年代人们建立了各种 HMM 的派生模型。另外，80 年代中期重新引起重视的神经网络（Neural Networks）研究，也给语音识别带来了新的活力。由于人工神经元网络具有刻划各种复杂分类边界的能力，十分适用于语音识别领域。目前，神经网络的研究还不是特别成熟，但是在某些方面已经显示了一定的威力。神经网络还可以与 HMM 综合应用于声学建模：由神经网络完成静态的模式划分问题，用 HMM 完成时间对准问题[Franzini 90, Morgan 90]，使神经网络更容易地应用于连续语音识别系统。

汉语语音听写机的目的是将语音流转换成相应的文字序列。对于听写机来说，语言模型是另一个十分重要的问题。由于声学信号的动态改变，瞬时和随机性，单靠声学模式的匹配与判断，是很难完成语音到语言文字的正确转换的。较高层次的语言知识的利用可以大幅度地减小声学模式匹配的模糊性，从而提高听写机的识别正确率。概括来讲，语言模型可以分为两类，基于统计的语言模型（Statistical Language Model）基于知识的语言模型（Knowledge-based Language Model）。在当前的技术条件下，基于统计的语言模型在实际应用中处于主流地位。它通过对大量实际语料的统计获得词之间的连接信息，从而评价一个词串是否为语言中合理的语句。这在一定程度上回避了基于规则的语言模型其规则集难以严

格和完备，以及语义规则难于形式化等困难。因此，现阶段实用语言模型中的规则模型主要作为统计模型的补充，对统计模型的结果进行校验和改进。N-gram 统计模型是最初引入而且应用最广泛的一种语言模型，该模型最初由 Jelinek 等人提出[Jelinek 83]。但是，N-gram 模型面临的最大困难是训练语料过于稀疏。针对这一困难，Nadas 给出了图灵估计变形的概率估计方法[Nadas 85]，Katz 给出了一种基于图灵估计的退化频度估计算法[Katz 87]等，力求在一定程度上解决训练数据稀疏的问题。

1.3 本文的主要工作和论文安排

本文的主要工作是对汉语语音听写机的关键技术进行研究并加以实现，同时对其性能进行分析、对比和评价。

作者主要完成了如下的工作：

(1) 研究了经典隐式马尔可夫模型（HMM）及现有的一些派生模型，提出了一种高斯混合分段模型（GMSM，Gaussian Mixture Segmentation Model）作为汉语语音听写机的声学识别模型。
(2) 研究了 GMSM 实现中的一些具体问题，并对其性能进行了实验和分析。
(3) 确定了汉语听写机语言模型数据库的层次结构。
(4) 分析了基于退化（Back off）的 n-gram 语言模型在实现中的一些问题，提出了修正退化频度估计算法。同时，对相应的搜索策略进行了研究，给出了实用的搜索方法。
(5) 实现了基于修正退化频度估计算法的 Tri-gram 语言模型，对其性能进行了实验和分析。
(6) 与全组师生一起，实现了包括 GMSM 和修正退化频度估计语言模型的连接词汉语语音听写机 ST98 原型。

本文的篇章安排如下：第二章简述声学特征与识别基元的选取；第三章探讨声学模型，提出 GMSM；第四章描述语料处理方法与语言模型的数据库；第五章探讨语言模型，提出修正退化频度估计算法；第六章介绍实验数据及分析，给出结论。

第二章 声学特征与识别基元

2.1 声学特征

特征提取与选择是语音识别的一个重要环节。特征提取解决了时域语音信号的数字表示问题，而特征选择则通过选取有效的特征为模式划分部分提供数据。特征提取与选择的好坏直接影响到识别器的性能。

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(CEP)是语音信号的又一个特征，有基于线性预测分析(LPC)的倒谱即 LPCC，有基于 Mel 频率弯折的倒谱即 MFCC。

基于听觉模型的特征参数提取如感知线性预测(PLP)分析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，并取得了一定的成功。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。

在本文的研究工作中，选取目前普遍采用的 LPCC 和自回归倒谱特征。

2.1.1 基于 LPC 的倒谱系数 LPCC

倒谱(Cepstrum，简称倒谱或 CEP)参数是重要的语音特征参数，它是对语音进行同态处理的产物。处理过程用公式表示为[Rabiner 78]

$$CEP(t) = DFT^{-1}\left(ln\left|DFT\left(Frame(t)\right)\right|\right)。\tag{2.1.1}$$

其中 $Frame(t)$ 表示第 t 帧语音，$DFT(\cdot)$ 和 $DFT^{-1}(\cdot)$ 分别表示离散傅里叶变换和离散傅里叶逆变换。从公式可以看出，倒谱参数 $CEP(t)$ 是时域特征，但由于进行了同态分析，因此只需要很少的前几阶就可以包容语音信号的大部分信息，起到了数据压缩效果。

线性预测系数(LPC)是对加窗语音波形频谱的全极点(all-pole)近似，在利用自相关(auto-correlation)算法求得 LPC 后，使用下面的公式可以计算出该帧语音的 LPC 倒谱(**LPCC**, Linear Predictive Coding Cepstrum)系数[Davis 1980]为

$$LPCC_t(d) = LPC_t(d) + \sum_{k=1}^{d-1} \frac{k-d}{d} LPCC_t(d-k) \cdot LPC_t(k), \quad d = 1,2,\ldots,D \text{ (2.1.2)}$$

其中 $LPC_t(d)$ 和 $LPCC_t(d)$ 分别表示第 t 帧语音的第 $d(d = 1,2,\ldots,D)$ 维 LPC 系数和 LPCC 系数。

2.1.2 自回归分析

自回归分析(**ARA**, auto-regression analysis)或线性回归分析(**LRA**, linear regression analysis)或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，ARA 具有下面的形式[Furui 86]

$$ARC_t(d) = G \cdot \sum_{n=-n_0}^{n_0} n \cdot CEP_{t+n}(d), \quad 1 \le d \le D \text{ (2.1.3)}$$

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

$$G = 1 \Big/ \sqrt{\sum_{n=-n_0}^{n_0} n^2} \text{ 。} \text{ (2.1.4)}$$

自回归倒谱也称为 Delta Cepstrum [Soong 86, Rabiner 88]。

在本文的工作中，取 n_0=2。

2.2 识别基元的选取

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。这两个方面需要综合考虑。

由于汉语是一个音节性强的语音，一个汉字对应于一个音节，所以采用音节作为声学识别的基元是一种自然的选择。除音节之外，采用半音节作为识别基元也是一种较好的方法，采用半音节作为识别基元可以减少识别基元的数目，同时保持一定的稳定性，对于声学搜索也有益处。在本文对声学模型研究的实验中，采用的是汉语全音节作为识别基元。

第三章 声学模型研究与实现

对语音识别基元进行声学建模是汉语听写机中声学层面处理的关键步骤。声学模型用来描述识别基元对应的特征矢量序列的产生过程。通过声学建模，可以估计待识别特征矢量序列所对应的语音识别基元，从而完成特征矢量序列到语音识别基元的识别转换。

声学建模的方法主要可以分为两类，即基于概率统计模型的方法和基于人工神经元网络的方法。在当前的汉语听写机系统中，前者占据着主导的地位。从理论上来讲，只要神经元网络的层数和每层的节点数选取比较适合，基于人工神经元网络的声学模型及模式分类器可以在高维空间中划分非常复杂的分界面[Pao 92]，这正是语音识别中的声学分类器所需要的。然而，正是因为这些问题尚没有很好地解决，加上人工神经元网络的训练过程控制困难，训练数据量大，从而限制了其性能的发挥。目前，该类模型主要处于研究阶段，也在实际的汉语听写机声学建模中作为一种辅助手段加以应用。

对于基于概率统计的建模方法，经过多年发展，其主流仍是隐式马尔可夫模型（HMM, Hidden Markov Model）及其各种变形。为克服经典 HMM 较高的时空复杂度及其自身的一些局限，从经典 HMM 中派生了各种比较实用的模型，例如链式隐马尔可夫模型，中心距离连续概率模型（CDCPM）[zheng 97]，以及本文提出的高斯混合分段模型（GMSM）等。本章主要分析经典 HMM 及一些 HMM 派生模型原理与实现。

3.1 隐式马尔可夫模型（HMM）

用 HMM 描述语音信号的产生过程是 80 年代语音信号数字处理技术的一项重大进展，用此模型解决语音识别中声学层面的模式划分问题取得了巨大的成功。

3.1.1 HMM 定义

HMM 是基于马尔可夫链的。马尔可夫过程是一个随机过程$\{S(t): t \in T\}$，它具备这样的性质，即已知 t 时刻过程所处的状态$s_t = S(t)$，在 t 时刻以后过程将要到达的状态与 t 时刻以前过程所处的状态无关，这个性质也称为过程的无后效性或马尔可夫性。

马尔可夫过程$\{S(t): t \in T\}$可能取值的全体构成状态空间，可以是连续的或离散的；马尔可夫过程的指标集 T 也可以是连续的或离散的。

对一个状态空间 I 和指标集 T 离散的随机过程$\{S(t): t = 0,1,2,\cdots\}$，若满足

$$P\{S(t+1) = s | S(0) = s_0, S(1) = s_1, \cdots, S(t) = s_t\} = P\{S(t+1) = s | S(t) = s_t\} \quad (3.1.1)$$

则称之为马尔可夫链。马尔可夫链在 t 时刻的一步条件转移概率

$$a_{ij}(t) = P\{S(t+1) = j | S(t) = i\} \quad (3.1.2)$$

称为 t 时刻状态 i 到状态 j 的转移概率。显然有

$$a_{ij}(t) \geq 0, \quad i, j \in I \quad (3.1.3)$$

$$\sum_{j \in I} a_{ij}(t) = 1, \quad i \in I \quad (3.1.4)$$

隐式马尔可夫模型(HMM)由两个相互关联的过程相互作用而成：一个是状态空间有限的马尔可夫链，一个是随机函数集。HMM 在任何时刻 t 下所处的状态 s_t 隐藏在系统内部，不为外界所见，外界只能得到系统在该状态下提供的实 R^Q 空间中的一个随机矢量，该随机矢量的发生概率由当前状态相关的随机函数给出。HMM 的状态转移由状态转移概率矩阵$\{a_{ij}\}$控制。

一个 HMM 由下面一些参数表征。(1) $\overline{\mathbf{N}} = \{1,2,3,\cdots,N\}$：模型的状态集合，$s_t = s(t) \in \overline{\mathbf{N}}$ 表示系统在 t 时刻所处的状态；(2) $A(t) = \{a_{ij}(t)\}_{N \times N}$：状态转移概率矩阵，$a_{ij}(t) = P\{s(t+1) = j | s(t) = i\}$；(3) $B = \{b_j(\cdot)\}_{N \times 1}$：观察符号输出概率(密度)矩阵，$b_j(x) = P_d\{output = x | state = j\}$；其中 $P_d\{\cdot\}$ 表示事件概率或概率密度。(4) $\pi = \{\pi_i\}_{N \times 1}$：初始概率分布，$\pi_i = P\{s(1) = i\}$。

状态概率转移矩阵一般是时间的函数，如果与时间无关，那么相应的 HMM 称为齐次的，此时

$A = \{a_{ij}\}_{N \times N}$：状态转移概率矩阵，$a_{ij} = P\{s(t+1) = j | s(t) = i\}$。

一个有 N 个状态的齐次 HMM 可以表示为 $\Lambda = \{\pi, A, B\}$。

3.1.2 HMM 的基本问题及解决

用 HMM 来完成语音识别的研究时，需要解决以下三个问题。

1. 训练。若有一个 HMM，需要根据该系统所给的若干观察序列 O 确定它的三项特征参数。所有的输出构成一个学习样本集合，其中每个观察序列 O 称为一个学习样本。设有 M 个样本，此集合可以记为$\{O^{(m)}, m = 1 \sim M\}$。确定 HMM 特征参数的准则是最大似然准则。
2. 计分。若已知一个 HMM 的三项特征参数，需要对系统可能产生的任何观察序列 O 计算其产生的概率。
3. 状态解码。同样已知三项特征参数，若得到了该系统产生的某个观察序列 O，需要估计该系统产生此序列 O 时最可能经历的状态序列。

“向前－向后”(Forward-Backward)算法或 Baum-Welch 算法[Baum 72]通过引入向前部分概率函数($1 \le j \le N$)

$$\alpha_t(j) = P\{o_1, o_2, \cdots, o_t, s_t = j | \Lambda\} = \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} b_j(o_t), 2 \le t \le T \tag{3.1.5}$$

$$\alpha_1(j) = \pi_i b_i(o_1) \tag{3.1.6}$$

及向后部分概率函数($1 \le i \le N$)

$$\beta_t(i) = P\{o_{t+1}, o_{t+2}, \cdots, o_T | s_t = i, \Lambda\} = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j), 1 \le t \le T-1 \tag{3.1.7}$$

$$\beta_T(i) = 1 \tag{3.1.8}$$

利用一组迭代公式可以解决第一个问题[Baum 72, Huang 89]。

$$\overline{\pi}_i = \frac{\alpha_1(i)\beta_1(i)}{\sum_{i=1}^{N} \alpha_T(i)} \tag{3.1.9}$$

$$\overline{a}_{ij} = \frac{\sum_{t=1}^{T-1} \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{t=1}^{T-1} \alpha_t(i) \beta_t(i)} \tag{3.1.10}$$

$$\overline{b}_i(x) = f(\Lambda, O) \quad (\text{是} O \text{和} \Lambda \text{的函数}) \tag{3.1.11}$$

其中$\bar{b}_i(\cdot)$的迭代公式视 HMM 的不同类型有所不同。

第一个问题解决之后，第二个问题由向前部分概率可以得到：

$$P\{\mathbf{O}|\Lambda\}=\sum_{i=1}^{N}\alpha_T(i) \tag{3.1.12}$$

Viterbi 解码算法[Viterbi 67]可以用来解决第三个问题。令：

$$\Phi_1(j)=\pi_j b_j(o_1),\quad 1\le j\le N \tag{3.1.13}$$

进行如下递归($2\le t\le T, 1\le j\le N$)：

$$\Phi_t(j)=\max_{1\le i\le N}\left[\Phi_{t-1}(i)\cdot a_{ij}\right]b_j(o_t) \tag{3.1.14}$$

及

$$\Psi_t(j)=\arg\max_{1\le i\le N}\left[\Phi_{t-1}(i)\cdot a_{ij}\right]b_j(o_t) \tag{3.1.15}$$

最后得到：

$$s_T^{(ML)}=\arg\max_{1\le i\le N}\left[\Phi_T(i)\cdot 1\right]\cdot 1 \tag{3.1.16}$$

$$s_t^{(ML)}=\Psi_{t+1}(s_{t+1}^{(ML)}),\quad 1\le t\le T-1 \tag{3.1.17}$$

记$S^{(ML)}=\left\{s_t^{(ML)}|1\le t\le T\right\}$，此为最大似然(ML, Maximum Likelihood)状态序列，那么该 HMM 产生这个状态序列的概率为：

$$\begin{aligned}P_d^{(ML)}&=\Phi_T(s_T^{(ML)})=\pi_{s_1^{(ML)}}\cdot b_{s_1^{(ML)}}(o_1)\prod_{t=2}^{T}a_{s_{t-1}^{(ML)}s_t^{(ML)}}\cdot b_{s_t^{(ML)}}(o_t)\\&=\left(\pi_{s_1^{(ML)}}\prod_{t=2}^{T}a_{s_{t-1}^{(ML)}s_t^{(ML)}}\right)\cdot\left(\prod_{t=1}^{T}b_{s_t^{(ML)}}(o_t)\right)=P\left\{S^{(ML)}|\Lambda\right\}\cdot P_d\left\{O|S^{(ML)},\Lambda\right\}\end{aligned} \tag{3.1.18}$$

事实上，由全概率公式，产生该观察序列的概率(密度)为：

$$\begin{aligned}P_d\{O|\Lambda\}&=\sum_S P_d\{O,S|\Lambda\}=\sum_S P_d\{O|\Lambda,S\}\cdot P\{S|\Lambda\}\\&=\sum_S\left(\pi_{s_1}\prod_{t=2}^{T}a_{s_{t-1}s_t}\right)\cdot\left(\prod_{t=1}^{T}b_{s_t}(o_t)\right)\\&=\sum_S\left(\pi_{s_1}\cdot b_{s_1}(o_1)\prod_{t=2}^{T}a_{s_{t-1}s_t}\cdot b_{s_t}(o_t)\right)\end{aligned} \tag{3.1.19}$$

其中$S=\left\{s_t|1\le t\le T\right\}$是任意一种状态序列。因此 Viterbi 算法给出的是该和式中的一项，是最大似然(ML, Maximum Likelihood)状态序列。

语音识别中所有的 HMM 一般都是从左向右结构，即$a_{ij}=0,\ j<i$。从左

向右结构又分为无跳跃式($a_{ij}=0,\ j<i, j>i+1$)和有跳跃式。

下面，如无特殊说明，都假定是在无跳跃式的从左向右结构上讨论问题。

3.1.3 HMM 的分类

根据观察输出概率矩阵中的函数$b_j(x)$是基于 VQ、连续密度还是二者的综合，HMM 又分为离散 HMM (DHMM, Discrete HMM) 、连续密度 HMM (CDHMM, Continuous Density HMM, 简称 CHMM) 和半连续 HMM (SCHMM, Semi-Continuous HMM)[Huang 1989]。这三种 HMM 的 A 矩阵具有相同的特性，由于：

$$\begin{aligned} P_d\{O|\Lambda\} &= \sum_S P_d\{O,S|\Lambda\} = \sum_S P\{S|\Lambda\}\cdot P_d\{O|\Lambda,S\} \\ &= \sum_S\left(\pi_{s_1}\prod_{t=2}^{T}a_{s_{t-1}s_t}\right)\cdot\left(\prod_{t=1}^{T}b_{s_t}(o_t)\right) \end{aligned} \tag{3.1.20}$$

因此，本节中我们仅对

$$P_d\{O|\Lambda,S\} = \prod_{t=1}^{T}b_{s_t}(o_t) \tag{3.1.21}$$

中的$b_{s_t}(o_t)$进行讨论。

1. 离散 HMM（DHMM）

DHMM 是基于矢量量化(VQ, Vector Quantization)技术的。它把特征向量空间分成若干个子空间，每个子空间用一个中心向量来表示，表征这个中心向量的是一个码字(codeword)，所有码字的集合构成码本(codebook)。

在计算概率$b_{s_t}(o_t)$时，取而代之的是计算$b_{s_t}\left(V(o_t)\right)$，这里$V(\cdot)$表示把向量量化后所对应的码字代号。

这里对$b_{s_t}(v)$的估计比较容易，通过某种计数的方法就可以实现：

$$\bar{b}_i(v) = \frac{\sum_{t=1}^{T}\alpha_t(i)\cdot\beta_t(i)}{\sum_{t=1}^{T}\alpha_t(i)\cdot\beta_t(i)}\Bigg|_{V(o_t)=v} \tag{3.1.22}$$

式中 v 表示码字序号。

很显然，由于一个子空间里的所有向量都用一个码字来代替，量化误差会很

大，因此 DHMM 的描述误差也就较大。但是，随着计算机处理能力的增强，通过大规模码本的应用，可以在一定程度上克服这个问题，同时充分利用 DHMM 运算复杂度较低的优势。

2. 连续 HMM（CHMM）

为了克服 DHMM 对特征空间描述上的不精确性，CHMM 应运而生。

CHMM 的主要目的是对特征空间进行比较精确的描述。

对一个概率密度进行估计的准确程度取决于训练数据量的多少。当训练数据量足够大时，利用类似于直方图统计的 Parzen 窗估计法[Parzen 62]可以估计得很精确，缺点是选择统计半径的大小比较困难、会浪费很大存储空间和对训练数据量的敏感性。

什么描述方法可以作到既能少占存储空间，又能降低估计复杂度呢？

一种好的方法是使用混合高斯密度(MGD, Mixture Gaussian Density)[Wilpon 89, Huang 89]，即

$$b_n(x) = \sum_{m=1}^{M} g_{nm} \cdot N(x; \mu_{nm}, \Sigma_{nm}) \tag{3.1.23}$$

其中

$$\sum_{m=1}^{M} g_{nm} = 1 \tag{3.1.24}$$

这是用 M 个混合高斯密度对第 n 个状态的特征空间进行估计。理论上可以证明，当 M 足够大时，MGD 可以比较准确地描述特征向量的概率密度。

3. 半连续 HMM（SCHMM）

虽然 MGD 描述方法中所要存储的参数不多(每个混合的中心向量 μ_{nm}、协方差矩阵 Σ_{nm} 和混合增益 g_{nm})，但当 M 很大时由于每个 $b_n(x)$ 都需要存储 M 组这样的参数，因此比较浪费空间。SCHMM[Huang 89]结合 VQ 技术和连续密度描述的特点比较好地解决了这个问题。

$$\begin{aligned} b_{s_t}(o_t) &= f\{o_t | s_t\} = \sum_{l=1}^{L} f\{o_t | V_l, s_t\} \cdot P\{V_l | s_t\} \\ &= \sum_{l=1}^{L} f\{o_t | V_l, s_t\} \cdot b_{s_t}^{(D)}\{V_l\} = \sum_{l=1}^{L} f\{o_t | V_l\} \cdot b_{s_t}^{(D)}\{V_l\} \end{aligned} \tag{3.1.25}$$

其中 $\{V_l | 1 \le l \le L\}$ 是表征特征空间的码本，$b_{s_t}^{(D)}\{V_l\}$ 是输出离散码字 V_l 的输出概率，$f\{\cdot | V_l\}$ 为以码字 V_l 为中心的子空间中的特征向量概率密度的高斯逼近。

这种描述方法改变一下形式成为

$$b_n(o_t) = \sum_{l=1}^{L} g_{nl} \cdot f\left\{o_t | V_l\right\} \tag{3.1.26}$$

这就是栓柱式 MGD(TMGD, Tied Mixture Gaussian Density)[Bellegarda 90]。

在这样的描述中，所有模型都公用 L 个类似于码字的密度函数，记录一个模型中不同状态的概率密度函数$b_n(x)$只需要一组系数$G = \left\{g_{nl} | 1 \le n \le N\right\}$即可。虽然 SCHMM 或 TMGD 对特征空间的描述节省了很大的存储空间，而且效果很好，但是由于所有模型的所有状态的特征空间的描述都依赖于这 L 个分布，因此其描述不如 MGD 来得精确，尤其在码本选得不合适时更是如此。

3.1.4 HMM 的局限

尽管 HMM 技术对现代语音识别作出了巨大的贡献，该技术本身仍存在一些固有的局限。这些局限性限制了经典 HMM 在语音识别系统中的实际应用，同时也是其他派生声学模型要重点解决的问题。HMM 的局限性主要表现在：

1. 观察序列中的特征矢量之间相互独立[Rabiner 89]的假设，即有下式：

$$P(O_1, O_2 \cdots O_T) = \prod_{j=1}^{T} P(O_j) \tag{3.1.27}$$

2. 马尔可夫链的假设 [Rabiner 89]。随着计算机处理能力的增强，可以采用高阶 HMM 等方法考虑系统当前状态与前 n 个时刻的状态有关（n>1），从而克服马尔可夫链对语音信号描述时过强的假设。

3. 末状态不可跨越。由于 A 矩阵的约束条件

$$\sum_{j=1}^{N} a_{ij} = 1 \tag{3.1.28}$$

导致末状态 N 不可跨越，这对于连续语音识别十分不利。柴海新[chai 94]提出的双层自动机网络模型可以较好地解决这一问题。

4. HMM 的时间和空间复杂度较高。

3.2 中心距离连续概率模型（CDCPM）

中心距离连续概率模型（CDCPM, Center-Distance Continuous Probabilistic Model）是从 HMM 衍变而来的一种新的概率统计模型，可以用于非特定人或特定人、孤立词或连续语音的识别[Zheng 96]。

一个 HMM 通常由状态转移概率矩阵 A、观察概率密度函数矩阵 B 和初始概

率分布矢量π表征。而 CDCPM 只保留 HMM 模型中的 B 矩阵，且 B 中的 PDF 被一个一维的 PDF 取代。这种取代可以在保证良好性能的前提下降低时空复杂度。

3.2.1 中心距离正态分布

N 个状态的从左向右连续 HMM 模型的第 n 个状态的观察输出 PDF 可采用下列形式的混合 Gauss 密度(MGD)：

$$b_n(\mathbf{x}) = \sum_{m=1}^{M} g_{nm} \mathrm{N} \quad (\mathbf{x}; \mu_{nm}, \Sigma_{nm}) \tag{3.2.1}$$

其中 $\mathrm{N} \quad (\mathbf{x}; \mu, \Sigma)$ 表示均值矢量为μ、协方差矩阵为Σ的 D 维正态分布密度函数，那么一个连续混合密度的 HMM 需要选择下列参数：状态数 N、混合密度数 M、特征矢量的维数 D、状态转移矩阵 A、混合成分的均值矢量 $\mu_{nm} = (\mu_d^{(nm)})$、混合成分的协方差矩阵 $\Sigma_{nm} = (\sigma_{pq}^{(nm)})$ 以及混合增益 g_{nm}。其中 $1 \le n \le N$, $1 \le m \le M$, $1 \le d,p,q \le D$。

现在考虑均值矢量为 $\mu = (\mu_1, \mu_2, \cdots, \mu_D)$、协方差矩阵为 $\Sigma = (\sigma_{pq})$ 的 D 维正态密度函数:

$$p(\mathbf{x}; \mu, \Sigma) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp(-\frac{1}{2}(\mathbf{x} - \mu) \Sigma^{-1} (\mathbf{x} - \mu)^T) \tag{3.2.2}$$

如果我们仅考虑Σ阵的对角部分,那么公式将变成(记 $\sigma_i^2 = \sigma_{ii}$):

$$p(\mathbf{x}; \mu, \Sigma) = \frac{1}{(2\pi)^{D/2} \prod_{d=1}^{D} \sigma_d} \exp(-\sum_{d=1}^{D} (x_d - \mu_d)^2 / 2\sigma_d^2) \tag{3.2.3}$$

同时定义两个矢量 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 之间的加权 Euclidean 距离为:

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{d=1}^{D} w_d (x_{1d} - x_{2d})^2} \tag{3.2.4}$$

改写公式(3.2.3)得到一个新的 PDF:

$$p(\mathbf{x}; \mu, \sigma_w) = \frac{2}{(2\pi)^{1/2} \sigma_w} \exp(-d^2(\mathbf{x}, \mu) / 2\sigma_w^2) \tag{3.2.5}$$

这是一个关于$d(\mathbf{x},\mu)$的而不是关于$\boldsymbol{x}$的 PDF。事实上，若一维的随机变量 ξ 服从正态分布 N(μ , σ)，记$\eta=|\xi-\mu|$，那么 η的 PDF 可以表示为：

$$p(y;\sigma)=\frac{2}{\sqrt{2\pi}\sigma}\exp(-y^2/2\sigma^2),y\geq 0 \tag{3.2.6}$$

或 $$p(x;\mu,\sigma)=\frac{2}{\sqrt{2\pi}\sigma}\exp(-d^2(x,\mu)/2\sigma^2) \tag{3.2.7}$$

(3.2.7)式和(3.2.5)式分别是一维和多维的情形，我们统一记为 $\mathrm{N}_{CD}(x;\mu,\sigma)$。由于这种分布描述的是服从正态分布的随机矢量与其均值矢量之间距离的分布，故称之为中心距离正态(Center-Distance Normal，CDN)分布。由(3.2.6)，η的均值为:

$$\mu_\eta=\int_0^\infty y\ p(y;\mu,\sigma)dy=\frac{2\sigma}{\sqrt{2\pi}} \tag{3.2.8}$$

于是 CDN 的参数μ和σ可以由下面的公式从训练数据库中估计出来:

$$\mu=\frac{1}{S}\sum_{n=1}^{S}x_n,\quad \sigma=\frac{\sqrt{2\pi}}{2}\mu_\eta=\frac{\sqrt{2\pi}}{2}\left[\frac{1}{S}\sum_{n=1}^{S}d(x_n,\mu)\right] \tag{3.2.9}$$

其中 S 是样本的数目，$x_1,x_2,\cdots,x_S$ 是 S 个训练样本。

3.2.2 CDCPM 训练及识别

每个识别单元需要一个 CDCPM。描述一个 CDCPM 需要下列参数：模型的状态数 N、混合密度数 M、特征矢量维数 D、混合成分 CDN 分布参数 $\mu_{nm}=(\mu_d^{(nm)})$ 和 σ_{nm} 以及混合的增益 g_{nm}。其中 $1\leq n\leq N$, $1\leq m\leq M$, $1\leq d\leq D$。第 n 状态的观察 PDF 有如下的形式:

$$b_n(\mathbf{x})=\sum_{m=1}^{M}g_{nm}\mathrm{N}_{CD}(\mathbf{x};\mu_{nm},\sigma_{nm}) \tag{3.2.10}$$

训练步骤为：

(1) 使用非线性分段(NLS)方法把训练数据库中的每个观察特征序列分为 N 段(对应于 N 个状态)；
(2) 把每个观察序列中第 n 段的特征矢量都集中起来，使用诸如 LBG 算法的聚

类算法分为 M 类；

(3) 利用公式(3.2.9)对每一个混合成分估计分布参数 μ_{nm} 和 σ_{nm}。

Bayes 学习算法可以用于训练 CDCPM。这里给出了公式(3.2.10)的另外一种形式，即基于最近邻(NN)准则的计分方法：

$$b_n(\mathbf{x}) = \max_{1\le m\le M} \mathrm{N}_{CD}(\mathbf{x};\mu_{nm},\sigma_{nm}) \tag{3.2.11}$$

实验结果是公式(3.2.11)比(3.2.10)的实用效果更好[Zheng 96]。在识别时对观察矢量序列得分的计算，有两种方法。

第一种，在识别前预分段。给定观察矢量序列 $\mathbf{O} = (o_1, o_{2,\cdots,} o_T)$，在算分前首先分段，然后按下面的公式计算它与某个 CDCPM 的匹配得分:

$$score(\mathbf{O};\mu,\sigma) = \prod_{t=1}^{T} b_n(o_t | o_t \in segment\ n) \tag{3.2.12}$$

其中 $b_n(\cdot)$ 可以由公式(3.2.10)或(3.2.11)计算。

这种算分策略很有用，并证明对孤立词识别是有效的。但在连续语音识别中不容易实现，因此就有了自动分段策略。

第二种，识别时自动分段。

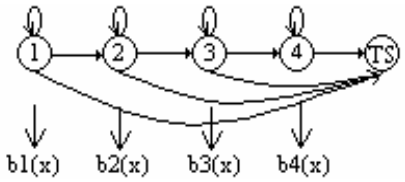

图 3.1　连续语音识别中 4 状态的 CDCPM 的表示

图 3.1 是一个用于连续语音识别的 4 状态从左向右 CDCPM，其中终结状态 TS 可以是一个静音态 SILENCE 也可以是一个垃圾态 GARBAGE。CDCPM 没有状态转移矩阵，它采用如下的方法控制状态转移：

在 t=1 时刻 CDCPM 总处在状态 1，即 state(1)=1。若 t 时刻 CDCPM 处在状态 i 时，即 state(t)=i, 则下一状态只能是状态 i、状态 i+1 或状态 TS，这取决于 $b_n(o_{t+1})$ 的值。这里有简单的决策准则:

$$state(t+1) = \arg\max_{n=i,i+1,TS} b_n(o_{t+1}) \tag{3.2.13}$$

在某种意义下这个决策准则工作得很好，但它忽略了状态驻留长度。公式(3.2.13)可以更准确地修改为:

$$state(t+1)=\begin{cases} i, & DUR^{(i)} < -TSH \\ \arg\max\limits_{n=i,i+1,TS} b_n(o_{t+1}), & -TSH \le DUR^{(i)} \le TSH \\ \arg\max\limits_{n=i+1,TS} b_n(o_{t+1}), & TSH < DUR^{(i)} \end{cases} \tag{3.2.14}$$

其中 $DUR^{(i)}$ 是状态 i 的(经过驻留长度的统计均值和方差)标准化了的驻留长度, TSH 是标准化的驻留长度阈值，一般地取 TSH=1 或 1.5.

3.3 高斯混合分段模型（GMSM）的研究与实现

针对经典 HMM 时空复杂度高，尤其是 Viterbi 解码过程的复杂性，本文给出一个基于高斯混合分布和观察序列分段的声学模型，称为高斯混合分段模型（GMSM，Gussian Mixture Segmentation Model）。该模型采用高斯混合分布描述观察矢量在特征空间的分布，同时对 HMM 的训练和 Viterbi 解码过程进行简化，采用对观察序列进行线性或非线性分段代替 HMM 中的状态，用帧同步搜索代替状态解码。

经过这样的简化，GMSM 与同样采用高斯混合分布的 CHMM 相比，识别速度有较大提高，而正识率与 CHMM 不相上下，详细结果请参阅本文第六章的内容。

3.3.1 高斯混合分布及其参数估计

设 K 段从左至右高斯混合分段模型（GMSM）中第 k 段的观察输出特征矢量集 $Z_k=\{\vec{z}_1,\vec{z}_2,\cdots,\vec{z}_T\}$（在本小节中将 Z_k 简记为 Z），我们使用如下的概率密度函数描述特征矢量 $\vec{z}_j$ 的分布:

$$p(\vec{z}_j|\theta)=\sum_{i=1}^{M}\left(p_i\,p(\vec{z}_j|\theta_i)\right) \tag{3.3.1}$$

其中：● $\vec{z}_j \in Z$ 是 d 维特征矢量，j＝1,2,$\cdots$,T。

● M 是高斯混合分布的阶，即高斯混合概率密度函数中单个高斯分布的个数。单个高斯分布的概率密度函数为：

$$p(\vec{z}_j|\theta_i)=(2\pi)^{-d/2}|R_i|^{-1/2}\exp\left(-\frac{1}{2}(\vec{z}_j-\vec{\mu}_i)^T R_i^{-1}(\vec{z}_j-\vec{\mu}_i)\right) \tag{3.3.2}$$

其参数θ_i包括均值矢量$\vec{\mu}_i$及协方差矩阵R_i。

- p_i是第 i 个高斯分布的权重，满足$\sum_{j=1}^{M} p_i = 1, p_i \geq 0.$
- θ是高斯混合模型的参数，它包括p_i和θ_i， i＝1,2,…, M。

给定集合 Z，使用高斯混合模型描述频谱矢量的分布，需确定模型参数θ，使得在该参数下矢量集 Z 发生概率$p(Z|\theta)$最大。Dempster 和 Laird 给出了一种 EM 算法[Dempster 77]求解一般高斯混合分布的参数θ。本文给出该算法在语音声学建模时的简便形式。

设 Z 中矢量互相独立，有

$$p(Z|\theta) = \prod_{j=1}^{T} p(\vec{z}_j|\theta) \tag{3.3.3}$$

设$p(Z|\theta)$对θ可微，当$p(Z|\theta)$取极大值时满足：

$$f = \nabla_\theta\left(\ln p(Z|\theta)\right) = 0 \tag{3.3.4}$$

解上式，即可得到估计$\vec{\mu}_i$，R_i和$p(\omega_i)$的迭代算法（i＝1,2,…, M）。以$\vec{\mu}_i$为例：

$$\begin{aligned} f &= \nabla_{\mu_i}\left(\ln p(Z|\theta)\right) \\ &= \sum_{j=1}^{T} p(\vec{z}_j|\theta)^{-1}\nabla_{\mu_i}\left(\sum_{i=1}^{M} p_i p(\vec{z}_j|\theta_i)\right) \\ &= \sum_{j=1}^{T} p(\vec{z}_j|\theta)^{-1} p_i \nabla_{\mu_i} p(\vec{z}_j|\theta_i) \\ &= \sum_{j=1}^{T} p(\vec{z}_j|\theta)^{-1} p_i p(\vec{z}_j|\theta_i) R_i^{-1}(\vec{z}_j - \vec{\mu}_i) \\ &= 0 \end{aligned} \tag{3.3.5}$$

令

$$P_{i,j} = p(\vec{z}_j|\theta)^{-1} p_i p(\vec{z}_j|\theta_i) \tag{3.3.6}$$

于是得到

$$\vec{\mu}_i = \sum_{j=1}^{T}\left(P_{i,j}\cdot\vec{z}_j\right)\Big/\sum_{j=1}^{T} P_{i,j}\ . \tag{3.3.7}$$

将等式左侧的 $\vec{\mu}_i$ 改写为 $\hat{\vec{\mu}}_i$，即得到估计 $\vec{\mu}_i$ 的估值 $\hat{\vec{\mu}}_i$ 的迭代算法。

类似地，可以得到 p_i 和 R_i 的估值 $\hat{p}_i$ 和 $\hat{R}_i$ 的迭代算法。

$$\hat{\vec{\mu}}_i = \sum_{j=1}^{T}\left(P_{i,j}\cdot\vec{z}_j\right)\Big/\sum_{j=1}^{T}P_{i,j} \tag{3.3.8}$$

$$\hat{R}_i = \sum_{j=1}^{T}\left(P_{i,j}(\vec{z}_j-\vec{\mu}_i)(\vec{z}_j-\vec{\mu}_i)^T\right)\Big/\sum_{j=1}^{T}P_{i,j} \tag{3.3.9}$$

$$\hat{p}_i = T^{-1}\cdot\sum_{j=1}^{T}P_{i,j} \tag{3.3.10}$$

需要注意的是，在推导(3.2.10)时，有约束条件 $\sum_{j=1}^{M}p_i=1, p_i\geq 0.$ 即要在满足此条件的同时进行确定(3.2.3)式的极值。

3.3.2 观察特征矢量序列的分段

GMSM 对观察特征矢量序列的分段主要用于训练模型时确定用高斯混合分布描述的观察输出特征矢量集 Z，该分段问题与 CDCPM 训练时的分段问题类似。

设某个需要建模的识别基元的学习样本集合为 $\{O^{(m)}, m=1\sim M\}$，其中每个特征矢量序列 $O^{(m)}$ 构成一个学习样本，共有 M 个样本（本小节使用符号 M 表示总学习样本数，区别于 3.3.1 节中的 M）。则需要解决的问题是给出一种特征矢量序列的分段准则，将每个样本序列 $O^{(m)}=O_1^{(m)},O_2^{(m)}\cdots O_{T_m}^{(m)}$ 划分为 K 个子集 $O_{(k)}^{(m)}=\{O_{s_k}^{(m)},O_{s_k+1}^{(m)},\cdots O_{s_{k+1}-1}^{(m)}\}$，其中 $1\leq k\leq K+1$，$1\leq S_k\leq T_m+1$，$S_k<S_{k+1}$。这样，就可以确定用高斯混合分布描述的第 k 段观察输出特征矢量集 Z_k：

$$Z_k=\{O_{(k)}^{(m)}, m=1\sim M\} \tag{3.3.11}$$

分段准则主要有以下几个方面：

1. 线性分段

线性分段的准则比较简单，即将每个样本序列的T_m个观察特征矢量按照矢量的个数平均分为 K 段，每段有T_m/K 个矢量。这种分段方法对于识别基元比较小的情况（例如以汉语音节或半音节作为识别基元）非常适合，因为它规则简单，运算量小，且在识别基元小时能够保证各段特征矢量集大小比较均匀，不会出现某个矢量集的矢量数过少的情况，利于用高斯混合概率密度函数来描述。

2. 非线性分段

非线性分段要求首先定义两个观察特征矢量O_i和O_j之间的距离$d(O_i,O_j)$。常用的距离度量有欧氏距离，加权欧氏距离等。令某个学习样本$O^{(m)}$的特征矢量序列累积距离

$$D^{(m)}=\sum_{i=1}^{T_m-1}d(O_i^{(m)},O_{i+1}^{(m)}) \tag{3.3.12}$$

非线性分段以$D^{(m)}/K$为依据将此序列分为 K 段，使得每段包含的特征序列累积距离$D_{(k)}^{(m)}=\sum_{i=S_k}^{S_{k+1}-2}d(O_i^{(m)},O_{i+1}^{(m)})$均等于$D^{(m)}/K$。这样的分段方法适于识别基元较大的情况（例如在孤立词识别中以整词为识别基元），这种情况下每段的特征矢量数较多，可以用概率模型比较充分地描述。同时，非线性分段保证了每一段内的特征矢量变化比较均匀，符合概率模型在同一段内忽略时序特征的初衷。

3. 对分段数 K 的调整

以上的两种分段方法都是根据预先确定的分段数 K 对学习样本的观察特征矢量序列进行分段。然而，对于不同的待建模识别基元，其特征矢量的序列的长度，变化趋势等都不尽相同。因此，为了体现这种差异，可以在线性分段或非线性分段的基础上对每个待建模识别基元的分段总数 K 进行调整。调整的原则是使某个待建模识别基元的第 k 个观察输出特征矢量集Z_k（3.3.11 式）与第 k＋1 个观察输出特征矢量集Z_{k+1}之间有足够的差异。

设第 k 个观察输出特征矢量集Z_k有L_k个观察矢量。定义Z_k内部离散程度的度量：

$$G_k=\sum_{i=1}^{L_k}d(O_i^{(k)},\overline{O}^{(k)}) \tag{3.3.13}$$

其中，$O_i^{(k)} \in Z_k$ 为 Z_k 中的矢量，$\overline{O}^{(k)}$ 为 Z_k 中矢量的均值矢量。

定义 Z_k 与 Z_{k+1} 之间的离散程度的度量：

$$F_k = \sum_{i=1}^{L_k} d(O_i^{(k)}, \overline{O}^{(k+1)}) \tag{3.3.14}$$

给定合并阈值 Δ，要求 $G_k / F_k > \Delta$，否则将 Z_k 与 Z_{k+1} 合并为一段。这样，对于声学特性比较稳定的识别基元，就可以根据其自身的特点采用较少的段数进行描述，进一步减小识别的计算复杂度，同时增大了段内的矢量数，对于训练模型也有益处。

3.3.3 GMSM 的识别

使用 GMSM 对识别基元进行识别时，使用帧同步算法[Lee, 89]搜索特征序列的最佳分段。设识别器共有 B 个识别基元的模型，则识别结果由下式的最大后验概率确定：

$$\mathrm{b} = \underset{1 \le \mathrm{b} \le \mathrm{B}}{\operatorname{argmax}} \ \mathrm{p}(\mathrm{O} \mid \lambda_{\mathrm{b}}) \tag{3.3.15}$$

其中，$\mathrm{p}(\mathrm{O} \mid \lambda_{\mathrm{b}})$ 为待识别观察特征矢量序列 O 在第 b 个识别基元的模型下帧同步搜索给出的发生概率值。

一个典型的帧同步搜索包括以下的几个问题：

- 计算当前特征矢量在当前段的概率得分
- 确定当前特征矢量在本段停留或转移到下段
- 根据识别基元的有限状态网络进行路径搜索及合并
- 剪枝策略
- 回溯求得最优路径

每段驻留长度的控制可以采用类似于 CDCPM 中(3.2.14)式的方法。随着特征序列时间指标 t 的增加，从左至右搜索的帧同步算法会成指数规模扩展出新的路径，保留全部的路径然后求出最优解往往是不可能的。因此，就需要制定相应的剪枝规则。实际搜索时，可以采用两种方法控制搜索规模：

1. 设定得分阈值

选定一个常数 $\alpha (0 < \alpha < 1)$，在当前时刻 t，在所有可以扩展的路径中选择得

分值最大的一条，设其得分为 S_{Max}，则只保留得分在 S_{Max} 和 αS_{Max} 之间的路径，其余路径被放弃。

2. 设定保留路径数阈值

选定一个保留路径数的上限值 β。当扩展出的路径数超过此阈值时,按路径得分的大小顺序仅保留 β 条路径。

3.3.4 GMSM 实现的一些具体问题

1. 方差矩阵估计的简化

确定了某个待建模识别基元的第 k 段观察输出特征矢量集 Z_k（3.3.11 式）后，就需要使用(3.3.8)到(3.3.10)的迭代公式估计混合高斯模型的参数。这些参数包括：

均值矢量集 $\hat{\vec{\mu}}_i$ (i=1,2,⋯M，M 为高斯混合分布的阶)，方差矩阵集 $\hat{R}_i$ (i=1,2,⋯M)和单高斯分布的权重集 $\hat{p}_i$ (i=1,2,⋯M)。注意，我们在推导高斯混合分布的参数估计时，并未假定特征矢量各维之间的独立性，即并不要求方差矩阵 $\hat{R}_i$ 为对角阵。然而，在实际应用时，由于计算复杂性的限制，矩阵 $\hat{R}_i$ 将在特征矢量各维独立的假设下用对角阵代替。此时(3.3.9)式变为如下的形式：

$$\hat{\vec{R}}_i = \sum_{j=1}^{T}\left(P_{i,j}(\vec{z}_j - \vec{\mu}_i)^{@}\right) \Big/ \sum_{j=1}^{T} P_{i,j} \tag{3.3.16}$$

其中符号 $(\vec{z}_j - \vec{\mu}_i)^{@}$ 表示矢量 $(\vec{z}_j - \vec{\mu}_i)$ 各维自身平方后构成的矢量。

2. 迭代初值的选取

一般来说，迭代算法在整个初值的可取值范围内都保证收敛是很困难的，同时，不同的初值选取对于迭代过程的收敛速度和最终收敛到的局部极值点都有很大的影响。因此，对于迭代算法来说，迭代初值的选取是十分重要的。在 GMSM 的训练过程中，我们采用下列的初值确定方法：

首先对某个待建模识别基元的第 k 段观察输出特征矢量集 Z_k（3.3.11 式）进行 LBG 聚类[Linde 80]，将其划分为 M 个子集 $Z_i^{(k)}(i = 1,2,\cdots M)$（M 即为高斯混合

分布的阶)。然后分别确定初值：

- $\hat{\vec{\mu}}_i$ 的初值设为相应子集 $Z_i^{(k)}$ 所包含矢量的均值。
- $\hat{\vec{R}}_i$ 的初值有两种选取原则：**1.** 各维均采用 LBG 给出的类内离散度度量，即方差矢量各维初值相同。**2.** 对所有训练数据的各维方差进行统计，得到一个各维的加权系数，然后对第一种初值进行加权，即方差矢量各维初值不同。后者从理论上说收敛性要好一些，但由于不同初值可能导致最终收敛到不同的局部极值，这两种方法我们都进行了实验，详见第六章实验结果。
- $\hat{p}_i$ 的初值亦有两种方法：**1.** 令 $\hat{p}_i = 1/M$ 的等权重初始化。**2.**令 $\hat{p}_i$ 正比于 LBG 聚类结果中每一类的矢量数，同时满足 $\sum_{j=1}^{M} p_i = 1, p_i \geq 0.$ 实验证明两种方法对模型性能影响不大，后者收敛速度更快。

3. 训练数据不足和奇异数据的解决

在实际的模型训练过程中，发现由于某些识别基元的学习样本集过小，训练数据不足，LBG 给出空类或单矢量类，导致高斯混合分布参数估计迭代过程失效。另外，由于 LBG 结果的某些类内包含个别奇异特征矢量，这些矢量与本类的中心矢量间的距离很大，导致迭代算法在针对这些奇异矢量进行参数估计的迭代调整时产生下溢。针对这些问题，训练中采用下面的算法进行控制：

```
bIterationSuccess = false;
while ((M>=1) && (!bIterationSuccess))
{
    ActualM = LBGCluster(M); //LBG 聚类，共 M 类。返回剔除空类的类数
    InitGMSMParameter(ActualM);   //根据 LBG 结果初始化 GMSM 参数
    {                             //GMSM 迭代过程
`       …
        bIterationSuccess = true;
        if (( nDiscardedFaultyVec / nTotalVec ) > CONST_THRE)
          bIterationSuccess = false;   //奇异矢量被抛弃。对抛弃的矢量计
                                       //数，若超过一定比例认为迭代过程
                                       //失败
        …
    }
    if (!bIterationSuccess) M /= 2;    //若迭代失败，减少 GMSM 的阶，相
                                       //应增加训练量
}
```

该算法实际上是使得高斯混合分布的阶 M 对训练矢量集的大小作出自适应，以比较合理的模型进行描述。这样，对于不同的识别基元模型以及同一识别基元内部的不同分段，高斯混合模型的阶均可能不同。

4. 迭代终点的确定

确定迭代终点可以采用绝对阈值或相对阈值。

设 GMSM 的参数$\xi = \{\hat{\vec{\mu}}_i, \hat{\vec{R}}_i, \hat{p}_i \mid i = 1,2,\cdots M\}$，定义两组 GMSM 参数$\xi_u$和$\xi_v$之间的差异度量$\omega_{u,v}$：

$$\omega_{u,v} = \sum_{i=1}^{M}\left(\sum_{d=1}^{D}\left(\left|\hat{\vec{\mu}}_i^{(u)}(d) - \hat{\vec{\mu}}_i^{(v)}(d)\right| + \left|\hat{\vec{R}}_i^{(u)}(d) - \hat{\vec{R}}_i^{(v)}(d)\right|\right)\bullet w_d + \left|\hat{p}_i^{(u)} - \hat{p}_i^{(v)}\right|\right) \quad (3.3.17)$$

其中，D 为参数矢量的维，w_d为第 d 维的加权值，它是所有参加训练的特征矢量的第 d 维分量方差的倒数。$\hat{\vec{\mu}}_i^{(u)}(d)$和$\hat{\vec{R}}_i^{(u)}(d)$分别表示第 u 组 GMSM 参数均值和方差矢量的第 d 维分量。

- 绝对阈值

  给定阈值ω，当本次迭代训练的结果ξ_u和上次迭代训练的结果（或 GMSM 参数初值）ξ_{u-1}间的差异$\omega_{u,u-1} < \omega$时，停止迭代。

- 相对阈值

  给定阈值η（$\eta \in (0,1)$，例如η＝0.01），设首次迭代训练结果ξ_1与 GMSM 参数初值ξ_0间的差异度量为$\omega_{1,0}$。当本次迭代训练的结果ξ_u和上次迭代训练的结果ξ_{u-1}间的差异$\omega_{u,u-1} < \eta \bullet \omega_{1,0}$时，停止迭代。

相对而言，采用相对阈值比较合理。但根据实际实验，由于各个识别基元收敛情况比较一致，两种方法结果无明显差异。

5. 提高识别速度的处理

GMSM 与采用(3.2.11)式 NN 计分方法的 CDCPM 相比，在识别速度上有一定损失，主要是在使用(3.3.1)式取对数的形式

$$\ln p(\vec{z}_j \mid \theta) = \ln\left(\sum_{i=1}^{M}\left(p_i p(\vec{z}_j \mid \theta_i)\right)\right) \tag{3.3.18}$$

来计算特征矢量在混合高斯分布下的对数得分时，无法预先存储方差参数矢量 $\vec{\hat{R}}_i$ 各维的对数值来避免在识别过程的对数运算。为进一步提高识别速度，在假设混合高斯分布中各个单高斯分布差异较大的前提下，将(3.3.18)式简化为：

$$\ln p(\vec{z}_j \mid \theta) \approx \max_{1\le i\le M}\left(\ln\left(p_i p(\vec{z}_j \mid \theta_i)\right)\right) \tag{3.3.19}$$

这样，可以预先存储对数运算的结果，从而提高识别的速度。采用(3.3.18)式和(3.3.19)式计分的 GMSM 的性能对比实验详见第六章的结果。

第四章 语料处理与语言模型数据库

4.1 语料处理

针对汉语的特点，汉语语言的基础加工与收集过程主要包括：语词的定义与切分，基本词频信息统计和词语分类[Zhang 97]。由于我们的语言模型很大程度上取决于统计的结果，因而影响其效果的主要因素有：语料的来源、词表的选择、语料库的标注、语言知识在统计中的应用。下面我们分别对它们一一加以描述。

第一、语料的来源。汉语的语法规则千变万化，在不同的领域内词的搭配也是各有不同，为了反映出汉语语言的变化规律，我们的语料库就应该涵盖较大的范围，加以区分，然后视不同的应用领域进行利用，产生合适的语言模型。

第二、词表的选择。由于听写系统的语言模型调整的基础是词，因而词表的选择相当关键，如何选择使其能够基本上满足一般的使用要求，同时又不过分庞大以致于浪费大量的搜索时间、空间。进行语料处理首先要确定词表，以便对语料进行分词。

第三、语料库的标注。在建立了比较合适的词表之后，就需要对语料库进行标注工作。语料的标注工作不同于语音的标注工作，它的规律性更弱，不确定性也更强。在标注的过程中，不但要准确地标出词的分界，而且还应标出词的属性，如词性、词义等，其目的是为了将来能用来产生更高阶的语言和语义模型。标注完成的方式也可分为手工标注和自动标注。手工标注的工作同语音数据库的标注一样，需要较大量的人力完成，而自动标注，目前也处于研究阶段。

第四、语言知识在统计中的应用。许多语言工作者在长期的研究中，发现了一些汉语语言所特有的规律，这些规律不光可以用来指导语言模型的生成，还可以用来指导我们的语料库的建立工作，如降低词表的冗余度，提高语料库标注的准确性等。

在以上的几个方面中，分词是汉语语言模型中的一个特殊问题。在汉语中，词的界限不象西方语言那样明显，因此分词就成为语言模型处理的一个首要问题。语言模型的分词技术的主要目的是确定语言模型处理的最小语言单元边界。由于汉语语言较高的灵活性，“ 词 ”这一概念并没有一个十分严格的标准。著名语言学家朱得熙先生从词的组织结构上考虑词与非词的界限将词定义为 :最小的能独立活动的有意义的语言成分[Zhu 82]。归纳起来，“ 词 ”的定义主要有三个关键因素：语言的最小单位；固定的发音形式；一定的意义。但是，这样的描述性定义在是不能够简单用于实际的词切分模型的，三个因素有矛盾或重叠之处。因此，必须建立计算机语言处理中“ 词 ”的具体标准。

进行计算机自动分词的方法主要步骤是：

- 最大匹配：根据基础词表进行正向、反向或双向扫描进行最大匹配，给出各种可能的切分序列。
- 基于统计的切分调整：利用基础词表的统计信息，例如 Uni-gram，Bi-gram 或 Tri-gram 等模型对各种匹配结果进行评价和调整。
- 基于词法知识的切分调整：利用一定的词法知识对各种匹配结果进行评价和调整。同基于知识的语言模型一样，这种方法也具有知识的形式化表示不完全的困难。

汉语听写机中分词包括两个方面的内容，即大规模训练语料的分词和识别过程中对声学模型给出的候选音节序列进行分词解码。对于后者的分词过程，可以把基于统计的切分调整与基于词法知识的切分调整与相应语言模型的处理统一起来进行。

以上讨论的方法都需要一个基础词典进行分词的处理。目前有一种无词典指导的语料加工与语言信息自动处理方法也值得注意[Nagao 96]。该方法通过对语料中共同前缀串的搜索和计数来统计 n 元字符的 n-gram 同现频度信息。借助这些信息可以从中定义词语或短语、抽取词频和进行语料中长字符序列信息的统计和处理。与基于基础词典的方法相比，这种方法克服了基础词典对于分词处理的限制，同时简化了依据基础词典分词后再进行统计的步骤，具有一定的特色。但是，这种方法一定程度上抹杀了词的自然概念，代之以同前缀字符串，为后续基于知识的语言模型的处理带来了困难，因此需要进一步探索。

4.2 语言模型数据库

语言模型数据库（Language Model Database）记录了汉语的文字表述信息，目的是提供语言模型建模的必要数据。本节描述实际语言模型建模时采用的数据库结构。

4.2.1 整体结构

语言模型的数据库整体上采用基本词表与扩展属性表集结合的层次结构，在基本词表中放置有关词汇的最基本信息，同时给出对应的扩展属性表标志。

1. 基本词表

基本词表主要供基于统计的语言模型使用。确定基本词表词条首先由人工给出初始词表，然后根据训练语料库和分词评价函数由计算机调整确定。基本词表同时包括对应扩展属性表集的标志字。

2. 扩展属性表集

扩展属性表集由一系列的专门属性表构成，主要供基于规则的语言模型使用。例如，动词属性表，专有名词属性表，其他词性词语属性表，标点符号属性表，其他语法单位属性表等。这些扩展属性表提供基于规则的语言模型所需要的基本语言知识，是对基本词表信息的扩充。扩展属性表集可以根据需要逐步添加完善。

4.2.2 具体表项结构

1. 基本词表

基本词表主要包括：

- 词号 (1 DWORD)
- 词中音节总数 (1 WORD)
- 词形 (10 WORDs max) (排序键)
- 音节编号串 (10 WORDs max)
- 音调编号串 (10 BYTEs max)
- 分类（自动聚类）类别标识 (DWORD)
- 扩展属性表标志字 (DWORD)

2. 扩展属性表集

扩展属性表集中各属性表的结构根据自身特点有所不同。属性表中包含的信息内容主要由基于规则的语言模型中规则所需信息确定。其中一部分属性表可以考虑以合作的形式由第三方提供。

2-1 动词属性表

动词属性表主要包括：

- 词形
- 基本词表词号
- 动词类型
- 论元数目
- 义项数目
- 义项序号
- 释义
- 各论旨属性

专有名词属性表主要包括：

- 词形
- 基本词表词号
- 专有名词可能连接的前缀
- 专有名词可能连接的动词

2-2 专有名词属性表

2-3 其他词性词语属性表

其他词性词语属性表主要包括：

- 词形
- 基本词表词号
- 词性
- 前缀
- 后缀

2-4 标点符号属性表

标点符号属性表主要包括：

- 词形
- 基本词表词号
- 在此标点之前成对使用的标点
  （例如：左双引号+右双引号）
- 在此标点之后成对使用的标点
- 可在此标点前连用的标点表
  （例如：冒号+左双引号）
- 可在此标点后连用的标点表
- 可在此标点前连用的词的词性
- 可在此标点后连用的词的词性

2-5 其他语法单位属性表

其他语法单位（语义块）属性表主要包括：

- 词形
- 基本词数
- 基本词表词号串
- 语法单位分类号
  （例如：动词短语等）
- 前缀
- 后缀

第五章 语言模型研究与实现

对于听写机系统来说，其输入为声学信息，最终的输出是具有明确意义的文本。它与进行孤立词识别的语音识别系统的重要差别之一就在于输入听写机的待识别声学信息单元之间具有明确的连接关系。这种关系主要是非声学的信息，如语法、语义、语用等方面。在很大程度上，人正是有效地利用了这些高层次的知识才获得了对语音的正确的识别和理解，而单纯依靠孤立的发音往往是难以进行有效交流的。因此，听写机系统中必然要有用来描述和处理这种关系的模型。这种基于非声学的高层知识的模型，被统称为语言模型。

本章概述语言模型的特点、类别，同时提出一种基于修正退化频度估计算法的 n-gram 语言模型，讨论了其实现的具体问题。

5.1 语言模型所包含的具体知识

语言模型作为一种高层次的模型，它应当包含的具体知识主要有：

- **语法（syntax）**：确定一个词序列是否合乎规则，是否是一个合法的句子或一个合法句子的一部分。
- **语义（Semantics）**：确定一个句子是否有意义。
- **语用（Pragmatics）**：根据背景知识和使用环境的了解，确定句子在其上下文或语境中是否合适。
- **韵律（Prosodics）**：根据句子的声调、语调，以及节奏速度等来对句子的句法或语义等作出某种判断。

计算机建立语言模型的主要任务就是实现以上的各种知识的形式化表示，从而使计算机能够利用这些知识进行高层次的语言处理。限于目前的条件和水平，对于语言模型的研究和实现主要集中在语法这一层次，同时可以考虑初步拓展到语义的范畴；而对于语用和韵律知识的表示和实际应用尚十分困难。

5.2 汉语语言模型的特点与难点

从信息处理和语言知识的角度分析，汉语与西方语言相比具有显著的不同，西方语言主要具有如下特点：

- 西方语言的语言单元即词之间具有明显的空格分界，词的定义十分明确。而汉语的语言单元之间没有明确的分界。
- 西方语言的同音词较少，而汉语的发音只有约 400 个无调音节或 1200 个有调

音节，所以同音字、同音词比较多。

- 西方语言的明确的语法变化标志较多，如单复数、时态等等。汉语没有这些明确的标志。
- 西方语言具有较为固定的语法、语义表示形式。而汉语只有一小部分书面语言能够借用西方语法规则对其进行形式化表示。绝大部分语言模式目前还难以找到结构化的表示方式。
- 汉语一词多意的现象更加严重。

以上的特点也就决定了汉语语言模型在研究和实现上有自己的难点：

- 汉语语言模型中分词的特殊问题。汉语词切分成为汉语理解与处理首要的问题。
- 大量的同音字、同音词给语言模型的处理增加了负担。
- 十分灵活和自由的语言表述使得寻找形式化的语言规则来描述汉语语法、语义限制比较困难。

因此，我们在研究中不能单纯地全套照搬西方语言模型研究的途径，而是应该在借鉴的基础上寻找有效的汉语语言模型的建立方法。

5.3 语言模型的类别

根据对高层次的语言信息描述和处理方法的不同，语言模型主要分为两类：即基于统计的语言模型和基于知识（规则）的语言模型[Li 95]。

5.3.1 基于统计的语言模型

基于统计的语言模型是从大规模的真实语料库中获得语言单元之间的统计关系，从而为某一个特定的语言单元序列提供一个对应的评价值，即提供该序列在统计意义下出现的概率估计。统计模型的基本思想是：大规模的语料库中隐含了各种先验知识，语言模型的任务是由计算机自动或半自动地从大规模的语料库中发现和学习这种知识，并客观地对语料库中的语言事实作出表述。这种基于统计的语言模型使用的基本手段是发展比较成熟的概率论和统计学，是当前自然语言建模与处理的主流技术。

5.3.2 基于知识的语言模型

基于知识的语言模型的核心是利用人们对某种语言的语法、语义等方面规律的先验的总结来进行语言的分析与处理。一般来说，先验知识的表示途径包括基

于文法的规则表示，基于语义关系的语义网络或神经元网络，以及结合了概率信息的随机文法规则等。

然而，无论是怎样的先验知识表达方式，现阶段都不可能把理解自然语言所需要的全部知识或规则枚举出来，尤其是对语法、语义、语用规则十分灵活的汉语更是如此。知识在数量上是不可穷尽的，在表达上有具有高度不确定性和随意性。因此，现阶段基于知识的语言模型的研究目的主要是希望它能够在一定的语言范畴内对于基本的统计模型进行一定程度上的补充和完善，以获得系统性能最大程度上的提高。

5.3.3 统计语言模型与知识语言模型的比较

经典的统计语言模型给出的结果是一个语言单元序列出现的概率，而经典的知识语言模型的结果是作出某个语言单元序列是否合理的二元判断。经典的基于统计的语言模型与经典的基于知识的语言模型相比，具有这样一些优点：

- 更加直观、松弛的语言信息处理模式。在语言模型的实际使用中，给出一个最可能的结果往往比给出一个“对”或者“错”的判断更加有用。
- 相对简单有效的信息获取方式。统计语言模型的基础是大规模的语料库，其信息的获取方式是基于真实语料的无监督的学习过程。这就绕过了语法知识的人工编辑、整理与归纳总结的过程。
- 对不同语言领域的适应性强。基于统计的语言模型在语言领域发生变化时，对新的语言领域可采用相对固定的训练方法获取知识，而基于先验知识的语言模型则可能需要进行大的规则体系改动才能适应。

同时，统计模型也存在以下的不足：

- 统计模型并没有真正从语法、语义、语用等理解自然语言的本质角度进行知识的刻画，只是从大规模的语料库中提取统计信息，从而不可避免地会出现一些完全不合自然语言规律的错误。
- 统计模型需要十分庞大的训练语料。而事实上准备满足统计模型语料规模要求的语料库是不可能实现的，这就带来了有关数据平滑的各种问题
- 目前可以实现的统计模型存在着只能描述近邻语言单元之间关系的严重不足。事实上，在自然语言中具有大量的远邻语言单元间的关系信息，例如各种成对使用的连词，一些常用语等。这些关系没用得到有效的描述。

因此，单靠其中任何一种语言模型都是不能十分有效地描述和处理高层次的语言知识的。如何将两者结合起来，根据现有条件在有所侧重的基础上作到两者相辅相成，是一个对于理论研究和实践工作都意义重大的课题。此外，近阶段发

展起来的一些结合了概率统计信息的知识语言模型也可以给出一个语言单元序列符合先验语法、语义等规则的合理性概率，在一定程度上向统计语言模型靠拢，这样的模型也是较有潜力的。

5.4 几种典型的语言模型简述

5.4.1 n-Gram 语言模型

n-Gram 语言模型是具有代表性的语言模型。其最大的优点是求解方法简单有效，能通过对大量训练语料的统计计算出语言单元之间的连接概率，从而给出整 个语言单元序列的出现概率。然而，真正使 n-Gram 模型实用化，必须解决 n 值的选取与训练语料稀疏的问题。理论上说，n 值越大，n-Gram 模型对语言单元序列出现概率的估计越准确。但随着 n 值的增大，n-Gram 模型的参数空间将成指数增长（V^n,V 为词表大小），所需的训练语料亦将是呈指数增长。这样，会导致实际训练过程中训练语料的严重稀疏，不足以反映有意义的统计规律。此时，必须采取措施解决数据稀疏的问题。解决此问题的主要途径是：

- 基于频度折扣的数据平滑技术
- 基于等价类的数据平滑技术
- 基于词语相似性映射的数据平滑技术

5.4.2 概率上下文无关文法 （SCFG）

SCFG 是一种语法知识的语言模型，其基本内容是一个引入了统计机制的形式语言 G。之所以在形式语言的严格语法规则中引入统计机制，主要是希望对真实语法知识不确定性进行初步适应。SCFG 的文法 G 可以表示为一个四元式<T,N,S,R>，其中：

T：终结符集。既语言处理中的语法单元。

N：非终结符集。代表一定的语法概念或语法单元的组合。

S：开始符号。即语言处理中最终感兴趣的语法范畴，通常为一个“ 句子 ”。

R：产生式或规则。在 SCFG 中，产生式的形式为<A→α,p>，p 为应用此产生式进行推演的概率。它是 SCFG 中引入统计机制的标志。

SCFG 对观察语言单元单元进行解码的过程可以看作状态为短语的 HMM 解码过程。与 HMM 的前向、后向概率相类似，SCFG 引入了 inner 和 outer 概率进行解码运算。

对 SCFG 的实际应用可以与 n-Gram 进行结合，实现一个统一了词连接信息与局部特殊语法规则的语言模型。基于大规模语料统计得到的词连接概率反映语序组织的一般规律，SCFG 作为短语或局部句型的描述来约束局部的语序组织，两者相互补充，可以获得高鲁棒性的语序自组织求解能力。

5.4.3 基于短语句型的规则模型

基于短语句型的规则模型是一种比较严格的规则模型。它的基础是汉语语言学对于一些汉语常见句型的明确的规则化表示。前面已经提到，这种规则很难穷尽汉语语言的全部规则信息，然而，在一定范畴内确定的语言规则对于语言模型的整体性能的提高是很有帮助的。

因为基于短语句型的规则模型难以完全描述真实语言中的全部语法知识，它的主要应用是作为统计语言模型的补充。应用此类语言模型需要解决下列主要问题：

- 句型的形式化表示
- 句型匹配规则
- 根据已有的语言单元序列对即将出现的语言单元进行预测

目前，进行汉语语言学研究的学者在汉语句型的形式化表示方面已经进行了大量的工作，对汉语常见句型的描述有不少进展。但是，把这些研究成果应用于听写机中的语言模型还十分少见。所以，进行这方面的研究有较大挑战性同时也很有意义。

5.5 修正退化频度估计算法的 n-gram 语言模型研究与实现

5.5.1 图灵估计

设 n_r 为统计语料中恰好出现了 r 次的不同词串的数目。N 为语料的总词串数目。则有：

$$N = \sum_r r n_r \tag{5.5.1}$$

根据最大似然估计，设词串（m 元组） $w_1^m = \{w_1, w_2, \cdots, w_m\}$ 在训练语料中出现了 $C(w_1^m) = r$ 次，则对该词串出现概率的估计为：

$$P_{ML} = r / N \tag{5.5.2}$$

且满足：

$$\sum_r n_r P_{ML} = 1 \tag{5.5.3}$$

当 m 较大时（例如 $m = 3$，即 Tri-gram 模型），训练语料中 $C(w_1^m) = 0$ 的可能性便大大增加，这样 $P_{ML} = 0$。$P_{ML} = 0$ 有两种含义，（1） w_1^m 语义上是正确的，但实际语料中没有统计到；（2） w_1^m 是语义上不可能出现的 m 元组。前者仅仅是因为训练语料不足所致。基于这样的考虑，需要对最大似然估计进行修改，对这样的零概率事件加以平滑。图灵估计[Good 53]就是一种常用的平滑估计，其基本思路是在保证总概率为 1 的前提下，把出现次数大于零的词串的概率估计非线性地折扣到零概率上。简单地讲，即：

$$P_T = r^* / N \tag{5.5.4}$$

其中

$$r^* = (r+1) n_{r+1} / n_r \tag{5.5.5}$$

定义系数

$$d_r = r^* / r \tag{5.5.6}$$

为折扣系数，则图灵估计也可以写作

$$P_T = d_r P_{ML} \tag{5.5.7}$$

图灵估计仍然满足

$$\sum_r n_r P_T = 1 \tag{5.5.8}$$

5.5.2 基本退化频度估计算法

目前已经有一些基于图灵估计的频度估计算法，如 Nadas 给出的图灵估计变形[Nadas 85]，Katz 给出的一种基于图灵估计的退化频度估计算法[Katz 87]等。其中，

退化频度估计算法是一种性能较好的方法。它对 $P(w_m \mid w_1^{m-1})$ 进行递归估计，实现统计语言模型的零概率重估。当 $C(w_1^m) > 0$ 时，由图灵估计，$P(w_m \mid w_1^{m-1})$ 的估计 $P_s(w_m \mid w_1^{m-1})$ 定义为：

$$P_s(w_m \mid w_1^{m-1}) = d_{C(w_1^m)} C(w_1^m) / C(w_1^{m-1}) \tag{5.5.9}$$

而当 $C(w_1^m) = 0$ 时，所有这样的词串的图灵估计之和为：

$$\tilde{\beta} = n_0 P_T = n_1 / N = n_1 / C(w_1^{m-1}) \tag{5.5.10}$$

设 $P_s(w_m \mid w_2^{m-1})$ 已知，则可以根据 $P_s(w_m \mid w_2^{m-1})$ 对 $\tilde{\beta}$ 进行加权分配，作为零概率平滑估计 $P_s(w_m \mid w_1^{m-1})$。定义

$$\alpha = \tilde{\beta} \Big/ \sum_{w_m : C(w_1^m)=0} P_s(w_m \mid w_2^{m-1}) \tag{5.5.11}$$

则零概率退化频度递归估计式为

$$P_s(w_m \mid w_1^{m-1}) = \alpha P_s(w_m \mid w_2^{m-1}) \quad （C(w_1^m) = 0） \tag{5.5.12}$$

5.5.3 修正退化频度估计算法

在实际语言模型的构造中，上述的退化模型仍存在着一些缺陷，因而难以直接应用。本文提出以下的修正退化频度估计算法。

1. 对 d_r 的修正

若采用(5.5.6)式对 d_r 的定义直接进行计算，设 $\tilde{r} = \max_{n_r \neq 0} r$，则有 $d_{\tilde{r}} = 0$。这样使得出现次数最多的词串的图灵估计变为了 0，显然不符合实际的情况。若直接使用分段线性弯折的方法对 d_r 进行修正，即给定常数 k （例如取 $k = 5$），认为 $r > k$ 时的极大似然估计已经可以较好地反映实际的词串连接概率，从而令 $d_r \mid_{r>k} = 1$；对于 $1 \le r \le k$ 的情况，令 $(1 - d_r \mid_{1 \le r \le k}) = \mu(1 - r^* / r)$（$\mu$ 为待定比例常数），同时满足 $\sum_{w_m} P_s(w_m \mid w_1^{m-1}) = 1$，得到：

$$d_r = (r^* n_1 - (k+1) r n_{k+1}) / (r n_1 - (k+1) r n_{k+1}) \quad (1 \le r \le k, n_1 \neq 0) \tag{5.5.13}$$

则由于在实际的语料统计中，数据的稀疏会经常使得 $n_r = 0 \quad (1 \le r \le k+1)$，无法直接计算由(5.5.5)式定义的 r^*，因而无法应用(5.5.13)式对 d_r 进行修正。本文使用下面的方法修正 d_r：

- **首先对 n_r 进行平滑。即令：**

$$\tilde{n}_r = \sum_{w_1^{m-1}:C(w_1^{m-1})>0} n_r(w_1^{m-1}) \Big/ \sum_{w_1^{m-1}:C(w_1^{m-1})>0} 1 \qquad (1 \le r \le k+1) \tag{5.5.14}$$

式中 $n_r(w_1^{m-1})$ 表示在 w_1^{m-1} 词前缀下恰好出现了 r 次的不同词串 w_1^m 的计数。因为 $1 \le r \le k+1$ 时属于统计数据稀疏的情况，经过上述的平滑后可以将对应于不同 w_1^{m-1} 词前缀下 w_1^m 的出现次数信息进行综合，获得统计语料 n_r 的整体分布规律。同时减少了 $n_r = 0 \quad (1 \le r \le k+1)$ 的情况。同时，这样的平滑大大减小了折扣系数 d_r 的存储空间，而基本不影响该统计语言模型的性能。

- **对 d_r 的计算公式进一步修正。**

经过 n_r 的平滑后，在 $1 \le r \le k+1$ 时 $\tilde{n}_r$ 仍然可能为 0，此时需要对(5.5.13)式进一步进行修正。定义函数 $\tilde{\theta}(r) = r + \min(\theta \mid \theta \ge 0 \text{且} n_{r+\theta} > 0)$，跳过使 $n_r = 0$ 的 r。将(5.5.13)式改写为：

$$\tilde{d}_r = \frac{\tilde{\theta}(r)\tilde{n}_{\tilde{\theta}(r)} \Big/ (r\tilde{n}_r) - \tilde{\theta}(k)\tilde{n}_{\tilde{\theta}(k)} \Big/ \tilde{\theta}(1)\tilde{n}_{\tilde{\theta}(1)}}{1 - \tilde{\theta}(k)\tilde{n}_{\tilde{\theta}(k)} \Big/ \tilde{\theta}(1)\tilde{n}_{\tilde{\theta}(1)}} \quad (1 \le r \le k, \text{且} \tilde{n}_r \ne 0) \tag{5.5.15}$$

注意，这样的修正仍然满足 $\sum_{w_m} P_s(w_m \mid w_1^{m-1}) = 1$ 的要求。

2. 对 α 的修正

在实际语料的统计中，因为数据稀疏的情况严重，$C(w_1^m) = 0$ 的词串数量相当大，所以使用(5.5.11)式计算 α 时 $P_s(w_m \mid w_2^{m-1})|_{w_m:C(w_1^m)=0}$ 的计算量很大。这样，在实际计算时，常常使用(5.5.11)式的等价形式，即：

$$\alpha = \tilde{\beta} \Big/ (1 - \sum_{w_m:C(w_1^m)>0} P_s(w_m \mid w_2^{m-1})) \tag{5.5.16}$$

然而，在上述 $\tilde{d}_r$ 的分段弯折下，当给定 $w_2^{m-1}, \forall w^m$ 有 $C(w_2^m) > k$ 时，会出现

$\sum_{w_m:C(w_1^m)>0} P_s(w_m \mid w_2^{m-1}) = 1$的情况，从而无法对零概率进行加权分配。此时可以对零概率分配系数α进行下列修正：令

$$\tilde{\alpha} = \tilde{\beta} \Big/ (1 - \sum_{w_m:C(w_1^m)>0,且C(w_2^m)>\tilde{\theta}(0)} P_s(w_m \mid w_2^{m-1})) \tag{5.5.17}$$

5.5.4 修正退化频度估计模型的搜索算法

1. 问题的描述

设待识别的词串共有 L 个词，其中每个词由声学模型最多给出 N 个候选。用$w_{n,l}(1 \le l \le L, 1 \le n \le N)$表示第$l$个词的第 n 个候选。搜索算法的目的是在矩阵$(w_{n,l})_{(1\le l\le L, 1\le n\le N)}$中确定一条从左至右的“最佳”路径，使得 n-gram 模型对该路径评分最高。以 Tri-gram 模型为例，就是要给出指标序列$\{a_l \mid 1 \le l \le L, 1 \le a_l \le N\}$，使得下式取到最大值：

$$P = P(w_{a_1,1})P(w_{a_2,2} \mid w_{a_1,1}) \prod_{l=3}^{L} P(w_{a_l,l} \mid w_{a_{l-2},l-2}, w_{a_{l-1},l-1}) \tag{5.5.18}$$

2. Beam Search 算法

Beam Search 是一种寻找次优路径的方法。设置一个有限规模的路径保存缓冲区，针对待搜索路径的矩阵$(w_{n,l})_{(1\le l\le L, 1\le n\le N)}$从左至右进行路径的扩展。在扩展路径时对保存在缓冲区内的路径得分进行排序，同时抛弃缓冲区内由于路径扩展而无法存放的低分路径。

这种方法运算复杂度低，容易实现，但它存在着一些最优或次优路径在搜索的早期阶段即被截断的危险，即最终搜索的结果将受到搜索早期情况的极大影响，造成l指标小的候选词与l指标大的候选词的在搜索地位上的不平等。事实上，Beam Search 在声学层面应用时也有同样的问题，柴提出的建立反向索引再进行一次从右至左的 Beam Search，然后综合两次结果的方法[Chai 96]可以在一定程度上解决这个问题。

3. 图最优路径搜索及改进

使用图最优路径搜索算法，可以保证找到最优的路径。然而，图最优路径搜索只能找到最优的一条路径，而不能给出最优的前 k 条路径候选，这是它不用于声学层的主要原因之一。但是，语言层面在听写机中处于要求给出最终结果的位置，上述问题不再是图最优路径搜索算法应用于语言层面的障碍。以 Tri-gram 模

型为例，图 5.1 表示了图最优搜索算法的搜索过程：

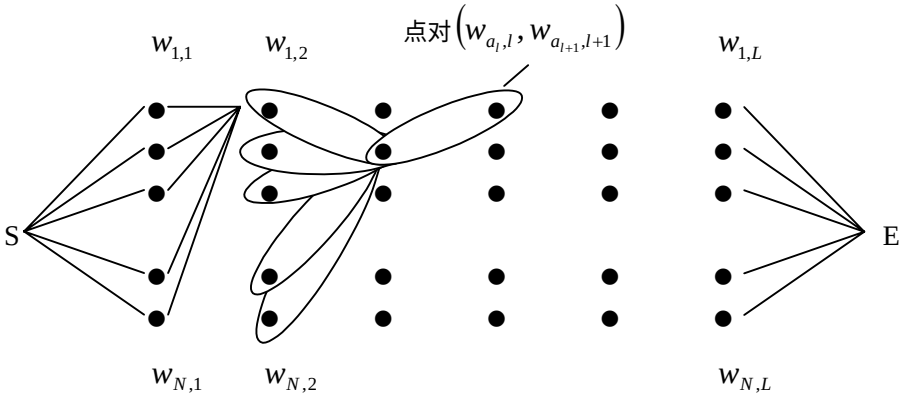

图 5.1　图最优搜索算法的搜索过程

在 Trigram 模型中，因为每次涉及的基本计分单位是连续 3 个词的 Trigram 得分，所以矩阵 $(w_{n,l})_{(1\le l\le L,1\le n\le N)}$ 图搜索算法对应的图节点应为一个点对 $(w_{a_l,l}, w_{a_{l+1},l+1})$。为保证搜索出全局最优路径，需要保留所有到达当前点对 $(w_{a_l,l}, w_{a_{l+1},l+1})$的路径中得分最高的路径。这样，该搜索算法的时间和空间复杂度分别为$O(N^3L)$和$O(N^2L)$。由于复杂度与是 N 的平方或三次方的关系，所以复杂度对 N 值的增大十分敏感。而为了保证声学模型给出的候选中包括正确的结果，N 值不可能过小。所以，考虑一种折衷的方法降低这样的敏感性，即：对于当前词 $w_{a_{l+1},l+1}$ 涉及的 N 个词对 $(w_{a_l,l}, w_{a_{l+1},l+1})(a_l = 1,2,\cdots N)$，仅保留前 M$(1 \le M \le N)$个最佳的结果。这样，时间和空间复杂度分别降低为$O(MN^2L)$和$O(MNL)$。当 M=N 时，是最优搜索；在极端情况下 M 降低为 1 时，该算法的时间复杂度与 Beam Search 相当，但是它更有可能获得较优的搜索结果，因为它对任何 l 值的各个候选词都保留有一条路径，一定程度上克服了 l 指标小的候选词与 l 指标大的候选词的在搜索地位上的不平等问题。

第六章 实验结果与结论

6.1 GMSM 的测试

GMSM 实验测试所用数据库是由国家 863-306 主题组委托中国科技大学、中国科学院声学所、社会科学院语言研究所等构建的汉语普通话连续语音数据库，以下简称 863 数据库。数据采样是 PC 上利用 16 位标准声卡完成的，采样率为 16KHz，采样精度是 16bits。特征抽取的分辨率是 32ms 帧宽，16ms 帧移和 8ms 帧移。

6.1.1 训练集和测试集的选取

对 863 数据库经过人工切分和标音之后，计算得到 LPC-CEP 和 AR-CEP 特征参数，存成“ *.CEP ”文件。第一批光盘数据有 1560 个句子，这些句子分成 ABC 三组，每组的句子数目大致相等。第二批光盘数据有 120 个无意义音节串，分别由三男三女完成，每人 20 个音节串，作为 D 组。ABCD 四组中男声和女声的数据量相当。为了实验的速度和一定的代表性，在大部分实验中我们选取了所有的男声作为训练和识别数据。

Total :

class A: 00 02--06 18 20--22 24--26 [13 men]
class B: 01 07--11 29--34 36 [13 men]
class C: 12--16 39 41--45 49 [12 men]
class D: m1--m3 [3 men]

Train Set: (398 syllables , 180065 samples)

class A: 04--06 18 20--22 24--26 [10 men]
class B: 09--11 29--34 36 [10 men]
class C: 12--16 39 41 44--45 49 [10 men]

注：每个 class 的前 9 个人的按名称排列的前 60 个文件，以及每个 class 的第 10 个人的按文件大小排列的前 60 个文件除外，这些文件构成了 Test Set I 的 Part II。

Test Set I: (397 syllables , 70593 samples)

Part I: (397 syllables , 49286 samples)

class A: 00 02 03 [3 men]

class B: 01 07 08 [3 men]

class C: 42 43 [2 men]

Part I 涉及人的数据没有参加训练。

Part II: (397 syllables, 21307 samples

Train Set ABC 各组的前 9 个人的按名称排列的前 60 个文件，以及第 10 个人的按文件大小排列的前 60 个文件。

Part II 涉及人的一部分数据参加了训练。

Test Set II: (182 syllables, 355 samples)

class D: m1--m3 [3 men]

Test Set II 涉及人的数据没有参加训练。

6.1.2 GMSM 测试结果

本小节的记号：

K：GMSM 分段数，CDCPM 或 HMM 状态数。

M：GMSM 或 CHMM 中高斯混合分布的阶，CDCPM 中的混合数。

U：GMSM 训练收敛阈值。

TT：训练数据的帧移(ms)。

TR：识别数据的帧移(ms)。

未加特别说明，本小节给出的是在测试集 Test Set I 下的测试结果，覆盖 397 个音节的 70593 个测试样本，总规模约是训练集的一半，包括了未参加训练的人的样本和参加训练的人在训练集 Train Set 外的样本。该测试集规模较大，数据来源较全面，因而结果是比较客观的。此外，以下未加特别说明时，GMSM 的计分策略均采用(3.3.18)式的“ 标准 ”方法。

1. GMSM 与 CDCPM（K=6, U=1.0, TT=TR=16ms）

表 6.1 GMSM 与 CDCPM 模型性能(%)

前n名	1	2	3	4	5	6	7	8	9	10
CDCPM(M=16)	52.25	69.95	78.26	82.92	86.06	88.13	89.82	91.08	92.11	93.93
CDCPM(M=8)*	40.26	55.93	64.59	70.20	74.12	77.21	79.59	81.43	82.90	84.10
GMSM(M=16)	62.60	77.15	83.37	86.75	89.02	90.62	91.69	92.54	93.27	93.88
GMSM(M=8)	61.06	76.25	82.88	86.51	88.88	90.52	91.73	92.69	93.41	94.05

*注：此结果的测试集是 Test Set I 的 Part II。

2. 帧移的变化(K=6, M=16, U=1.0, TR=16ms)

表 6.2 GMSM 与 CDCPM 在帧移变化下的性能(%)

前n名	1	2	3	4	5	6	7	8	9	10
CDCPM(TT=16ms)	52.25	69.95	78.26	82.92	86.06	88.13	89.82	91.08	92.11	93.93
CDCPM(TT=8ms)	51.91	69.45	77.67	82.30	85.33	87.52	89.23	90.50	91.51	92.35
GMSM(TT=16ms)	62.60	77.15	83.37	86.75	89.02	90.62	91.69	92.54	93.27	93.88
GMSM(TT=8ms)	58.79	73.91	80.65	84.50	86.91	88.61	89.96	91.00	91.87	92.54

3. GMSM 简化计分策略的测试(K=6, U=1.0, TR=16ms)

表 6.3 GMSM 标准与简化*计分策略的性能 (%)

前n名	1	2	3	4	5	6	7	8	9	10
标准 (M=16, TT=8ms)	58.79	73.91	80.65	84.50	86.91	88.61	89.96	91.00	91.87	92.54
简化 (M=16, TT=8ms)	57.68	73.22	80.31	84.16	86.65	88.44	89.86	90.91	91.73	92.43
标准 (M=8, TT=16ms)	61.06	76.25	82.88	86.51	88.88	90.52	91.73	92.69	93.41	94.05
简化 (M=8, TT=16ms)	60.26	75.73	82.52	86.28	88.64	90.39	91.65	92.60	93.33	93.94

*注：“ 标准 ”指使用(3.3.18)式计分，“ 简化 ”指使用(3.3.19)式计分。

4. GMSM 训练收敛阈值的变化(K=6, M=16, TT=TR=16ms)

表 6.4 GMSM 训练收敛阈值变化对模型性能的影响(%)

前n名	1	2	3	4	5	6	7	8	9	10
U=10.0	60.92	76.02	82.54	86.12	88.48	90.17	91.35	92.26	92.99	93.60
U=1.0	62.60	77.15	83.37	86.75	89.02	90.62	91.69	92.54	93.27	93.88
U=0.01	62.66	77.02	83.30	86.71	88.92	90.46	91.66	92.56	93.24	93.81

5. GMSM 与经典 CHMM (K=6, M=16, TT=TR=16ms)

本测试的训练集是 Train Set 加 Test Set I，包括相应的女声数据。测试集是 Test Set II，也包括相应的女声数据。识别基元边界由一个切分算法确定，可能因此引入附加的错误。

表 6.5 GMSM 与 CHMM 模型性能(%)

前n名	1	2	3	4	5
GMSM	50.37	64.32	71.50	79.00	80.20
CHMM	49.93	63.87	72.56	79.01	81.41

6.2 修正退化频度估计算法的测试

6.2.1 语料数据库与测试数据库

在进行实际语言模型的构造时，采用了较大规模的语料库，包括 1993 年人民日报全文，市场报摘编，以及新华社文稿摘编共约 4 千万词的语料库。基本词表 24713 个词，全部语料库进行了分词和词号的标注。

测试使用的数据库为训练语料集外随机选取的语句。所有语句按照基本词表进行了分词，共包含 6070 个词。句子中的每一个词 w 都随机给出 $q-1$ 个干扰候选，加上词 w 本身共 q 个候选。此外，我们还将此语言模型挂接在实际的声学模型之后，进行了实际语句的口呼测试。口呼测试用句亦在训练语料集外随机选取，共 36 句，407 词。

6.2.2 修正退化频度估计算法测试结果

1. n-gram 统计模型困惑度测试

困惑度是衡量一个基于 n-gram 的统计语言模型的基本参数，定义为：

$$H = -\frac{1}{N-n+1}\sum_{i=n}^{N}\log(P_s(w_i \mid w_{i-n+1}^{i-1})) \tag{6.2.1}$$

式中 n 为 2 或 3 分别对应 Bi-gram 或 Tri-gram；N 为测试文本的总词数。H 值越小，表示模型的不确定度越低，性能越好。表 6.6 给出了基本 n-gram 模型和修正退化频度估计模型的困惑度 H。

表 6.6　修正退化频度估计模型困惑度测试*

	基本 n-gram 模型	修正退化频度估计模型
Bi-gram	132	104
Tri-gram	97	82

*注：该测试使用人民日报 93 全文作为测试用语料。

2. 汉语语句识别结果

该项测试采用 3-gram 修正退化频度估计模型，随机选取训练语料集外的语句进行语句识别。待识别语句共包含 6070 个词。每个词给出 q 个候选。设待识别语句的长度为 l 个词，则语言模型给出的语句识别结果 $\tilde{w}_1^l$ 由下式确定：

$$\tilde{w}_1^l = \arg\max_{w_1^l} \prod_{i=1}^{l-2} P_s(w_{i+1}^{i+2} \mid w_i) \tag{6.2.2}$$

测试时(5.5.15)式中的k值取 5。由于构建的语言模型所依据的词表较大，最终模型所占的空间也较大。在本项测试中，设定了一个计数阈值T，在进行语料库统计时，所有出现次数r满足$0 \le r \le T$的词（即出现次数很少的词）均按照出现 0 次对待。这样，虽然牺牲了一些模型的性能，但可以大幅度地缩小模型所占的存储空间。事实上，实际进行的统计表明$T=1$时的模型占用空间是$T=0$时的 1/3 左右。搜索采用 5.5.4 节所述的图最优路径搜索算法。

表 6.7 给出了不同的q值和T值下全部测试文本的正确识别率。

表 6.7　修正退化频度估计模型汉语语句正确识别率

	$T=0$	$T=1$
$q=10$	94.4%	92.5%
$q=50$	93.5%	90.2%
$q=100$	91.1%	88.1%

3. 口呼汉语语句识别测试

将上述 Tri-gram 修正退化频度估计模型挂接在声学模型之后，进行实际口呼汉语语句的集外测试。口呼测试用句共 36 句，407 词，均是在训练语料集外随机选出。每个词声学模型设定给出最多不超过 50 个的候选词。候选词中可能不包括正确的待识别词。这样，对于该语言模型的语句正确识别率有一定影响。实际测试的结果如表 6.8 所示（$T=0$）。搜索中使用 5.5.4 节中提高速度的改进算法，令 M=1，即仅保留一个最佳点对的路径。

表 6.8　口呼识别正确率

	正确识别率（%）
声学模型给出的首选	63.2%
声学模型加修正退化频度估计语言模型	87.7%

6.3 结论

1. 本文提出的 GMSM 在进行汉语全音节识别时具有较高的正确率。GMSM 采用分段确定状态，因而比经典 CHMM 识别效率更高，是一种比较实用的声学模型。
2. 由于训练集中某些识别基元的训练样本数过少，拟采用减小帧移的方法增加它们的训练量。然而，以 8ms 帧移进行训练的模型总体性能均较 16ms 帧移低，这有可能是因为这样使得大多数训练样本充足的识别基元的训练数据量过大，导致段内特征矢量在特征空间内的分布离散度加大，而模型中高斯混合分布的阶相对过小，模型描述能力不足所至。因此，仍然需要在语音库的设计中尽量均衡不同识别基元间的训练量（参见附录 1 中 GMSM 对不同识别基元的识别情况）。
3. GMSM 标准计分策略(3.3.18)与简化计分策略(3.3.19)对模型性能影响不大，但是简化计分策略可以较大幅度地提高识别速度。
4. 随着 GMSM 训练收敛阈值的减小，GMSM 模型参数的迭代终值更加稳定，其性能有所提高。然而过小的训练收敛阈值会使 GMSM 的训练过程大幅延长，因此取比较合适的阈值（例如 U=1.0）即可。
5. 测试结果证明，本文描述的修正退化频度估计模型是一种构造比较简单，且易于实用的基于 n-gram 的统计语言模型。同时，该统计语言模型可以作为基于规则的语言模型处理的基础，在现有结果上继续进行校验和改进，从而提高整个语音识别系统的性能。
6. 基于修正退化频度估计算法的 n-gram 语言模型的性能对于计数阈值 T 的选取不是十分敏感，这对于进一步降低模型对空间的要求十分有利。

参考文献

1. **Baum, L.E., (1972)** "An inequality and associated maximization technique in statistical estimation of probabilistic functions of Markov processes," *Inequalities*, 3, 1972
2. **Bellegarda, J.R., Nahamoo, D., (1990)** "Tied Mixture Continuous Parameter Modeling for Speech Recognition," *IEEE Trans. on ASSP*, vol.ASSP-38, No.12, pp.2033-2045, Nov. 1990
3. **柴海新， 吴文虎， 方棣棠， (Chai 1994)** "连续语音识别的研究和汉语数字连呼系统的实现" 杨家沅编：语音识别与合成(NCMMSC-94 论文集)，277-287。四川：四川科学技术出版社，1994
4. **柴海新，方棣棠， (Chai 1996)** "汉语大词表识别策略和非特定人听写机实验系统" 软件学报，1996 年 10 月
5. **Davis, S.B., Mermelstein, P., (1980)** "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on ASSP*, Aug. 1980
6. **Dempster, A., Laird, N., Rubin, D., (1977)** "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society B*, 39(1): 1~38, 1977
7. **Franzini, M.A., Lee, K.F., (1990)** "A New Hybrid Method for Continuous Speech Recognition", *ICASSP-90*, 1990
8. **Furui, S., (1986)** "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. on Acoust., Speech, and Signal Processing*, 34(1): 52-59, Feb., 1986.
9. **Good, I., (1953)** "The population frequencies of species and the estimation of population parameters," *Biometrika*, vol. 40, No. 3 and 4, pp. 237-264, 1953
10. **Huang, X.D., Jack, M.A., (1989)** "Semi-continuous hidden Markov models for speech signals," *Computer Speech and Language*, 3:239-251, 1989
11. **Jelinek, F., Mercer, R.L., Bahl, L.R., (1983)** "A maximum likehood approach to continuous speech recognition", *IEEE Trans. On Pattern Analysis and Machine Intelligence*, Vol. PAMI-5, pp179-190, Mar. 1983
12. **Katz, S., (1987)** "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *IEEE Trans. Acoust., Speech, Signal processing*, vol. ASSP-35, pp. 400-401, Mar. 1987
13. **LEE, CHIN-HUI *et al.*, (1989)** "A Frame-Synchronous Network Search Algorithm for Connected Word Recognition", *IEEE trans. acoustics, speech, and signal processing*, Vol. 37, No.11, Nov. 1989
14. **李志敏, (Li 1995)** "大字表语音识别系统中的语言模型"，硕士学位论文，清华大学计算机科学与技术系，1995
15. **Linde, Y., Buzo, A., Gray, R.M., (1980)** "An algorithm for vector quantization," *IEEE Trans. On COM*, 28(1), Jan. 1980
16. **Morgan, N., Bourlard, H., (1990)** "Continuous Speech Recognition Using Multilayer

Perception with Hidden Markov Models", *ICASSP-90*, 1990

17. **Nadas, A., (1985)** "On Turing's formula for word probabilities," *IEEE Trans. Acoust., Speech, Signal processing*, vol. ASSP-33, pp. 1414-1416, Dec. 1985
18. **Nagao, M., Mori S., (1996)** "A New method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text data of Japanese", *Invited Speech in ICCC'96 Singapore*, **June 1996**
19. **包约翰,著.马颂德,等,译. (Pao 1992)** 自适应模式识别与神经网络.北京:科学出版社,1992
20. **Parzen., E, (1962)** "On estimation of a probability density function and mode," *Annals of Mathematics and Statistics*, 33, 1962
21. **Rabiner, L.R., Schafer, R.W., (1978)** Digital Processing of Speech Signals. USA: Prentice-Hall, Inc., 1978
22. **Rabiner, L.R., Wilpon, J.G., Soong, F.K., (1988)** "High performance connected digit recognition using hidden Markov models," *ICASSP-88*, 1988
23. **Rabiner, L.R., (1989)** "A Tutorial on Hidden Markov Models and Selected Applications in Speech recognition", *IEEE Proceedings*, Vol. 77, No.2, pp.257-286, Feb. 1989
24. **Soong, F. K., Rosenberg, A.E., (1986)** "On the use of instantaneous and transitional spectral information in speaker recognition," *ICASSP-86*, 877-880, 1986
25. **Vintsjuk, T.K, (1968)** "Recognition of Words of Oral Speech by Dynamic Programming", *Kibernetica*, Vol. 81, No. 8, 1968.
26. **Viterbi, A.J., (1967)** "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. on IT*, 13(2), Apr. 1967
27. **Wilpon, J.G., Lee, C.H., Rabiner, L.R., (1989)** "Application of hidden Markov models for recognition of a limited set of words in unconstrained speech," *ICASSP-89*, 3: 254-257
28. **Wilpon, J.G., Miller, L.G., Modi, P., (1991)** "Improvements and Applications for Key Word Recognition Using Hidden Markov Modeling Techniques," *ICASSP-91*, pp. 905-908
29. **张树武, (Zhang 1997)** "汉语语言处理及语言模型研究"，博士学位论文，中国科学院自动化研究所，1997
30. **郑方， 吴文虎， 方棣棠， (Zheng 1996)** "CDCPM 及其在语音识别中的应用 "， 软件学报，863 高技术项目智能主题专刊，7:69-75, 1996
31. **Zheng, F., Chai H., Shi, Z., Wu W., Fang, D, (1997)** "A Real-World Speech Recognition System Based on CDCPMs," *ICCPOL-97*, vol.1, pp.204-207, 1997
32. **朱得熙， (Zhu 1982)** "语法讲义", 商务印书馆 1982

附 录

1. K=6,U=10,TT=TR=16ms,M=16 的 GMSM 模型（参见 6.1.2 节）对 Test Set I 不同音节的识别情况。

[音 节]	(有效样本数 / 总样本数)	首选(%)					十选(%)					十选外(%)
[shang]	Results(433 / 434 Samples) :	76.67	87.99	92.15	94.69	97.00	97.92	97.92	98.38	98.61	99.08	---- 0.92
[hai]	Results(345 / 347 Samples) :	76.23	87.83	92.75	95.07	95.65	97.10	97.39	97.68	97.68	97.97	---- 2.03
[de]	Results(2744 / 2744 Samples) :	53.68	67.24	74.38	78.53	82.43	84.66	86.22	88.05	89.58	90.67	---- 9.33
[gong]	Results(579 / 579 Samples) :	86.87	92.06	94.13	95.85	96.72	97.75	98.27	98.45	98.62	98.62	---- 1.38
[ren]	Results(637 / 638 Samples) :	76.45	86.03	89.80	91.84	93.72	94.51	94.98	95.92	96.23	96.55	---- 3.45
[shi]	Results(2430 / 2435 Samples) :	81.52	93.87	97.08	97.90	98.11	98.35	98.64	99.01	99.09	99.30	---- 0.70
[fu]	Results(378 / 378 Samples) :	85.71	91.80	93.65	94.44	95.24	96.56	96.83	97.09	97.35	97.35	---- 2.65
[ke]	Results(334 / 334 Samples) :	73.65	91.02	94.91	96.11	97.01	98.50	98.50	98.50	98.50	98.80	---- 1.20
[kun]	Results(38 / 38 Samples) :	42.11	44.74	47.37	52.63	57.89	57.89	60.53	71.05	73.68	73.68	---- 26.32
[nan]	Results(162 / 162 Samples) :	55.56	72.84	79.01	82.72	85.19	86.42	88.89	91.36	92.59	95.06	---- 4.94
[er]	Results(537 / 539 Samples) :	91.06	94.04	95.53	96.09	96.46	96.83	97.02	97.21	97.21	97.21	---- 2.79
[ou]	Results(98 / 98 Samples) :	63.27	73.47	78.57	80.61	80.61	82.65	85.71	85.71	86.73	87.76	---- 12.24
[zhou]	Results(122 / 122 Samples) :	49.18	68.03	77.87	81.97	85.25	86.07	86.07	88.52	92.62	92.62	---- 7.38
[qi]	Results(541 / 542 Samples) :	72.83	87.43	92.61	94.82	96.12	97.04	97.60	97.97	98.15	98.15	---- 1.85
[huo]	Results(200 / 200 Samples) :	71.50	89.00	92.50	94.50	96.50	97.50	97.50	98.00	98.50	99.00	---- 1.00
[chang]	Results(384 / 384 Samples) :	67.71	87.50	92.97	95.31	96.09	97.40	97.66	98.44	98.70	98.96	---- 1.04
[chu]	Results(395 / 397 Samples) :	73.16	84.05	88.86	90.38	91.39	91.90	92.91	93.67	93.67	94.18	---- 5.82
[le]	Results(643 / 643 Samples) :	53.97	71.07	79.63	85.38	88.34	90.20	91.60	92.85	94.25	95.18	---- 4.82
[chuang]	Results(56 / 56 Samples) :	14.29	25.00	39.29	48.21	57.14	66.07	67.86	69.64	76.79	78.57	---- 21.43
[hui]	Results(592 / 593 Samples) :	85.30	92.91	94.59	95.95	96.96	97.30	97.47	97.47	97.64	97.80	---- 2.20
[yuan]	Results(512 / 512 Samples) :	75.59	88.09	91.60	93.95	95.70	97.07	97.46	98.24	98.63	98.83	---- 1.17
[zhi]	Results(933 / 935 Samples) :	73.74	89.28	94.21	95.82	96.68	97.11	97.32	97.43	97.96	98.29	---- 1.71
[wai]	Results(158 / 158 Samples) :	37.34	56.33	67.72	75.32	81.01	86.71	89.87	93.04	94.30	94.94	---- 5.06
[zhong]	Results(979 / 981 Samples) :	72.22	83.86	88.56	91.52	92.85	94.69	95.71	96.12	96.42	96.94	---- 3.06
[guo]	Results(1030 / 1030 Samples) :	73.69	88.64	92.62	95.73	96.89	97.67	97.96	98.45	98.83	99.42	---- 0.58
[dan]	Results(164 / 165 Samples) :	32.32	48.78	60.37	67.68	75.00	79.88	81.10	83.54	86.59	88.41	---- 11.59
[xuan]	Results(77 / 77 Samples) :	20.78	57.14	67.53	74.03	80.52	85.71	87.01	88.31	88.31	88.31	---- 11.69
[shou]	Results(289 / 289 Samples) :	69.55	82.70	85.47	88.58	91.35	92.73	93.08	93.43	93.77	93.77	---- 6.23
[lin]	Results(116 / 116 Samples) :	37.07	52.59	58.62	64.66	72.41	78.45	81.03	83.62	87.07	88.79	---- 11.21
[li]	Results(823 / 823 Samples) :	72.90	86.88	91.62	94.78	96.23	97.33	97.57	98.06	98.18	98.42	---- 1.58
[wen]	Results(263 / 263 Samples) :	57.79	77.57	84.79	87.45	88.59	91.25	92.40	93.54	93.92	94.30	---- 5.70
[he]	Results(734 / 735 Samples) :	86.92	92.64	95.23	97.00	97.55	97.82	98.09	98.37	98.50	98.77	---- 1.23
[dong]	Results(311 / 311 Samples) :	60.45	79.10	86.82	89.71	91.64	92.28	93.89	94.86	95.50	96.14	---- 3.86
[jiong]	Results(4 / 4 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----100.00
[ji]	Results(1207 / 1208 Samples) :	67.44	86.66	94.95	96.85	97.35	98.09	98.59	99.01	99.17	99.25	---- 0.75
[bai]	Results(213 / 213 Samples) :	49.77	70.89	79.34	87.79	92.02	92.96	93.43	94.37	94.84	95.77	---- 4.23
[ge]	Results(703 / 704 Samples) :	58.46	77.95	84.35	88.05	88.90	89.62	90.47	90.90	91.32	91.89	---- 8.11
[zi]	Results(439 / 439 Samples) :	75.17	91.34	95.90	96.13	96.58	97.04	97.27	97.49	97.49	97.49	---- 2.51
[dui]	Results(346 / 347 Samples) :	67.34	81.21	87.28	91.91	95.09	96.53	97.69	98.27	98.27	98.27	---- 1.73
[jin]	Results(728 / 730 Samples) :	56.18	78.71	86.95	93.54	96.70	97.94	98.35	98.49	98.76	98.90	---- 1.10
[ru]	Results(116 / 117 Samples) :	54.31	74.14	80.17	86.21	89.66	90.52	90.52	91.38	93.10	93.10	---- 6.90
[xia]	Results(194 / 197 Samples) :	61.34	84.02	88.66	89.69	92.27	93.30	94.85	96.91	97.42	97.94	---- 2.06
[yi]	Results(1943 / 1945 Samples) :	80.19	90.43	94.60	96.55	97.89	98.51	98.92	99.23	99.43	99.49	---- 0.51
[lun]	Results(71 / 71 Samples) :	32.39	45.07	54.93	59.15	64.79	67.61	71.83	73.24	77.46	80.28	---- 19.72
[bi]	Results(173 / 174 Samples) :	49.13	71.68	84.97	90.75	92.49	94.80	95.38	96.53	97.11	97.11	---- 2.89
[sai]	Results(88 / 88 Samples) :	34.09	54.55	77.27	86.36	93.18	93.18	93.18	93.18	94.32	95.45	---- 4.55
[wo]	Results(295 / 295 Samples) :	64.07	81.36	87.46	90.17	93.90	95.93	96.61	97.29	97.29	97.97	---- 2.03
[men]	Results(334 / 334 Samples) :	71.56	86.83	90.12	93.71	95.51	95.81	96.71	97.01	97.01	97.31	---- 2.69
[da]	Results(605 / 605 Samples) :	82.15	91.74	96.20	98.02	98.84	99.01	99.34	99.34	99.34	99.50	---- 0.50
[duo]	Results(211 / 211 Samples) :	44.08	57.82	65.88	70.62	78.20	83.41	86.26	88.15	88.63	89.10	---- 10.90
[shu]	Results(300 / 300 Samples) :	67.67	84.00	89.33	91.00	92.67	93.33	94.33	95.00	95.00	96.00	---- 4.00
[ling]	Results(221 / 221 Samples) :	53.85	70.59	78.73	81.00	85.52	87.33	87.78	89.59	90.95	91.40	---- 8.60
[dao]	Results(457 / 457 Samples) :	50.33	68.93	77.46	82.06	86.21	88.84	90.15	91.25	92.56	93.65	---- 6.35
[gan]	Results(148 / 148 Samples) :	47.97	63.51	70.27	72.97	77.70	79.73	82.43	84.46	87.84	88.51	---- 11.49
[bu]	Results(798 / 799 Samples) :	63.53	81.83	89.35	91.98	93.61	94.74	95.61	96.24	96.99	97.99	---- 2.01
[zuo]	Results(356 / 357 Samples) :	77.53	90.17	94.10	95.79	96.63	97.47	98.31	99.16	99.44	99.44	---- 0.56
[qin]	Results(63 / 63 Samples) :	11.11	22.22	28.57	31.75	36.51	57.14	61.90	71.43	79.37	85.71	---- 14.29
[ken]	Results(11 / 11 Samples) :	0.00	0.00	0.00	0.00	9.09	9.09	9.09	9.09	9.09	9.09	---- 90.91
[jing]	Results(575 / 577 Samples) :	55.30	78.26	85.74	88.87	91.48	93.39	94.96	95.65	96.17	96.87	---- 3.13
[ye]	Results(399 / 399 Samples) :	69.17	84.46	89.72	92.98	93.98	94.99	95.74	95.99	96.49	97.74	---- 2.26
[zhu]	Results(403 / 405 Samples) :	61.04	75.93	87.59	90.82	92.56	93.55	94.29	94.54	95.29	96.53	---- 3.47
[zhe]	Results(639 / 643 Samples) :	59.00	72.93	78.87	82.63	86.38	88.26	90.92	92.02	92.64	92.96	---- 7.04
[zheng]	Results(429 / 430 Samples) :	66.43	80.19	86.95	91.84	93.94	95.34	95.57	96.27	97.44	97.44	---- 2.56

```
[  peng] Results(   43 /    43 Samples) : 44.19  58.14  62.79  69.77  74.42  76.74  76.74  76.74  79.07  79.07 ---- 20.93
[    bo] Results(  117 /   117 Samples) : 29.91  51.28  64.10  71.79  77.78  82.05  84.62  87.18  88.89  92.31 ----  7.69
[    fa] Results(  532 /   533 Samples) : 84.40  93.98  96.99  97.56  98.68  98.68  99.06  99.25  99.25  99.25 ----  0.75
[  zhan] Results(  299 /   300 Samples) : 64.55  78.60  84.28  88.29  90.97  92.31  93.98  94.65  95.32  95.99 ----  4.01
[   yao] Results(  307 /   307 Samples) : 62.54  81.43  87.30  90.88  94.79  95.77  97.07  97.39  97.39  97.72 ----  2.28
[   yin] Results(  224 /   225 Samples) : 43.30  67.86  79.02  86.16  88.39  94.20  95.98  95.98  97.32  97.77 ----  2.23
[    di] Results(  447 /   447 Samples) : 60.40  78.75  86.13  90.16  90.60  91.72  92.62  93.29  93.29  93.51 ----  6.49
[    ti] Results(  273 /   273 Samples) : 72.89  85.35  92.31  96.34  97.07  97.80  98.53  98.53  98.90  99.27 ----  0.73
[    xi] Results(  490 /   495 Samples) : 65.10  88.78  95.31  96.73  97.35  98.16  98.16  98.37  98.78  98.98 ----  1.02
[  jian] Results(  642 /   642 Samples) : 59.03  79.28  87.69  92.68  94.24  95.17  96.42  97.20  97.51  97.66 ----  2.34
[   she] Results(  261 /   261 Samples) : 62.84  81.99  86.21  88.51  90.04  90.80  93.10  94.25  95.02  95.79 ----  4.21
[  deng] Results(  152 /   152 Samples) : 37.50  57.24  69.08  73.03  75.66  78.29  80.26  83.55  86.84  89.47 ---- 10.53
[   pei] Results(   57 /    57 Samples) : 49.12  64.91  68.42  77.19  82.46  84.21  87.72  87.72  87.72  87.72 ---- 12.28
[   tao] Results(   46 /    46 Samples) : 10.87  28.26  36.96  43.48  47.83  58.70  71.74  76.09  78.26  80.43 ---- 19.57
[   gai] Results(  128 /   128 Samples) : 53.13  71.88  78.13  85.16  85.16  87.50  88.28  89.06  89.84  89.84 ---- 10.16
[  zhuo] Results(   23 /    23 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[   you] Results(  826 /   826 Samples) : 76.76  86.08  90.19  93.58  95.28  95.76  96.49  96.73  97.09  97.46 ----  2.54
[ cheng] Results(  551 /   552 Samples) : 75.86  88.38  92.38  93.47  95.28  95.83  96.19  97.28  97.28  97.28 ----  2.72
[  xiao] Results(  286 /   288 Samples) : 75.52  89.51  94.06  96.50  98.95  98.95  98.95  99.30  99.30  99.30 ----  0.70
[  zhen] Results(  146 /   147 Samples) : 41.10  63.70  70.55  77.40  81.51  85.62  87.67  90.41  90.41  93.84 ----  6.16
[   wei] Results(  752 /   754 Samples) : 68.48  87.10  92.42  95.08  95.74  96.54  97.34  97.61  97.87  98.14 ----  1.86
[  teng] Results(   12 /    12 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[   fei] Results(  147 /   147 Samples) : 80.95  87.76  90.48  91.84  94.56  95.24  97.28  99.32  99.32  99.32 ----  0.68
[   tui] Results(   46 /    47 Samples) : 39.13  47.83  47.83  54.35  63.04  67.39  69.57  69.57  69.57  69.57 ---- 30.43
[ xiang] Results(  416 /   420 Samples) : 71.15  88.22  95.91  97.60  98.08  99.04  99.04  99.04  99.04  99.28 ----  0.72
[    ju] Results(  388 /   388 Samples) : 67.01  85.05  91.75  94.07  95.36  96.13  96.91  97.68  97.94  98.71 ----  1.29
[   lao] Results(  124 /   124 Samples) : 42.74  59.68  69.35  75.81  79.84  83.06  84.68  87.90  87.90  89.52 ---- 10.48
[   hua] Results(  305 /   305 Samples) : 85.90  94.10  96.07  97.38  97.70  98.36  98.36  99.34  99.34  99.34 ----  0.66
[   zei] Results(    5 /     5 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[  cong] Results(  102 /   102 Samples) : 50.00  68.63  78.43  85.29  88.24  91.18  93.14  94.12  94.12  95.10 ----  4.90
[   tou] Results(  154 /   154 Samples) : 68.18  81.17  82.47  85.06  87.01  87.66  88.31  89.61  89.61  90.91 ----  9.09
[   cai] Results(  206 /   206 Samples) : 85.44  93.69  96.60  98.06  98.54  98.54  99.51  99.51  99.51  99.51 ----  0.49
[  tian] Results(  201 /   202 Samples) : 52.74  69.15  78.11  81.59  84.58  88.06  91.04  92.04  93.03  95.02 ----  4.98
[  ming] Results(  197 /   198 Samples) : 52.79  78.68  85.28  88.32  90.86  92.39  93.40  94.42  95.43  95.43 ----  4.57
[   jiu] Results(  417 /   417 Samples) : 74.10  86.33  90.17  92.09  93.76  95.44  96.16  96.64  96.88  97.12 ----  2.88
[    mi] Results(   99 /    99 Samples) : 66.67  79.80  83.84  89.90  90.91  93.94  93.94  93.94  95.96  97.98 ----  2.02
[  qian] Results(  312 /   313 Samples) : 77.24  88.46  91.99  93.27  95.19  95.51  96.47  97.12  98.40  99.04 ----  0.96
[   ben] Results(   86 /    87 Samples) : 34.88  54.65  62.79  67.44  70.93  75.58  79.07  82.56  83.72  84.88 ---- 15.12
[   zai] Results(  715 /   717 Samples) : 82.38  91.47  94.83  96.08  96.78  97.34  97.90  98.46  98.88  99.16 ----  0.84
[  zeng] Results(   75 /    75 Samples) : 60.00  72.00  76.00  80.00  84.00  88.00  90.67  92.00  92.00  94.67 ----  5.33
[   jia] Results(  545 /   546 Samples) : 83.67  94.86  97.06  97.43  98.17  98.72  98.72  98.72  98.72  98.72 ----  1.28
[   xin] Results(  340 /   342 Samples) : 44.71  63.82  77.65  85.29  90.88  93.53  96.18  97.35  97.94  97.94 ----  2.06
[   tan] Results(  134 /   134 Samples) : 61.19  75.37  82.09  87.31  89.55  90.30  90.30  90.30  91.79  93.28 ----  6.72
[    wu] Results(  555 /   556 Samples) : 75.50  88.47  93.15  95.68  96.76  97.12  97.30  97.30  97.66  97.84 ----  2.16
[shuang] Results(   65 /    67 Samples) : 44.62  60.00  63.08  67.69  70.77  75.38  78.46  80.00  83.08  86.15 ---- 13.85
[  fang] Results(  481 /   481 Samples) : 76.92  88.57  91.68  93.97  95.22  95.84  96.67  97.30  98.13  98.54 ----  1.46
[   dou] Results(  141 /   141 Samples) : 41.84  56.74  63.12  67.38  73.76  77.30  78.01  80.85  85.11  85.11 ---- 14.89
[  biao] Results(  226 /   226 Samples) : 69.91  82.30  90.71  92.04  93.36  95.13  97.35  97.35  98.67  98.67 ----  1.33
[   kuo] Results(   61 /    61 Samples) : 21.31  32.79  40.98  49.18  54.10  62.30  67.21  68.85  70.49  73.77 ---- 26.23
[    mu] Results(  194 /   195 Samples) : 76.80  81.96  87.11  88.66  89.69  92.27  92.78  94.33  94.33  94.85 ----  5.15
[   bei] Results(  291 /   292 Samples) : 61.86  83.85  91.41  94.85  96.22  96.91  97.25  97.25  97.59  97.94 ----  2.06
[  zong] Results(  195 /   195 Samples) : 79.49  88.72  92.31  95.90  97.44  97.44  97.44  97.95  97.95  97.95 ----  2.05
[  tong] Results(  353 /   354 Samples) : 77.34  88.95  92.35  93.77  94.90  95.75  96.88  97.45  97.45  98.02 ----  1.98
[    ba] Results(  298 /   298 Samples) : 58.39  83.22  93.62  95.30  97.32  97.65  98.32  98.32  98.99  98.99 ----  1.01
[    du] Results(  184 /   185 Samples) : 46.74  61.41  67.39  76.63  83.15  88.04  91.85  92.93  93.48  96.20 ----  3.80
[   hou] Results(  185 /   185 Samples) : 57.84  74.05  83.24  87.57  89.73  90.81  92.43  93.51  94.05  94.05 ----  5.95
[    ci] Results(  217 /   217 Samples) : 63.59  85.25  95.39  96.77  98.16  99.08  99.08  99.08  99.08  99.08 ----  0.92
[   min] Results(  258 /   258 Samples) : 58.91  75.97  83.72  85.66  89.92  92.25  94.19  95.74  96.90  97.29 ----  2.71
[ jiang] Results(  331 /   335 Samples) : 67.07  86.71  92.45  94.56  96.68  97.58  97.89  98.19  98.19  98.49 ----  1.51
[    ze] Results(  107 /   107 Samples) : 56.07  70.09  75.70  83.18  85.98  87.85  90.65  90.65  90.65  91.59 ----  8.41
[  wang] Results(  149 /   149 Samples) : 58.39  73.83  81.21  84.56  88.59  91.95  92.62  93.96  94.63  94.63 ----  5.37
[  ding] Results(  179 /   179 Samples) : 41.34  55.31  65.36  70.95  77.09  81.01  83.24  85.47  88.27  89.94 ---- 10.06
[    re] Results(   45 /    45 Samples) : 20.00  28.89  37.78  44.44  51.11  57.78  62.22  66.67  75.56  80.00 ---- 20.00
[   lie] Results(  116 /   116 Samples) : 40.52  56.90  63.79  66.38  69.83  75.00  76.72  79.31  79.31  81.03 ---- 18.97
[  huan] Results(  135 /   135 Samples) : 62.22  77.04  83.70  85.93  87.41  90.37  90.37  90.37  92.59  93.33 ----  6.67
[  ying] Results(  189 /   190 Samples) : 51.32  71.96  83.07  87.83  89.95  90.48  91.01  91.53  93.12  94.18 ----  5.82
[  shuo] Results(  180 /   181 Samples) : 67.22  77.22  84.44  86.67  88.89  90.00  90.00  90.00  91.11  92.22 ----  7.78
[  neng] Results(  132 /   132 Samples) : 58.33  75.76  81.82  84.85  89.39  90.91  92.42  93.94  94.70  94.70 ----  5.30
[   hen] Results(   77 /    77 Samples) : 68.83  83.12  85.71  87.01  87.01  89.61  89.61  89.61  89.61  89.61 ---- 10.39
[   gao] Results(  194 /   195 Samples) : 62.89  76.29  80.93  84.54  88.14  89.69  90.21  91.75  93.81  93.81 ----  6.19
[  xing] Results(  472 /   473 Samples) : 63.35  78.60  84.96  88.14  91.10  92.37  92.80  93.86  95.76  96.61 ----  3.39
[  rong] Results(  109 /   109 Samples) : 47.71  77.06  84.40  86.24  89.91  91.74  91.74  92.66  94.50  94.50 ----  5.50
[   yue] Results(  169 /   170 Samples) : 55.62  75.15  79.29  83.43  85.21  88.17  89.94  89.94  91.12  91.72 ----  8.28
[   liu] Results(  175 /   175 Samples) : 60.00  82.86  90.29  93.71  95.43  96.57  97.14  97.14  98.29  98.29 ----  1.71
[    ri] Results(  214 /   214 Samples) : 79.44  89.25  92.06  94.86  96.26  96.26  96.26  96.26  96.26  96.26 ----  3.74
[   pai] Results(  118 /   118 Samples) : 46.61  72.03  82.20  86.44  89.83  92.37  94.07  94.92  94.92  95.76 ----  4.24
[  ting] Results(   70 /    70 Samples) : 17.14  30.00  44.29  60.00  70.00  77.14  78.57  82.86  82.86  85.71 ---- 14.29
```

```
[    en] Results(   71 /    71 Samples) : 18.31  23.94  29.58  39.44  43.66  50.70  56.34  57.75  60.56  63.38 ---- 36.62
[   zui] Results(  103 /   103 Samples) : 62.14  73.79  80.58  82.52  86.41  91.26  92.23  92.23  94.17  96.12 ----  3.88
[  pian] Results(   47 /    47 Samples) : 14.89  29.79  46.81  57.45  61.70  68.09  74.47  74.47  76.60  80.85 ---- 19.15
[    zu] Results(  152 /   152 Samples) : 76.97  87.50  91.45  93.42  94.74  95.39  96.05  96.71  97.37  98.03 ----  1.97
[   che] Results(   63 /    63 Samples) : 17.46  31.75  42.86  50.79  58.73  61.90  68.25  71.43  76.19  79.37 ---- 20.63
[   tuo] Results(   30 /    30 Samples) : 13.33  13.33  26.67  26.67  26.67  33.33  33.33  33.33  40.00  40.00 ---- 60.00
[   kai] Results(  196 /   197 Samples) : 68.88  79.59  85.71  89.29  93.88  95.41  95.92  96.94  96.94  97.45 ----  2.55
[    qu] Results(  408 /   409 Samples) : 75.00  90.93  94.12  96.08  96.81  97.30  98.04  98.28  98.53  98.77 ----  1.23
[    yu] Results(  613 /   615 Samples) : 84.50  91.68  94.45  96.08  96.90  97.39  98.04  98.69  99.02  99.02 ----  0.98
[  nian] Results(  378 /   378 Samples) : 62.70  80.69  89.42  93.65  94.97  96.30  97.35  97.88  98.15  98.41 ----  1.59
[   kao] Results(   41 /    41 Samples) : 24.39  48.78  60.98  65.85  70.73  73.17  78.05  78.05  78.05  80.49 ---- 19.51
[  duan] Results(   88 /    88 Samples) : 17.05  32.95  43.18  52.27  61.36  64.77  65.91  69.32  70.45  75.00 ---- 25.00
[   san] Results(  194 /   194 Samples) : 82.47  90.72  95.88  97.42  97.94  97.94  98.97  98.97  98.97  98.97 ----  1.03
[    hu] Results(  257 /   257 Samples) : 82.49  90.27  92.61  94.94  95.72  96.50  97.28  98.44  98.83  99.22 ----  0.78
[  ning] Results(    6 /     6 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[    su] Results(  117 /   117 Samples) : 44.44  73.50  81.20  86.32  88.89  93.16  95.73  95.73  95.73  95.73 ----  4.27
[    lu] Results(  139 /   139 Samples) : 48.92  59.71  67.63  74.10  80.58  85.61  86.33  88.49  90.65  92.09 ----  7.91
[  yang] Results(  160 /   160 Samples) : 55.00  79.38  88.75  92.50  93.75  94.38  96.25  96.88  97.50  98.75 ----  1.25
[  mian] Results(  179 /   179 Samples) : 67.04  82.12  88.27  91.62  92.18  95.53  96.09  97.21  97.21  98.32 ----  1.68
[  dian] Results(  240 /   240 Samples) : 43.33  65.42  75.83  81.67  86.25  91.67  94.17  95.42  96.67  96.67 ----  3.33
[  ping] Results(  198 /   198 Samples) : 68.18  82.32  86.87  90.40  92.93  94.44  95.45  95.96  95.96  97.47 ----  2.53
[   xue] Results(  175 /   175 Samples) : 57.14  80.57  89.71  93.71  94.86  94.86  95.43  96.00  96.57  97.14 ----  2.86
[  lian] Results(  155 /   158 Samples) : 56.13  69.68  76.13  78.71  81.94  87.10  88.39  90.32  92.90  93.55 ----  6.45
[   ban] Results(  194 /   194 Samples) : 33.51  50.00  68.04  72.16  77.84  84.54  86.08  88.66  91.75  92.78 ----  7.22
[   lai] Results(  292 /   292 Samples) : 67.47  82.19  89.04  91.10  92.81  95.21  95.55  95.55  96.58  96.58 ----  3.42
[   lei] Results(   33 /    33 Samples) :  9.09   9.09   9.09  15.15  21.21  24.24  30.30  30.30  33.33  36.36 ---- 63.64
[     e] Results(  135 /   135 Samples) : 47.41  71.85  77.04  82.96  88.15  89.63  91.11  92.59  94.07  94.07 ----  5.93
[  shui] Results(  235 /   236 Samples) : 75.74  80.00  84.26  92.34  94.04  95.74  96.17  96.17  97.02  97.87 ----  2.13
[  tiao] Results(   97 /    97 Samples) : 64.95  75.26  86.60  91.75  92.78  94.85  95.88  96.91  96.91  96.91 ----  3.09
[  bing] Results(  132 /   133 Samples) : 36.36  50.76  60.61  65.91  72.73  78.03  81.82  85.61  87.88  90.15 ----  9.85
[  dang] Results(  188 /   188 Samples) : 59.57  71.28  77.66  81.38  84.57  87.23  87.77  90.43  92.02  93.62 ----  6.38
[  xian] Results(  410 /   412 Samples) : 50.98  75.85  85.37  90.73  94.63  96.10  96.83  97.32  98.29  98.78 ----  1.22
[  yong] Results(  172 /   172 Samples) : 69.77  80.23  84.30  88.37  89.53  92.44  93.02  93.60  94.19  95.35 ----  4.65
[  suan] Results(   18 /    18 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   5.56 ---- 94.44
[   cao] Results(   35 /    35 Samples) : 40.00  57.14  71.43  80.00  82.86  88.57  88.57  91.43  91.43  91.43 ----  8.57
[  heng] Results(   13 /    13 Samples) :  0.00   0.00   0.00   7.69   7.69   7.69   7.69   7.69   7.69   7.69 ---- 92.31
[    ya] Results(  151 /   152 Samples) : 70.20  83.44  88.08  93.38  96.03  96.03  96.69  96.69  98.01  98.68 ----  1.32
[  bian] Results(  166 /   166 Samples) : 37.35  56.02  61.45  69.88  75.90  80.72  81.33  83.13  86.75  87.35 ---- 12.65
[ liang] Results(  294 /   294 Samples) : 79.25  92.18  95.58  96.26  96.94  97.28  97.62  98.30  98.30  98.98 ----  1.02
[   tai] Results(  132 /   132 Samples) : 43.18  65.15  78.03  87.12  91.67  93.94  96.21  96.21  96.97  97.73 ----  2.27
[  beng] Results(    5 /     5 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[   qie] Results(   52 /    52 Samples) :  7.69  17.31  30.77  42.31  48.08  57.69  61.54  67.31  67.31  67.31 ---- 32.69
[   xun] Results(   84 /    84 Samples) : 20.24  41.67  53.57  64.29  71.43  77.38  84.52  88.10  88.10  89.29 ---- 10.71
[  ruan] Results(   11 /    11 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   9.09   9.09   9.09   9.09 ---- 90.91
[   bao] Results(  235 /   237 Samples) : 59.15  75.32  82.98  85.53  90.21  91.91  93.19  93.19  94.04  95.74 ----  4.26
[ zhuan] Results(   93 /    93 Samples) : 35.48  47.31  49.46  53.76  55.91  59.14  62.37  62.37  66.67  68.82 ---- 31.18
[   man] Results(   72 /    72 Samples) : 23.61  43.06  55.56  58.33  63.89  70.83  76.39  80.56  83.33  84.72 ---- 15.28
[  shen] Results(  190 /   190 Samples) : 64.21  78.95  83.16  89.47  92.11  94.74  97.37  97.37  97.89  98.95 ----  1.05
[   dai] Results(  224 /   224 Samples) : 34.82  63.39  72.77  79.02  86.61  90.18  93.30  94.64  94.64  95.09 ----  4.91
[  tuan] Results(   90 /    90 Samples) : 54.44  64.44  72.22  74.44  76.67  78.89  80.00  80.00  80.00  82.22 ---- 17.78
[   yan] Results(  259 /   259 Samples) : 57.53  82.24  90.73  94.59  96.53  96.91  97.30  98.07  98.46  98.46 ----  1.54
[    an] Results(  209 /   210 Samples) : 61.24  74.16  79.90  84.69  87.08  89.00  90.43  92.34  92.82  93.30 ----  6.70
[   xiu] Results(   46 /    46 Samples) :  4.35  28.26  56.52  63.04  67.39  71.74  73.91  78.26  86.96  86.96 ---- 13.04
[    mo] Results(   93 /    95 Samples) : 51.61  56.99  62.37  69.89  73.12  77.42  78.49  80.65  81.72  82.80 ---- 17.20
[    na] Results(  127 /   127 Samples) : 62.20  81.89  86.61  89.76  92.13  93.70  94.49  95.28  96.06  96.85 ----  3.15
[   que] Results(   74 /    76 Samples) : 32.43  54.05  63.51  72.97  82.43  87.84  89.19  91.89  95.95  95.95 ----  4.05
[    ao] Results(  148 /   148 Samples) : 60.14  68.92  81.08  85.81  87.16  89.19  93.24  93.24  94.59  95.27 ----  4.73
[   yun] Results(  150 /   150 Samples) : 52.00  62.00  70.00  74.67  78.67  84.00  86.00  86.00  87.33  88.00 ---- 12.00
[   jue] Results(  117 /   117 Samples) : 58.12  71.79  77.78  87.18  89.74  91.45  92.31  92.31  93.16  93.16 ----  6.84
[    ta] Results(  399 /   400 Samples) : 78.20  91.48  94.24  96.24  97.49  97.99  98.75  99.00  99.00  99.50 ----  0.50
[  jiao] Results(  258 /   259 Samples) : 56.98  81.01  86.43  91.09  94.19  95.35  95.35  95.74  95.74  96.51 ----  3.49
[    ma] Results(  109 /   109 Samples) : 51.38  62.39  73.39  77.06  77.98  84.40  86.24  87.16  88.99  90.83 ----  9.17
[   kou] Results(   91 /    91 Samples) : 39.56  56.04  64.84  71.43  80.22  83.52  83.52  86.81  87.91  87.91 ---- 12.09
[   wan] Results(  212 /   212 Samples) : 37.26  58.96  71.23  75.47  80.66  84.43  88.21  90.57  92.45  93.40 ----  6.60
[   mei] Results(  267 /   268 Samples) : 73.78  86.89  94.01  95.51  97.38  97.75  98.50  98.50  98.88  98.88 ----  1.12
[zhuang] Results(   81 /    81 Samples) : 35.80  50.62  60.49  70.37  79.01  81.48  81.48  82.72  85.19  87.65 ---- 12.35
[  feng] Results(  158 /   158 Samples) : 69.62  86.08  87.97  91.14  93.67  94.30  94.94  96.20  96.84  96.84 ----  3.16
[  shan] Results(  116 /   116 Samples) : 47.41  63.79  74.14  82.76  86.21  87.93  88.79  88.79  90.52  93.10 ----  6.90
[  bang] Results(   56 /    56 Samples) :  1.79  12.50  21.43  23.21  32.14  39.29  44.64  44.64  46.43  51.79 ---- 48.21
[   cuo] Results(   19 /    19 Samples) :  0.00  21.05  31.58  36.84  42.11  42.11  42.11  42.11  47.37  47.37 ---- 52.63
[   luo] Results(   69 /    69 Samples) : 30.43  47.83  56.52  63.77  68.12  72.46  76.81  81.16  82.61  84.06 ---- 15.94
[  hang] Results(   71 /    71 Samples) : 50.70  61.97  70.42  74.65  74.65  77.46  84.51  84.51  87.32  88.73 ---- 11.27
[    ni] Results(   98 /    99 Samples) : 43.88  57.14  72.45  76.53  81.63  85.71  88.78  91.84  93.88  94.90 ----  5.10
[   fen] Results(  264 /   264 Samples) : 71.21  82.20  85.61  87.50  89.39  90.53  91.29  91.67  92.42  93.94 ----  6.06
[    pi] Results(  122 /   122 Samples) : 54.10  73.77  79.51  85.25  88.52  93.44  94.26  94.26  95.08  95.90 ----  4.10
[   gen] Results(   66 /    66 Samples) : 36.36  50.00  60.61  65.15  68.18  69.70  72.73  77.27  81.82  87.88 ---- 12.12
[  meng] Results(   50 /    50 Samples) : 32.00  38.00  46.00  56.00  62.00  64.00  64.00  66.00  70.00  72.00 ---- 28.00
```

```
[    te] Results(   87 /    87 Samples) : 49.43  60.92  75.86  80.46  82.76  85.06  85.06  85.06  85.06  86.21 ---- 13.79
[    ka] Results(   23 /    23 Samples) : 17.39  21.74  34.78  43.48  52.17  56.52  60.87  65.22  73.91  78.26 ---- 21.74
[  chen] Results(   59 /    60 Samples) : 25.42  38.98  49.15  62.71  64.41  69.49  71.19  71.19  71.19  76.27 ---- 23.73
[ zhang] Results(  225 /   225 Samples) : 61.78  75.56  82.22  88.44  92.89  95.56  95.56  97.33  98.67  99.11 ----  0.89
[  zhun] Results(   27 /    28 Samples) :  7.41   7.41  14.81  25.93  33.33  37.04  37.04  37.04  40.74  40.74 ---- 59.26
[  guai] Results(   18 /    18 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[  zhao] Results(   87 /    89 Samples) : 21.84  51.72  65.52  73.56  77.01  79.31  85.06  86.21  87.36  89.66 ---- 10.34
[  quan] Results(  244 /   245 Samples) : 79.92  90.16  92.62  94.67  95.49  95.90  96.31  96.72  97.54  97.54 ----  2.46
[   jie] Results(  486 /   486 Samples) : 69.75  86.01  90.95  95.68  97.74  97.94  98.15  98.35  98.56  98.56 ----  1.44
[  guan] Results(  435 /   435 Samples) : 75.86  91.03  94.94  95.86  96.78  97.47  97.93  98.62  98.62  98.85 ----  1.15
[   fan] Results(  145 /   146 Samples) : 56.55  73.79  79.31  84.83  87.59  89.66  91.03  91.72  93.79  95.17 ----  4.83
[   chi] Results(  153 /   153 Samples) : 35.95  59.48  87.58  91.50  92.81  93.46  94.12  94.12  94.77  95.42 ----  4.58
[   ran] Results(  104 /   104 Samples) : 31.73  48.08  65.38  71.15  75.96  79.81  83.65  88.46  89.42  91.35 ----  8.65
[   mao] Results(  102 /   102 Samples) : 57.84  68.63  80.39  83.33  89.22  90.20  90.20  92.16  92.16  92.16 ----  7.84
[   xie] Results(  279 /   279 Samples) : 53.76  72.04  83.51  88.53  91.76  93.55  94.27  95.70  96.06  96.42 ----  3.58
[ chuan] Results(  128 /   128 Samples) : 54.69  67.19  74.22  78.13  81.25  82.81  84.38  84.38  88.28  89.84 ---- 10.16
[   bie] Results(   61 /    61 Samples) : 18.03  29.51  39.34  42.62  44.26  47.54  47.54  49.18  54.10  55.74 ---- 44.26
[   can] Results(   89 /    90 Samples) : 60.67  82.02  87.64  88.76  91.01  94.38  96.63  96.63  96.63  96.63 ----  3.37
[ qiang] Results(  109 /   109 Samples) : 64.22  75.23  90.83  93.58  96.33  98.17  98.17  98.17  98.17  99.08 ----  0.92
[  hong] Results(   74 /    74 Samples) : 60.81  77.03  86.49  91.89  94.59  95.95  95.95  98.65  98.65  98.65 ----  1.35
[   zou] Results(   51 /    51 Samples) : 19.61  33.33  45.10  54.90  58.82  60.78  64.71  68.63  74.51  78.43 ---- 21.57
[  nong] Results(  105 /   106 Samples) : 57.14  78.10  82.86  89.52  93.33  94.29  94.29  95.24  95.24  96.19 ----  3.81
[   suo] Results(  101 /   101 Samples) : 59.41  79.21  90.10  92.08  94.06  96.04  97.03  98.02  98.02  98.02 ----  1.98
[ sheng] Results(  412 /   412 Samples) : 61.41  82.52  89.08  92.72  94.66  96.12  96.36  96.60  97.09  97.33 ----  2.67
[  gang] Results(   73 /    73 Samples) :  9.59  20.55  32.88  38.36  47.95  49.32  52.05  57.53  60.27  61.64 ---- 38.36
[    sa] Results(   43 /    43 Samples) : 48.84  65.12  74.42  76.74  79.07  81.40  83.72  88.37  88.37  88.37 ---- 11.63
[   lan] Results(   61 /    61 Samples) : 14.75  32.79  44.26  47.54  55.74  59.02  62.30  62.30  63.93  65.57 ---- 34.43
[   zan] Results(   19 /    19 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[  geng] Results(   64 /    65 Samples) : 39.06  56.25  62.50  70.31  78.13  81.25  85.94  85.94  89.06  89.06 ---- 10.94
[  liao] Results(   76 /    76 Samples) : 15.79  39.47  53.95  61.84  67.11  71.05  73.68  76.32  77.63  80.26 ---- 19.74
[   hao] Results(  151 /   151 Samples) : 72.85  87.42  93.38  95.36  95.36  96.69  96.69  98.01  98.68  98.68 ----  1.32
[ xiong] Results(   43 /    43 Samples) : 34.88  48.84  65.12  65.12  72.09  74.42  74.42  74.42  74.42  74.42 ---- 25.58
[   nai] Results(   13 /    13 Samples) :  0.00   0.00   0.00   7.69   7.69   7.69   7.69   7.69   7.69  15.38 ---- 84.62
[    cu] Results(   41 /    41 Samples) : 26.83  36.59  48.78  56.10  56.10  60.98  63.41  63.41  68.29  68.29 ---- 31.71
[   rao] Results(   11 /    11 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[    xu] Results(  176 /   176 Samples) : 53.41  80.11  90.91  93.75  95.45  97.16  98.30  98.30  98.30  98.30 ----  1.70
[   rui] Results(   19 /    19 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[   gou] Results(   59 /    59 Samples) : 16.95  30.51  44.07  47.46  55.93  59.32  61.02  62.71  66.10  67.80 ---- 32.20
[  qing] Results(  251 /   251 Samples) : 63.35  78.88  87.25  92.03  93.63  94.82  95.62  96.81  97.21  97.21 ----  2.79
[  kong] Results(   63 /    64 Samples) :  9.52  28.57  57.14  68.25  76.19  82.54  82.54  84.13  90.48  92.06 ----  7.94
[    gu] Results(  170 /   170 Samples) : 60.59  80.00  84.71  89.41  92.35  92.94  95.88  96.47  97.06  97.06 ----  2.94
[  zhua] Results(   14 /    14 Samples) :  0.00   7.14   7.14   7.14   7.14   7.14  14.29  21.43  21.43  21.43 ---- 78.57
[   gui] Results(   87 /    87 Samples) : 48.28  67.82  81.61  85.06  87.36  88.51  91.95  91.95  91.95  91.95 ----  8.05
[    ce] Results(   66 /    66 Samples) : 46.97  57.58  65.15  69.70  74.24  77.27  78.79  81.82  86.36  86.36 ---- 13.64
[  chao] Results(   35 /    35 Samples) : 42.86  54.29  65.71  68.57  71.43  77.14  77.14  82.86  85.71  85.71 ---- 14.29
[  shao] Results(   80 /    80 Samples) : 33.75  56.25  67.50  70.00  75.00  78.75  81.25  85.00  87.50  91.25 ----  8.75
[    ai] Results(  102 /   102 Samples) : 46.08  60.78  69.61  75.49  81.37  83.33  86.27  88.24  89.22  91.18 ----  8.82
[   hei] Results(   38 /    38 Samples) : 36.84  50.00  52.63  65.79  71.05  76.32  76.32  78.95  81.58  86.84 ---- 13.16
[   sui] Results(   57 /    57 Samples) : 21.05  49.12  59.65  66.67  70.18  70.18  75.44  75.44  75.44  75.44 ---- 24.56
[   mai] Results(   56 /    56 Samples) : 21.43  44.64  53.57  60.71  62.50  71.43  78.57  80.36  80.36  82.14 ---- 17.86
[   dun] Results(   30 /    30 Samples) :  6.67  10.00  16.67  23.33  36.67  40.00  43.33  53.33  53.33  60.00 ---- 40.00
[   jun] Results(  121 /   121 Samples) : 50.41  61.16  71.90  78.51  82.64  87.60  90.08  90.91  92.56  92.56 ----  7.44
[   mou] Results(   38 /    40 Samples) : 13.16  23.68  34.21  42.11  44.74  47.37  47.37  50.00  50.00  50.00 ---- 50.00
[   zao] Results(  118 /   119 Samples) : 55.08  72.88  76.27  80.51  83.90  85.59  88.14  88.98  89.83  94.07 ----  5.93
[  chan] Results(  163 /   163 Samples) : 69.94  82.82  87.12  93.87  96.32  97.55  98.16  98.16  98.16  98.77 ----  1.23
[ guang] Results(   79 /    79 Samples) : 65.82  77.22  81.01  87.34  87.34  87.34  89.87  91.14  91.14  93.67 ----  6.33
[  song] Results(   44 /    44 Samples) :  6.82  22.73  61.36  68.18  75.00  77.27  81.82  81.82  81.82  81.82 ---- 18.18
[    me] Results(   32 /    32 Samples) :  0.00   3.13   6.25   6.25   9.38   9.38  12.50  12.50  12.50  15.63 ---- 84.38
[ chong] Results(   69 /    69 Samples) : 40.58  59.42  60.87  68.12  73.91  82.61  82.61  82.61  88.41  92.75 ----  7.25
[  kang] Results(   16 /    16 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[   qiu] Results(  104 /   104 Samples) : 71.15  87.50  88.46  93.27  94.23  96.15  96.15  97.12  98.08  99.04 ----  0.96
[  kuai] Results(  102 /   102 Samples) : 52.94  60.78  75.49  78.43  79.41  84.31  86.27  87.25  90.20  92.16 ----  7.84
[    si] Results(  402 /   402 Samples) : 67.66  87.31  97.76  98.51  99.25  99.25  99.25  99.50  99.50  99.50 ----  0.50
[    nu] Results(   48 /    48 Samples) : 47.92  52.08  60.42  64.58  72.92  75.00  77.08  77.08  79.17  81.25 ---- 18.75
[   nei] Results(  108 /   108 Samples) : 53.70  66.67  75.00  84.26  85.19  85.19  88.89  90.74  91.67  91.67 ----  8.33
[   han] Results(   60 /    60 Samples) : 35.00  56.67  70.00  75.00  81.67  85.00  91.67  91.67  95.00  96.67 ----  3.33
[   tie] Results(   60 /    60 Samples) : 46.67  61.67  71.67  73.33  76.67  78.33  80.00  85.00  85.00  86.67 ---- 13.33
[    nv] Results(   46 /    46 Samples) : 34.78  39.13  47.83  54.35  60.87  65.22  65.22  65.22  65.22  67.39 ---- 32.61
[   kan] Results(   63 /    63 Samples) : 30.16  42.86  57.14  65.08  69.84  79.37  79.37  84.13  85.71  85.71 ---- 14.29
[  qiao] Results(   26 /    27 Samples) :  0.00   3.85  11.54  15.38  15.38  19.23  30.77  46.15  53.85  57.69 ---- 42.31
[ kuang] Results(   48 /    48 Samples) : 14.58  22.92  22.92  22.92  27.08  37.50  43.75  45.83  50.00  50.00 ---- 50.00
[  reng] Results(   26 /    28 Samples) :  7.69   7.69   7.69   7.69   7.69   7.69   7.69   7.69   7.69   7.69 ---- 92.31
[    tu] Results(  109 /   110 Samples) : 69.72  79.82  84.40  87.16  89.91  90.83  91.74  93.58  95.41  96.33 ----  3.67
[    lv] Results(   98 /    98 Samples) : 50.00  61.22  71.43  77.55  82.65  84.69  87.76  87.76  87.76  89.80 ---- 10.20
[    la] Results(   77 /    77 Samples) : 24.68  35.06  42.86  57.14  66.23  76.62  80.52  84.42  87.01  89.61 ---- 10.39
[   pan] Results(   56 /    56 Samples) : 10.71  17.86  28.57  32.14  33.93  42.86  48.21  51.79  53.57  55.36 ---- 44.64
[    po] Results(   87 /    88 Samples) : 47.13  59.77  66.67  72.41  78.16  80.46  81.61  82.76  85.06  87.36 ---- 12.64
```

[huai] Results(	18 /	18 Samples) :	0.00	0.00	5.56	22.22	22.22	22.22	22.22	22.22	22.22	22.22	----	77.78
[a] Results(	134 /	134 Samples) :	31.34	50.00	61.94	68.66	72.39	83.58	85.07	85.82	86.57	87.31	----	12.69
[se] Results(	113 /	113 Samples) :	47.79	65.49	71.68	78.76	84.07	86.73	92.04	92.92	92.92	92.92	----	7.08
[fou] Results(	15 /	15 Samples) :	0.00	0.00	6.67	6.67	6.67	6.67	6.67	6.67	6.67	6.67	----	93.33
[ceng] Results(	41 /	41 Samples) :	9.76	19.51	26.83	36.59	36.59	41.46	53.66	56.10	56.10	58.54	----	41.46
[piao] Results(	39 /	39 Samples) :	15.38	25.64	30.77	41.03	46.15	56.41	58.97	64.10	69.23	74.36	----	25.64
[wa] Results(	42 /	42 Samples) :	2.38	2.38	4.76	14.29	19.05	26.19	40.48	47.62	54.76	59.52	----	40.48
[zhui] Results(	23 /	23 Samples) :	4.35	8.70	8.70	17.39	26.09	26.09	34.78	34.78	39.13	39.13	----	60.87
[cha] Results(	105 /	105 Samples) :	61.90	75.24	80.95	82.86	86.67	89.52	90.48	91.43	92.38	93.33	----	6.67
[cun] Results(	112 /	112 Samples) :	75.89	81.25	86.61	90.18	91.96	92.86	93.75	93.75	94.64	94.64	----	5.36
[long] Results(	65 /	65 Samples) :	23.08	33.85	53.85	63.08	67.69	76.92	83.08	87.69	89.23	90.77	----	9.23
[huang] Results(	21 /	21 Samples) :	0.00	9.52	19.05	28.57	33.33	38.10	38.10	38.10	38.10	38.10	----	61.90
[miao] Results(	26 /	26 Samples) :	7.69	30.77	30.77	34.62	38.46	42.31	42.31	46.15	46.15	46.15	----	53.85
[pin] Results(	118 /	118 Samples) :	52.54	72.03	79.66	83.90	83.90	86.44	89.83	89.83	90.68	92.37	----	7.63
[sha] Results(	60 /	60 Samples) :	53.33	76.67	86.67	95.00	98.33	98.33	100.00	100.00	100.00	100.00	----	0.00
[pu] Results(	68 /	68 Samples) :	29.41	42.65	55.88	57.35	63.24	67.65	73.53	80.88	85.29	85.29	----	14.71
[cang] Results(	26 /	26 Samples) :	3.85	3.85	3.85	7.69	15.38	19.23	23.08	26.92	30.77	30.77	----	69.23
[ruo] Results(	19 /	19 Samples) :	5.26	5.26	5.26	5.26	5.26	5.26	5.26	5.26	10.53	10.53	----	89.47
[chai] Results(	10 /	10 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00
[gua] Results(	17 /	17 Samples) :	5.88	11.76	29.41	29.41	29.41	29.41	29.41	35.29	47.06	47.06	----	52.94
[qun] Results(	48 /	48 Samples) :	45.83	54.17	68.75	77.08	77.08	81.25	85.42	85.42	89.58	89.58	----	10.42
[sao] Results(	17 /	17 Samples) :	0.00	0.00	11.76	23.53	35.29	35.29	41.18	41.18	41.18	41.18	----	58.82
[dei] Results(	8 /	8 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00
[zuan] Results(	8 /	8 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00
[pao] Results(	24 /	24 Samples) :	4.17	8.33	8.33	8.33	8.33	8.33	12.50	16.67	29.17	29.17	----	70.83
[tang] Results(	18 /	18 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.56	----	94.44
[bin] Results(	14 /	14 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00
[ku] Results(	68 /	68 Samples) :	45.59	66.18	72.06	77.94	83.82	86.76	91.18	91.18	92.65	95.59	----	4.41
[kuan] Results(	45 /	45 Samples) :	13.33	17.78	22.22	24.44	33.33	35.56	44.44	46.67	51.11	57.78	----	42.22
[rang] Results(	38 /	38 Samples) :	0.00	5.26	5.26	13.16	13.16	13.16	15.79	23.68	26.32	34.21	----	65.79
[sun] Results(	32 /	32 Samples) :	21.88	34.38	46.88	46.88	56.25	59.38	59.38	59.38	59.38	65.63	----	34.38
[gei] Results(	56 /	56 Samples) :	17.86	21.43	32.14	42.86	51.79	57.14	62.50	67.86	73.21	76.79	----	23.21
[diao] Results(	47 /	47 Samples) :	4.26	14.89	21.28	25.53	29.79	36.17	44.68	48.94	53.19	55.32	----	44.68
[rou] Results(	29 /	29 Samples) :	6.90	13.79	20.69	24.14	27.59	34.48	41.38	55.17	55.17	65.52	----	34.48
[pa] Results(	18 /	18 Samples) :	0.00	5.56	11.11	11.11	16.67	16.67	27.78	27.78	27.78	27.78	----	72.22
[chun] Results(	10 /	10 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00
[zha] Results(	24 /	24 Samples) :	0.00	0.00	4.17	12.50	16.67	20.83	25.00	29.17	29.17	29.17	----	70.83
[lang] Results(	29 /	29 Samples) :	0.00	0.00	0.00	3.45	3.45	3.45	3.45	6.90	6.90	6.90	----	93.10
[niang] Results(	24 /	24 Samples) :	0.00	0.00	0.00	4.17	4.17	4.17	4.17	4.17	4.17	4.17	----	95.83
[ha] Results(	16 /	16 Samples) :	0.00	0.00	0.00	6.25	12.50	12.50	12.50	12.50	18.75	18.75	----	81.25
[luan] Results(	8 /	8 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00
[lue] Results(	19 /	19 Samples) :	5.26	5.26	5.26	15.79	15.79	15.79	15.79	15.79	21.05	21.05	----	78.95
[lou] Results(	36 /	36 Samples) :	11.11	11.11	13.89	16.67	25.00	27.78	30.56	30.56	30.56	30.56	----	69.44
[gun] Results(	10 /	10 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00
[niao] Results(	13 /	13 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00
[kua] Results(	10 /	10 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00
[zhai] Results(	29 /	29 Samples) :	0.00	0.00	3.45	13.79	13.79	13.79	13.79	17.24	17.24	17.24	----	82.76
[ca] Results(	16 /	16 Samples) :	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25	----	93.75
[die] Results(	15 /	15 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00
[za] Results(	24 /	24 Samples) :	12.50	16.67	20.83	33.33	41.67	41.67	50.00	54.17	54.17	54.17	----	45.83
[zen] Results(	7 /	7 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00
[nao] Results(	29 /	29 Samples) :	3.45	10.34	10.34	10.34	20.69	27.59	27.59	31.03	37.93	37.93	----	62.07
[sang] Results(	8 /	8 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00
[sen] Results(	19 /	19 Samples) :	0.00	0.00	0.00	5.26	15.79	15.79	21.05	26.32	26.32	31.58	----	68.42
[zang] Results(	18 /	18 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.56	5.56	5.56	----	94.44
[weng] Results(	11 /	11 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00
[nuo] Results(	12 /	12 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00
[nuan] Results(	11 /	11 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00
[diu] Results(	6 /	6 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00
[mie] Results(	15 /	15 Samples) :	0.00	0.00	0.00	6.67	6.67	6.67	6.67	13.33	13.33	20.00	----	80.00
[ang] Results(	17 /	17 Samples) :	0.00	0.00	0.00	0.00	11.76	11.76	29.41	29.41	35.29	35.29	----	64.71
[nie] Results(	4 /	4 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00
[cen] Results(	8 /	8 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00
[pang] Results(	11 /	11 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00
[chuai] Results(	4 /	4 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00
[ga] Results(	7 /	7 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00
[hun] Results(	16 /	16 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00
[niu] Results(	21 /	21 Samples) :	9.52	9.52	9.52	38.10	38.10	42.86	47.62	47.62	52.38	52.38	----	47.62
[juan] Results(	8 /	8 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00
[chou] Results(	15 /	15 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00
[ne] Results(	10 /	10 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	10.00	10.00	10.00	----	90.00
[leng] Results(	11 /	11 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00
[kui] Results(	12 /	12 Samples) :	8.33	8.33	16.67	16.67	25.00	33.33	33.33	50.00	50.00	50.00	----	50.00
[chui] Results(	4 /	4 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00
[miu] Results(	4 /	4 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00
[cuan] Results(	4 /	4 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00
[chuo] Results(	5 /	5 Samples) :	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	----	100.00

```
[   nue] Results(    5 /     5 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[   pie] Results(    4 /     4 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[   run] Results(    6 /     6 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[  seng] Results(    4 /     4 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[  shai] Results(    4 /     4 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[ shuai] Results(   10 /    10 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00  10.00  10.00  10.00  10.00 ---- 90.00
[   tun] Results(    4 /     4 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[ zhuai] Results(    4 /     4 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[  shun] Results(   10 /    10 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00  10.00  10.00  10.00 ---- 90.00
[   zun] Results(    9 /     9 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[ qiong] Results(    4 /     4 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[  keng] Results(    4 /     4 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[   cui] Results(    6 /     6 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[  mang] Results(   10 /    10 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00  10.00  10.00  10.00  10.00 ---- 90.00
[   pen] Results(    7 /     7 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[   qia] Results(    8 /     8 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[   dia] Results(    6 /     6 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[  nang] Results(    3 /     3 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[   cou] Results(    2 /     3 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[   nin] Results(    3 /     4 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[   pou] Results(    3 /     3 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[   sou] Results(    7 /     7 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[  shua] Results(    3 /     3 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[ shuan] Results(    3 /     3 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[    fo] Results(   14 /    14 Samples) :  0.00   7.14  21.43  21.43  21.43  21.43  21.43  21.43  21.43  21.43 ---- 78.57
[   nen] Results(    4 /     4 Samples) :  0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00 ----100.00
[Total] (  70462 Valid Samples):  62.60  77.15  83.37  86.75  89.02  90.62  91.69  92.54  93.27  93.88 ----  6.12
```

2. 口呼汉语语句识别结果举例。

(原句) # 中华人民共和国 具有 悠久 的 历史 。
(错误词数/总词数) 0/6 中华人民共和国 具有 悠久 的 历史 。

清华大学 计算机科学与技术系 语音实验室 开发 了 许多 语音识别 的 产品 。
2/10 清华大学 计算机科学与技术系 语音实验室 开放 的 许多 语音识别 的 产品 。

经 专家 鉴定 ， 其 性能 指标 达到 了 国际 先进 水平 。
4/13 经 专家 前面 ， 其 性能 指标 达到 的 国际 先进 食品 取得

中国 一贯 维护 奥林匹克 的 宗旨 和 原则 。
0/9 中国 一贯 维护 奥林匹克 的 宗旨 和 原则 。

阿富汗 国防部 发言人 说 ， 袭击 是 由 忠于 阿富汗 总理 的 武装 发起 的 。
3/16 阿富汗 国防部 发言人 少 ， 迄今 是 要 忠于 阿富汗 总理 的 武装 发起 的 。

直接 原因 是 俄罗斯 联邦 政府 停止 了 干预 外汇 交易 市场 。
0/13 直接 原因 是 俄罗斯 联邦 政府 停止 了 干预 外汇 交易 市场 。

过去 是 农村 依靠 城市 ， 现在 则 是 城乡 相互 依靠 、 相互 渗透 。
3/16 过去 是 农村 以后 城市 ， 现在 则 是 长江 相互 一个 、 相互 渗透 。

中国 一贯 积极 支持 南非 人民 的 正义 斗争 ， 支持 南非 通过 谈判 和平 解决 问题 。
2/18 中国 一贯 积极 支持 南非 人民币 的 正义 斗争 ， 直至 南非 通过 谈判 和平 解决 问题 。

美国 期货 市场 的 会员 席位 个人 都 可以 获得 。
1/11 美国 期货 市场 的 会员 西北 个人 都 可以 获得 。

新 产品 在 防治 棉铃虫 方面 居 国内 先进 水平 ， 填补 了 国内 农药 空白 。
3/17 新 产品 在 防治 棉铃虫 方面 居 国内 先进 水平 ， 填补 的 宏伟 目标 空白 。

报名 参加 比赛 的 代表团 已经 到位 ， 台湾 航空 模型 运动 队 也 来到 成都 参加 。
3/18 报名 参加 比赛 的 代表团 一些 宝贵 ， 台湾 航空 模型 运动 队 也 来到 成都 参展 。

本 地区 和 世界 其他 地区 相比 ， 政治 比较 稳定 ， 经济 充满 活力 。
3/16 本 地区 和 世界 其他 地区 相比 ， 正式 递交 文明 ， 经济 充满 活力 。

发表(已接收)论文

1. 牟晓隆，胡起秀，吴文虎，“ 与文本无关的复合策略说话人辨识系统 ”，清华大学学报（自然科学版），1997 年第 37 卷，第 3 期第 16－19 页
2. 郑方，牟晓隆，徐明星，武健，宋战江，“ 汉语语音听写机技术的研究与实现 ”，软件学报，1998 (已接收)
3. Xiaolong Mou, Qixiu Hu, Wenhu Wu, “Text-independent Speaker Identification Using Average Spectrum And GMM Approaches”, *Proc. International Conference on Multimodal Interface*, March 1996, Beijing.
4. 牟晓隆，詹津明，郑方，吴文虎，“ 基于修正退化频度估计算法的 n-gram 语言模型 ”，第五届全国人机语音通讯学术会议（NCMMSC’98）论文集，1998 年，哈尔滨
5. Zheng Fang, Xu Mingxing, Mou Xiaolong, Wu Jian,Wu Wenhu, Fang Ditang，“HarkMan - A Vocabulary-Independent Keyword Spotter for Spontaneous Chinese Speech”， *J. of Computer Science & Technology*, 1998 (Accepted)
6. Jin-ming Zhan, Xiaolong Mou, Shuqing Li, Ditang Fang, “A Language Model in a Large-Vocabulary Speech Recognition System”, *Oriental COCOSDA Workshop*, May 1998, JAPAN
7. 詹津明，牟晓隆，李树青，方棣棠，“ 一个大字表语音识别系统中的语言模型 ”，NCMMSC’98
8. 郑方，牟晓隆，徐明星，武健，宋战江，王重芳，吴文虎，“ 一个语文转换文本编辑器的实现 ”， NCMMSC’98

注: 1. 2. 5. 核心刊物
3. 6. 国际会议
4. 7. 8. 国内会议