电话语音识别中基于统计模型的动态通道 [1]

韩兆兵 张化云 张树武 徐 波

(中国科学院自动化研究所模式识别国家重点实验室 北京 100080)

摘 要: 与桌面环境相比，电话网络环境下的语音识别率仍然还比较低，为了推动电话语音识别在实际中的应用，提高其识别率成了当务之急. 先前的研究表明，电话语音识别率明显下降通常是因为测试和训练环境的电话通道不同引起数据失配造成的，因此该文提出基于统计模型的动态通道补偿算法 (SMDC) 减少它们之间的差异，采用贝叶斯估计算法动态地跟踪电话通道的时变特性. 实验结果表明，大词汇量连续语音识别的字误识率 (CER) 相对降低约 27%，孤立词的词误识率 (WER) 相对降低约 30%. 同时，算法的结构时延和计算复杂度也比较小，平均时延约 200 ms，可以很好地嵌入到实际电话语音识别应用中.

关键词: 电话语音识别，动态通道补偿，最大似然估计，最大后验估计

中图分类号: TP391.42 **文献标识码:** A **文章编号:** 1009-5896(2004)11-1714-07

Dynamic Channel Compensation Based on Statistical Model for Mandarin Speech Recognition over Telephone

Han Zhao-bing Zhang Hua-yun Zhang Shu-wu Xu Bo

(*National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing* 100080, *China*)

Abstract Automatic speech recognition in telecommunications environment still has a lower correct rate compared to its desktop pairs. Improving the performance of telephone-quality speech recognition is an urgent problem for its application in those practical fields. Previous works have shown that the main reason for this performance degradation is the var'ational mismatch caused by different telephone channels between the testing and training sets. In this paper, they propose an efficient implementation to dynamically compensate this mismatch based on a phone-conditioned prior statistic model for the channel bias. This algorithm uses Bayes' rule to estimate telephone channels and dynamically follows the time-variations within the channels. In their experiments on mandarin Large Vocabulary Continuous Speech Recognition (LVCSR) over telephone lines, the average Character Error Rate (CER) decreases more than 27% when applying this algorithm; in short utterance test, the Word-Error-Rate(WER) relatively reduced 30%. At the same time, the structural delay and computational consumptions required by this algorithm are limited. The average delay is about 200 ms. So it could be embedded into practical telephone-based applications.

Key words Telephone speech recognition, Dynamic channel compensation, Maximum-Likelihood (ML) estimation, Maximum A Posteriori (MAP) estimation

1 引言

众所周知，由于历史原因公用电话网络 (PSTN) 的语音带宽为 300~3400 Hz，这导致了电话语音识别性能的下降，但更为重要的是训练集和测试集电话通道的不匹配. 表 1 总结了 5 种通道条件下汉语语音识别率的变化. 这 5 种测试集的条件为: (1) 高质量的办公室下采集的

[1] 2003-06-12 收到，2004-03-23 改回
国家自然科学基金 (69835003) 资助项目

16 kHz 语音；(2) 用 DSP 工具包降采样的 8 kHz 语音；(3) 通过模拟电话线路的稳定线性滤波器采集的语音；(4) 在 PSTN 下录制的语音；(5) 通过 GSM 模拟器转录的语音. 可以发现，当通道为稳定线性时不变系统时，与 16 kHz 办公室条件下比较，CER 仅增加 2%—3%；但对于实际中的 PSTN 和 GSM 通道，识别率却下降约 9%—12%. 这很明显地表明 PSTN 和 GSM 通道不是理想的线性时不变系统，另外每次电话呼叫过程，都会产生不同的通道响应和噪声环境 [1]. 同时随着无线电话网络 (包括 GSM 和 CDMA) 的广泛应用，电话网络中的非线性也逐步增加，除此之外一些语音编码器 (例如 ETSI GSM 6.10) 也不是线性滤波器 [2,3]，这些非线性畸变进一步增加了问题的复杂性.

表 1　不同通道条件的 CER 比较

Database	语音带宽 (Hz)	CER (%)
基线系统	0–16000	12.2
降采样	0–8000	14.3
线性滤波	300–3400	15.4
PSTN 转录	300–3400	23.8
GSM 编码器	300–3400	21.5

最近几年，国际上提出许多电话通道衰减的补偿算法 [4−10]. 通常，电话通道被假定为线性时不变系统，这对于高质量的固定电话线路较适合，但对于低质量的长途电话和非线性线路 (如 GSM) 不精确，而且补偿效果有限. 因此，当电话线路中存在非稳定加性噪声和非线性畸变时，就不能简单地把它简化成线性时不变系统. 另外，这些算法通常需要整个呼叫过程结束后才可以进行通道补偿，增加了时延和计算复杂度，防碍在实际系统中的应用.

本文提出基于统计模型的动态通道补偿算法 (SMDC)，能够有效地补偿训练集和测试集中通道的差异. 第 2 节详细介绍基于 ML 准则 (ML-DC) 和 MAP(MAP-DC) 准则的 SMDC 算法；第 3 节是实验结果和分析；最后给出结论.

2 SMDC 算法

目前，被许多专家学者普遍接受的电话网络中的通道噪声模型如图 1 所示. 在时域，通道对于干净语音是卷积作用，但在 lg 域和倒谱域是加性作用. 一般地，通道 H_t 是一个非线性时变滤波器，背景噪声 N_t 为非稳态随机过程. 如文献 [7]，可以假设通道在短时信号分析中表现出近似线性的性质.

$$Y_t = X_t \cdot H_t + N_t \tag{1}$$

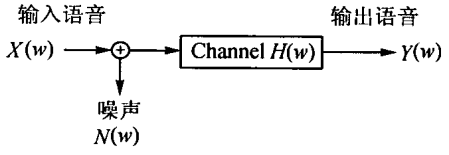

图 1　电话网络中的噪声环境模型

其中 X_t 是发送端的干净语音，Y_t 是接受端的带噪语音. 在倒谱域，这个电话模型又可以描述成下式：

$$\boldsymbol{y}_t = \boldsymbol{x}_t + \boldsymbol{b}_t \tag{2}$$

式中

$$\boldsymbol{b}_t = \lg(|H_t|^2) + \lg(1 + |N_t|^2/(|H_t|^2 \cdot |X_t|^2)) \tag{3}$$

通道偏置 b_t 如果满足如下两个条件可以认为是固定的常量：(1) $H_t(\omega)$ 时不变；(2) $N_t(\omega) << H_t(\omega) \cdot X_t(\omega)$. 然而在许多现实情况中，这两个条件很难满足. 但是在比较短的时间内 (通常

为几百毫秒内)，信噪比 SNR(SNR=$X_t(\omega)/N_t(\omega)$) 可认为是常量，同时频率响应 $H_t(\omega)$ 也可以假定为短时稳定量，这样通道偏置 $\boldsymbol{b}_t$ 也是常量. 因此，本文的理论框架是在短时窗内基于观测矢量和统计声学模型间的随机匹配得到时变通道参数，然后重构出与训练集相近的干净语音，参数估计准则包括 ML 和 MAP，及它们实验结果的对比.

2.1 ML 动态补偿

在 ML 准则下，偏置是根据时间点 t 为中心的短时窗 T 内的观测声学矢量 $\boldsymbol{O}_T$ 估计的.

$$\boldsymbol{O}_T = \{\boldsymbol{o}_{t-T/2}, \cdots, \boldsymbol{o}_t, \cdots, \boldsymbol{o}_{t+T/2}\} \tag{4}$$

为了避免复杂的前后向解码，用音子相关的码本 $\boldsymbol{\Omega}^S$ 代替 HMM 模型，码本的大小根据经验确定.

$$\boldsymbol{\Omega}^S = \{\boldsymbol{\omega}_n^S\}, \qquad n = 1, 2, \cdots, N \tag{5}$$

每个码字都是一个混合高斯模型 (GMM)：

$$\boldsymbol{\omega}_n^S = \{\alpha_{n,m}^S, \boldsymbol{\mu}_{n,m}^S, \boldsymbol{\Sigma}_{n,m}^S\}, \qquad m = 1, 2, \cdots, M \tag{6}$$

M 是每个 GMM 中的高斯个数. $\alpha_{n,m}^S, \boldsymbol{\mu}_{n,m}^S, \boldsymbol{\Sigma}_{n,m}^S$ 是第 (n,m) 个高斯分量的权重、均值和方差. 目标函数为

$$\hat{\boldsymbol{b}}_{t,\mathrm{ML}} = \arg\max_{\boldsymbol{b}_t}\{p(\boldsymbol{O}_T|\boldsymbol{\Omega}^S, \boldsymbol{b}_t)\} \tag{7}$$

可以用 EM 算法求解式 (7)，辅助函数定义为

$$Q_{\mathrm{ML}}(\hat{\boldsymbol{b}}_t|\boldsymbol{b}_t) = E\{\lg p(\boldsymbol{O}_T|\boldsymbol{\Omega}^S, \hat{\boldsymbol{b}}_t)|\boldsymbol{b}_t, \boldsymbol{\Omega}^S\} \tag{8}$$

通过递归的求解，最大似然准则下偏置 (ML-CD) 为

$$\hat{\boldsymbol{b}}_{t,\mathrm{ML}} = \left(\sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{m=1}^{M}\gamma(\boldsymbol{\Sigma}_{n,m}^S)^{-1}\right)^{-1} \bullet \left(\sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{m=1}^{M}\gamma(\boldsymbol{\Sigma}_{n,m}^S)^{-1}(\boldsymbol{o}_t - \boldsymbol{\mu}_{n,m}^S)\right) \tag{9}$$

此处，$\gamma = p(\boldsymbol{o}_t, n, m|\boldsymbol{\Omega}^S, \boldsymbol{b}_t)$ 是第 (n,m) 个高斯分量在 t 帧时的占有概率. 通过移动短时窗就可以估计通道偏置的时间变化轨迹.

显然在每个短时窗内多次递归的 EM 算法是相当耗时的，但幸好，EM 在每个音子相关的码本内收敛速度很快，如图 2 所示. 另外本文还研究了识别率随着递归次数的变化，结果表明当递归次数超过 2 后，识别率没有明显提高，因此在实验中可以对观测声学矢量 $\boldsymbol{Y}$ 在每个码本内仅估计一次. 一旦估计出通道偏置，通道的频率响应就可以重构出来. 图 3 是 2400 ms 语音的通道响应重构图，在高频和低频段都有明显的衰减，但在中间频率段是时变的，这也就是我们的假设依据.

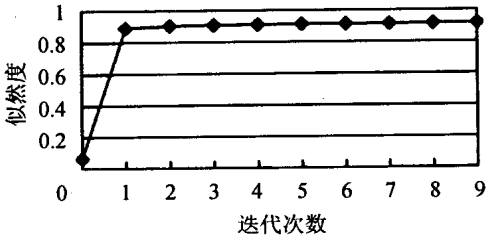

图 2 ML-DC 的收敛性

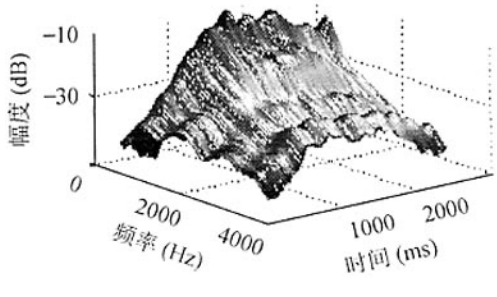

图 3 ML-DC 算法通道频率响应重构图

2.2 MAP 动态补偿

实验表明在 ML-DC 算法中最优的数据窗的长度约 800 ms，这就意味着至少有 400 ms 的时延. 如果缩短数据窗的长度，对于 ML 准则会出现数据稀疏问题，估计不鲁棒. 为了弥补 ML 准则的缺陷，我们引入 MAP 准则，用先验的通道统计模型减少估计误差.

文献 [9] 和实践经验表明，通道偏置对于不同的语音段具有明显的不同特性，因此可以用一系列 GMM 分布 $\boldsymbol{\Omega}^B$ 描述先验的通道偏置的统计信息.

$$\boldsymbol{\Omega}^B = \{\boldsymbol{\omega}_q^B\}, \qquad q = 1, 2, \cdots, Q \tag{10}$$

式中

$$\boldsymbol{\omega}_q^B = \{\alpha_{q,p}^B, \boldsymbol{\mu}_{q,p}^B, \boldsymbol{\Sigma}_{q,p}^B\}, \qquad p = 1, 2, \cdots, P \tag{11}$$

此处 Q 是 $\boldsymbol{\Omega}^B$ 中的所有模型数 (或者说码字)，P 是 GMM 的高斯个数. $\alpha_{q,p}^B$, $\boldsymbol{\mu}_{q,p}^B$, $\boldsymbol{\Sigma}_{q,p}^B$ 分别是第 (q,p) 个高斯的权重、均值和方差.

一旦取得先验的通道统计模型，MAP 动态通道补偿 (MAP-DC) 的公式就可以根据先验通道模型和声学的 GMM 模型在短时窗内最大化观测矢量得到. 在短时窗 T 内，基于 MAP 准则的偏置量 $\boldsymbol{b}_t$ 估计公式如下：

$$\hat{\boldsymbol{b}}_{t,\mathrm{MAP}} = \arg\max_{\boldsymbol{b}_t}\{p(\boldsymbol{O}_T|\boldsymbol{\Omega}^S, \boldsymbol{b}_t)p(\boldsymbol{b}_t|\boldsymbol{\Omega}^B)\} \tag{12}$$

EM 辅助函数定义为

$$\begin{aligned} Q_{\mathrm{MAP}}(\hat{\boldsymbol{b}}_t|\boldsymbol{b}_t) &= E\{\lg(p(\boldsymbol{O}_T|\boldsymbol{\Omega}^S, \hat{\boldsymbol{b}}_t)p(\hat{\boldsymbol{b}}_t|\boldsymbol{\Omega}^B))|\boldsymbol{\Omega}^S, \boldsymbol{b}_t\} \\ &= E\{\lg p(\boldsymbol{O}_T|\boldsymbol{\Omega}^S, \hat{\boldsymbol{b}}_t)|(\boldsymbol{\Omega}^S, \boldsymbol{b}_t\} + E\{\lg p(\hat{\boldsymbol{b}}_t|(\boldsymbol{\Omega}^B)|\boldsymbol{b}_t\} \end{aligned} \tag{13}$$

通过递归求解得到下式：

$$\begin{aligned} \hat{\boldsymbol{b}}_{t,\mathrm{MAP}} = &\left(\sum_{t=1}^{T}\left(\sum_{n=1}^{N}\sum_{m=1}^{M}\gamma(\boldsymbol{\Sigma}_{n,m}^S)^{-1} + \sum_{q=1}^{Q}\sum_{p=1}^{P}\rho(\boldsymbol{\Sigma}_{q,p}^B)^{-1}\right)\right)^{-1} \\ &\bullet\sum_{t=1}^{T}\left(\sum_{n=1}^{N}\sum_{m=1}^{M}\gamma(\boldsymbol{\Sigma}_{n,m}^S)^{-1}(\boldsymbol{o}_t - \boldsymbol{\mu}_{n,m}^S) + \sum_{q=1}^{Q}\sum_{p=1}^{P}\rho(\boldsymbol{\Sigma}_{q,p}^B)^{-1}\boldsymbol{\mu}_{q,p}^B\right) \end{aligned} \tag{14}$$

式中 $\gamma = p(\boldsymbol{o}_t, n, m|\boldsymbol{\Omega}^S, \boldsymbol{b}_t)$, $\rho = p(\boldsymbol{b}_t, p, q|\boldsymbol{\Omega}^B)$. γ 是第 (n,m) 个高斯在帧 t 时的占有概率，ρ 是 $\boldsymbol{\Omega}^B$ 中关于通道偏置 $\boldsymbol{b}_t$ 的第 (p,q) 个高斯先验概率.

在文献 [9]，只用到了 Viterbi 解码的最优状态和最近的先验通道统计量，而 MAP- DC 的两个平滑因子 γ 和 ρ，可以对几个相近的分布加权，换句话说 MAP-DC 采用软决策代替文献 [9] 的硬决策. 软决策可以弥补解码中的不正确的状态和不精确的先验通道统计引起的估计错误. 实验表明，软决策对于提高识别率还是比较有效的，但是会带来额外的计算量，不过可以通过调节 $\boldsymbol{\Omega}^S$ 和 $\boldsymbol{\Omega}^B$ 的大小控制计算量.

2.3 先验的通道偏置分布

在 MAP-DC 算法中，虽然能够克服 ML-DC 的数据稀疏问题，但也比较强地依赖先验的通道模型，因此精确的先验通道偏置分布是其关键问题. 为了得到包含足够的真实环境中的通道特性，我们收集了 30 多个小时电话语音，包含 3000 个不同的电话呼叫过程. 先验的统计分布可以用 ML-DC 对上述的数据统计得到.

一般，$\boldsymbol{\Omega}^B$ 可以不需要特殊的知识，用纯数据驱动的方法估计出. 所有的偏置矢量根据它们在声学空间上的距离进行聚类，当然也可以根据先验的问题集聚类. 所有的语音根据脚本对齐，偏置矢量根据不同的语音段聚类，例如声母、韵母、鼻音、滑音或者更为详细的音子分类或状态分类.

式 (3) 表明了偏置与通道、背景噪声和语音间的复杂关系. 如果 SNR 较高时, 偏置主要由通道决定; 如果 SNR 低时, 偏置由背景噪声和通道共同决定. 显然, 仅用单一的全局概率密度函数不足以精确描述. 另外以前的研究也表明, 区分有声段和无声段的统计量能够提高识别率 [9] . 因为汉语中不同的音子有明显的不同强度, 这促使我们分析不同的音子对通道的影响.

图 4 描述了在电话语音中不同的语音段的通道幅频响应的平均包络. 其中有 3 个声母段 ('y', 'sh', 'j'), 3 个韵母段 ('e', 'i', 'u') 和一个无声段. 这表明通道不仅在有声段和无声段有明显的区别, 而且对于不同的音子也有明显的不同特性.

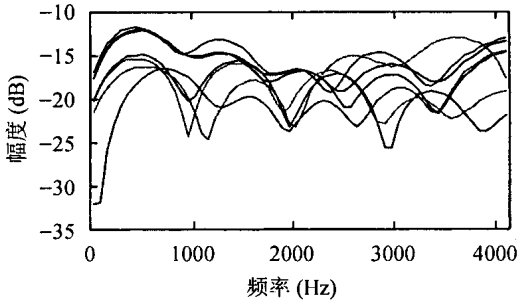

图 4 用 MAP-DC 算法分析对于不同音子的平均通道响应

图 5 为偏置分布的直方图分析, 分析了无声段、声母和韵母的第一和第二维偏置矢量. 显然, 仅用统一的 PDF 描述不同语音段的偏置变化, 会产生明显的错误, 所以本文中采用基于语音学知识的音子相关的通道模型.

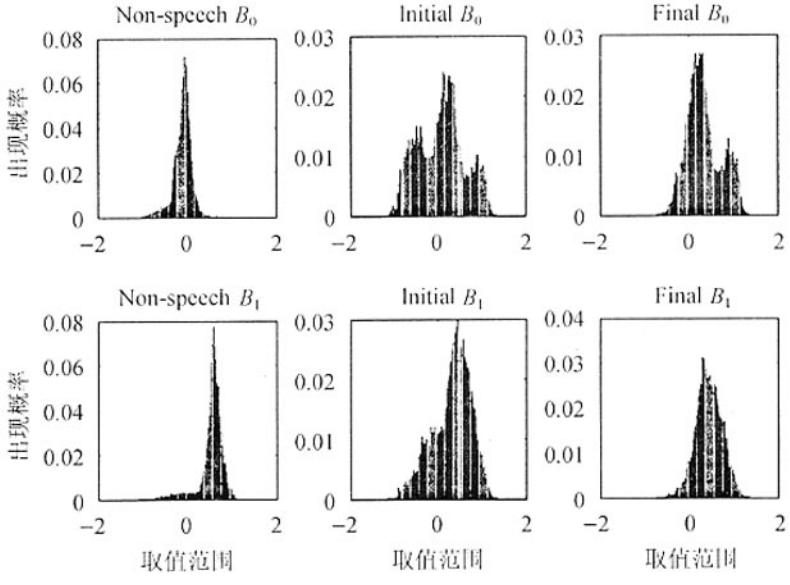

图 5 在不同类型语音段内通道偏置分布的直方图分析

为了更好地解释 MAP-DC 算法, 把 1.2 s 的干净语音通过固定电话网络转录. 用 MAP-DC 对转录后的 PSTN 电话语音估计出通道响应, 然后重构干净语音. 图 6 是干净语音、电话语音、电话通道及重构语音的频谱图.

3 实验结果及分析

为了验证所提出算法的有效性, 我们在 PSTN 和 GSM 电话通道上进行对比研究, 包括如下两个任务:

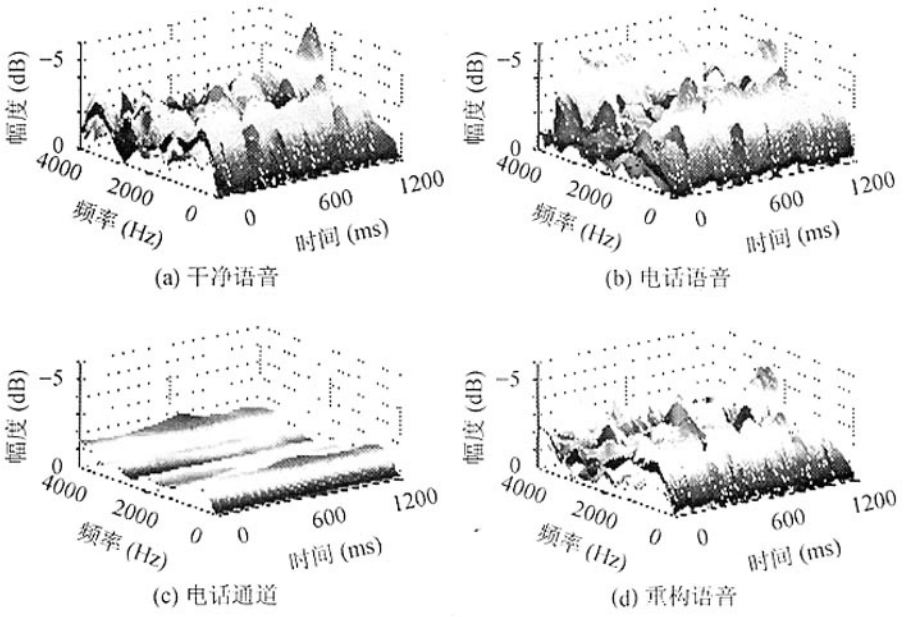

图 6　MAP-DC 算法中语音频谱分析

(1) 大词量连续电话语音识别— 863

词汇量约为 40,000 词．语言模型采用三元统计模型． 863 的测试集是在安静环境下录制的，包括 4 个说话人每人 60 句共 240 句话．为了得到电话质量的语音，把它们通过 PSTN 和 GSM 转录，并降采样得到电话语音的 863 测试集．每句包含 5~25 个汉字．

(2) 基于电话的在线股票信息查询— STOCK

词汇量约 1000 ．由于语言混淆性较小，语言模型没有用 N 元模型而是采用了语法相关的建模方法．约 1000 个不同的股票名称通过 10 个人在 PSTN 和 GSM 线路下录制的．每句包含 2~4 个汉字．

平台为中国科学院自动化所的 Pattek-Asr 3.1 软件包．声学模型是通过约 600 h 的语音库训练得到的，这些语音是通过高质量的麦克风录制，包括北京、武汉、哈尔滨等方言，并且脚本设计时经过了音子平衡． ML-DC 和 MAP-DC 中音子相关的码本也是用此语音库训练的．特征包括一维能量、一维基频、 12 维 MFCC 以及一阶和二阶差分．

长时倒谱归一化 (CMS) 作为所有实验的基本平台．本文比较了 ML-DC 和 MAP-DC 与其它常用的通道补偿技术的性能，这些技术包括 RASTA filtering[4], SBR(Signal Bias Removal)[5] 和 SM(Stochastic Matching)[2] ．表 2 给出了在 PSTN 和 GSM 条件下， 863 测试集的 CER ，STOCK 测试集的 WER ．

表 2　在不同任务下不同补偿技术的比较

方法	测试集			
	863PSTN (CER)	863GSM (CER)	STOCKPSTN (WER)	STOCKGSM (WER)
CMS	23.8	24.2	7.5	7.4
RASTA	24.6	27.2	7.4	7.6
SBR	20.4	24.0	5.4	6.5
SM	18.0	22.3	5.0	5.4
ML-DC	18.5	18.0	5.8	6.2
MAP-DC	16.8	16.0	3.5	4.0

表 2 表明，对于 863 任务 ML-DC 和 MAP-DC 都明显比其它方法有效，其中 MAP- DC 算法是最好的．尤其在短语识别 (STOCK) 中 MAP-DC 效果更为明显，这可能因为其测试语句比 863 任务通常要短的多， ML-DC 没有优势，而且在实验中发现 ML-DC 在这些句子的起始和结尾处常产生不正确的估计，如果引入先验的偏置分布限制 (MAP-DC) ，就可以明显减少错

误估计.

实验还表明，越精细的先验通道偏置分布，越能增加 MAP 估计的正确性，但同时也会引起计算量的增加。目前由于较有限的电话语音资源，很难估计出精细的通道分布。为了简化起见，令 Ω^S 和 Ω^B 大小一样，系统中包括 24 个声母、37 个韵母、一个静音和一个 garbage 共 63 个音子。Ω^S 和 Ω^B 分别包含 63 个 GMM 码本描述语音的统计量和通道偏置的统计量。

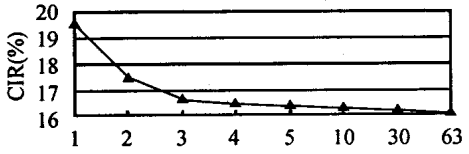

图 7 不同的 K 值识别率的变化

为了减少计算量，对观测矢量和模型中 GMM 的质心进行快速匹配，对于帧 t 语音 O_t，仅 K 个 Ω^S 中的 GMM 分布和 K 个 Ω^B 中通道分布参与运算。图 7 是 K 从 1 到 63 系统的性能变化，可以发现当 $K > 3$ 时识别率没有明显改善，因此在我们的实验中 K 取值为 3，这样可以在几乎不减少识别率的条件下明显减少计算量。

4 结论

本文提出了一个新颖的基于统计模型的动态通道补偿算法 (SMDC)，它有如下两个优点：(1) 中等规模的音子相关 GMM 代替了 HMM 模型 [6]，减少了计算量，识别性能明显提高；(2) 可以动态跟踪实际电话网络中的通道变化。SMDC 算法既可以应用在 PSTN 网络中，也可以应用在 GSM 网络中；既可以用 ML 准则，又可以用 MAP 准则；但 MAP 准则更为有效，尤其短语识别更为明显。虽然 MAP-DC 在本文的试验中优于 ML-DC，但是当先验的通道模型很难获得或者获得的不精确时，ML-DC 可能是更好的选择。实验结果表明 SMDC 算法与其它常用的通道补偿算法相比对于识别率的提高具有显著效果。

参 考 文 献

[1] Moreno P J, Siegler M A, Jain U, Stern R M. Continuous speech recognition of large vocabulary telephone quality speech. Proc. of the Eighth Spoken Language Systems Technology Workshop, Austin, Texas, 1995.

[2] Besacier L, Grassi S, Dufaux A, Ansorge M, Pellandini F. GSM speech coding and speaker recognition. Proc. of ICASSP 2000, Istanbul, Turkey, June 2000: 1085–1088.

[3] Huerta J M. Speech recognition in mobile environments. [Ph.D. Thesis]: School of Computer Science, Carnegie Mellon University, Apr. 2000.

[4] Hermansky H, Morgan N. RASTA processing of speech. *IEEE Trans. on Speech and Audio Processing*, 1994, 2(4): 578–589.

[5] Rahim M G, Juang Biing-Hwang. Signal bias removal by maximum likelihood estimation for robust telephone speech recognition. *IEEE Trans. on Speech and Audio Processing*, 1996, 4(1): 19–30.

[6] Sankar Ananth, Lee Chin-Hui. A maximum-likelihood approach to stochastic matching for robust speech recognition. *IEEE Trans. on Speech and Audio Processing*, 1996, 4(3): 190–202.

[7] Moreno P J. Speech recognition in noisy environments. [Ph.D. Thesis]: School of Computer Science, Carnegie Mellon University, April 22, 1996.

[8] Westphal M. The use of cepstral means in conversational speech recognition. Proc. of Eurospeech 97, Greece, 1997: 1143–1146.

[9] Chien Jen-Tzung, Wang Hsiao-Chuan, Lee Lee-Min. Estimation of channel bias for telephone speech recognition. In Proc. ICSLP'96, Philadelphia USA, 1996: 1840–1843.

[10] Veth J D, Boves L. Comparison of channel normalization techniques for automatic speech recognition over the phone. In Proc. ICSLP'96, Philadelphia USA, 1996: 2332–2335.

韩兆兵： 男，1978 年生，博士生，研究方向：语音识别、模式识别、声学建模.

张化云： 男，1974 年生，博士生，研究方向：信号处理、语音识别.

张树武： 男，1964 年生，副研究员，研究方向：语音识别、自然语言处理.

徐 波： 男，1966 年生，研究员，研究方向：信号处理、语音识别、自然语言处理.