学术论文

多带同步模型用于噪声环境下语音识别

孙　暐，吴镇扬
（东南大学无线电系，南京　210096）

[摘要]　根据人耳听觉特性，提出新的同步多带最大似然线性回归算法用于噪声环境下语音识别。该算法采用最大似然作为参数估计准则，利用各频带信号同步感知和噪声污染假定的方法进行语音模型补偿，有效地提高了识别系统在噪声环境下的识别性能。

[关键词]　隐马尔可夫模型；最大似然；多带同步模型；语音识别
[中图分类号]　TN912.34　　[文献标识码]　A　　[文章编号]　1009－1742（2006）03－0031－04

1　引言

由于语音信号易受到环境的影响，语音识别技术的环境鲁棒性引起了广泛研究[1~3]。目前语音识别环境鲁棒性采用的方法，一类是在识别恶化条件下，对参数进行优化估计，如噪声模型补偿方法[4]，随机匹配技术[5]，最大似然线性回归[6]（MLLR，maximum likelihood linear regression）；另一类是将人耳的感知特性应用于语音识别，如基于感知特性的特征参数[7,8]，独立子带分析[9,10]。

由于通常语音信号特征提取都要进行正交化处理，这会导致窄带噪声污染扩散，此外，识别对象以及噪声在不同频带上能量分布存在差异，因此，在噪声环境下，MLLR 算法采用全带单一线性变换的形式进行参数优化估计，过于简单，不能充分反映信号本身特性以及环境间的关系。而基于独立感知理论提出的多子带框架利用了信号的频谱特征[9]，但丢失了信号带间相关性的信息，且 Sangita 的子带算法（以下简称 S 算法）要求信号存在不受噪声影响的频带，而许多实际环境并不满足这一假设，实验表明，该算法在宽带噪声环境下识别性能较差，特别是在低信噪比环境下其性能的弱点更为突出。

分析可知，识别对象和噪声频率特性的差异，会对系统的识别性能产生很大的影响。Bregman 等听觉实验表明，人耳可以有选择地跟踪特定频率信号[11]。笔者结合听觉实验，提出同步多带最大似然线性回归算法（SMMLLR，synchronization multi-band maximum likelihood linear regression）。为便于信号处理上的简化，此前研究的子带分析[10]和多带分析（LMMLLR）都是采用对不同频率信号进行频带划分，各频带独立处理，而在 SMMLLR 中，采用噪声污染假定方法，在信号空间中进行多带分析来替代子带分析中各频带间独立感知假定，同时，在模型空间中对各频带信号采用帧同步分析。这种方法在利用有效频带进行识别的同时，引入带间相关性，有效地提高了整个识别系统在噪声环境下的识别性能，同时简化了模型。

2　算法分析

实验表明特定频带的信号受噪声污染严重，会导致该频带识别性能的急剧恶化。在自动语音识别中，为提高识别性能，通常在提取语音参数时要进行正交化处理（非正交的参数在纯净语音环境识别

[收稿日期]　2005－03－07；修回日期　2005－06－09
[基金项目]　国家自然科学基金资助项目（60272044）；“九七三”国家重点基础研究发展计划资助项目（2002CB312102）
[作者简介]　孙　暐（1974－），男，江苏吴江市人，东南大学博士研究生

性能相对较差），这又会导致噪声污染的扩散，使得整个识别系统性能恶化。人耳听觉实验表明，人可以跟踪特定频带信号加以识别。为此对信号空间采用噪声污染假定方法来减少噪声污染影响。为保留频带间信号的相关性，在信号空间进行信号合并，同时在模型空间各频带语音信号采用帧同步分析。为此，定义隐变量 f 表示互斥的频带合并策略。每个 f 排斥一种假定的噪声污染频带。$f \in [1, \cdots, 2^L - 1]$，其中 L 表示子带个数。设多带语音识别系统中噪声环境下的观测序列为 $\boldsymbol{Y} = \{\boldsymbol{y}_{ft},\ t \in [1, \cdots, T], f \in [1, \cdots, 2^L - 1]\}$，相应的纯净语音为 $\boldsymbol{X} = \{\boldsymbol{x}_{ft},\ t \in [1, \cdots, T],\ f \in [1, \cdots, 2^L - 1]\}$。语音信号模型采用连续密度隐 Markov 模型（CDHMM），模型参数为 $\Theta = \{\{a_{ij}\}, \{k_{jf}\}\{\boldsymbol{\mu}_{jfk}\}, \{\boldsymbol{\Sigma}_{jfk}\}\}$。由于采用帧同步分析，即在 HMM 中各频带信号同步转移，则多带模型中任一状态下信号观测概率密度为：

$$P(\boldsymbol{x}_t \mid s_t = j) = \sum_{f=1}^{2^L-1} \sum_{k=1}^{K} CP(k_t = k \mid s_t, f_t = f) \cdot N(\boldsymbol{x}_{ft} \mid \boldsymbol{\mu}_{jfk}, \boldsymbol{\Sigma}_{jfk}) \tag{1}$$

其中 $C = 1/(2^L - 1)$ 为常数，以保证概率归一性。$P(k_t = k \mid s_t, f_t)$ 为 t 时刻状态 s_t 频带 f_t 下第 k 个高斯混合密度的观测概率，$N(\boldsymbol{x}_{ft} \mid \boldsymbol{\mu}_{jfk}, \boldsymbol{\Sigma}_{jfk})$ 为多维高斯概率密度函数。

分析式（1）可知，如果噪声环境下直接采用式（1）进行同步概率估计，由于状态转移的连乘关系，会带来不同状态各频带概率估计交叉项。在纯净语音环境下，各频带都能对语音进行较好的识别，交叉项是在概率估计时对各频带取平均，所以可以获得较好的识别性能，同时利于模型的训练。但在噪声环境下，交叉项的存在会扩展噪声的影响，导致估计性能下降，所以引入 δ 算子对似然估计修正，限制受污染的频率信号的影响，以此减少噪声影响的扩散，如图 1 所示。

同时，为减少环境不匹配带来的性能恶化，假定 $\boldsymbol{Y}$ 与 $\boldsymbol{X}$ 间存在确定的映射关系，设 $g: \boldsymbol{x}_{ft} \to \boldsymbol{y}_{ft}$，$g(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$。假设环境变化仅对 CDHMM 观测概率的均值产生影响而其他参数保持不变，则有

$$P(\boldsymbol{y}_t \mid s_t = j) = \sum_{f=1}^{2^L-1} \sum_{k=1}^{K} CP(k_t = k \mid s_t, f_t = f) \cdot N(\boldsymbol{y}_{ft} \mid \boldsymbol{A}_{jfk}\boldsymbol{\mu}_{jfk} + \boldsymbol{b}_{jfk}, \boldsymbol{\Sigma}_{jfk})\delta(f_t = f_{t-1}) \tag{2}$$

其中，$\delta(f_t = f_{t-1}) = \begin{cases} 1 & f_t = f_{t-1}, \\ 0 & f_t \neq f_{t-1}。 \end{cases}$

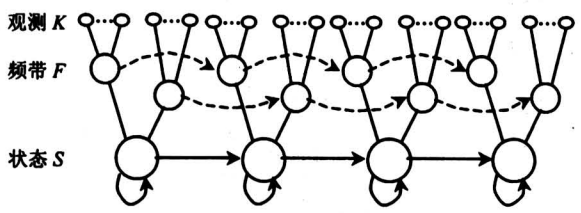

图 1 噪声环境下左右多带同步隐马尔可夫模型

Fig.1 Left-right multi-band synchronization HMM under noisy condition

此时，$P(\boldsymbol{Y} \mid \Theta, \Pi) = \sum_S \sum_F \sum_K \left[a_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} C \cdot P(k_t \mid f_t, s_t) O_{s_t f_t k_t}(\boldsymbol{y}_{f_t t}) \delta(f_t = f_{t-1}) \right]$。

采用最大似然准则来估计参数 Π，即

$$\Pi = \arg\max_{\Pi} P(\boldsymbol{Y} \mid \Theta, \Pi) \tag{3}$$

其中 Π 对应线性变换参数集。

设 $\{\boldsymbol{y}_{ft}\}$ 在 $s_t = j$，$f_t = f$，$k_t = k$ 时的高斯观测概率为

$$O_{jfk}(\boldsymbol{y}_{ft}) = N(\boldsymbol{y}_{ft} \mid \boldsymbol{A}_{jfk}\boldsymbol{\mu}_{jfk} + \boldsymbol{b}_{jfk}, \boldsymbol{\Sigma}_{jfk}) = (2\pi)^{-p/2} \mid \boldsymbol{\Sigma}_{jfk} \mid^{-1/2} \exp\{-(\boldsymbol{y}_{ft} - \boldsymbol{W}_{jfk}\tilde{\boldsymbol{\mu}}_{jfk})' \cdot \boldsymbol{\Sigma}_{jfk}^{-1}(\boldsymbol{y}_{ft} - \boldsymbol{W}_{jfk}\tilde{\boldsymbol{\mu}}_{jfk})/2\}, f \in [1, \cdots, 2^L - 1] \tag{4}$$

其中 $\boldsymbol{W}_{jfk} = [\boldsymbol{b}_{jfk}\boldsymbol{A}_{jfk}]$ 为状态 j 频带 f 混合高斯 k 对应的线性变换 g 的矩阵描述，$\tilde{\boldsymbol{\mu}}'_{jfk} = [1 \quad \tilde{\boldsymbol{\mu}}'_{jfk}]$，$p$ 为观测向量维数。

由于涉及隐变量，用 EM 算法[12]构造辅助函数求解。

$$Q(\Pi \mid \bar{\Pi}) = \sum_S \sum_F \sum_K P(S, F, K \mid \boldsymbol{Y}, \Theta, \bar{\Pi}) \cdot \{\ln a_{s_0} + \sum_{t=1}^{T} [\ln a_{s_{t-1}s_t} + \ln C + \ln P(k_t \mid s_t, f_t) + \ln O_{s_t f_t k_t}(\boldsymbol{y}_{f_t t})]\}\delta(f_t = f_{t-1}) = Q_a(\Theta, \bar{\Pi}) + Q_\Pi(\Theta, \bar{\Pi}, \Pi) \tag{5}$$

其中 Π 为被估线性变换参数，$\bar{\Pi}$ 为被估线性变换当前值，S 为状态序列，F 为频带序列，K 为混合高斯序列。$Q_a(\Theta, \bar{\Pi})$ 与 Π 无关，$Q_\Pi(\Theta, \bar{\Pi}, \Pi)$ 与 Π 相关。

$$Q_a(\Theta, \bar{\Pi}) = \sum_S \sum_F \sum_K P(S, F, K \mid \boldsymbol{Y}, \Theta, \bar{\Pi}) \cdot \{\ln a_{s_0} + \sum_{t=1}^{T} [\ln a_{s_{t-1}s_t} + \ln C + \ln P(k_t \mid s_t, f_t)]\} \cdot \delta(f_{s_t} = f_{s_t - 1}) \tag{6}$$

$$Q_\Pi(\Theta, \bar{\Pi}, \Pi) = \sum_S \sum_F \sum_K P(S, F, K \mid \boldsymbol{Y}, \Theta, \bar{\Pi}) \cdot$$

$$\left[\sum_{t=1}^{T}\ln O_{s_{f_t}k_t}(\boldsymbol{y}_{f_t})\right]\delta(f_{s_t}=f_{s_{t-1}}) \tag{7}$$

定义

$$\gamma_{jk|f}(t)=P(s_t=j,k_t=k\mid \boldsymbol{Y},\Theta,\bar{\Pi},f)=\frac{1}{P(\boldsymbol{Y}\mid\Theta,\bar{\Pi},f)}\sum_{S}\sum_{K}P(\boldsymbol{Y},s_t=j,k_t=k\mid\Theta,\bar{\Pi},f)$$

则

$$Q_{\Pi}(\Theta,\bar{\Pi},\Pi)=\sum_{f}c_f\sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{k=1}^{K}\gamma_{jk|f}(t)\ln O_{jfk}(\boldsymbol{y}_{ft}) \tag{8}$$

其中 c_f 为与频带有关的系数。

设不同频带采用不同的线性变换，通过上述分析可以看到 δ 算子的引入，对各频带线性变换参数可以进行单独估计。根据 EM 算法，考虑到实际应用中可能自适应数据较少，同频带高斯可采用相同的线性变换，设采用相同线性变换的高斯集合为 U_z，$z=1,\cdots,Z$。令

$$\frac{\partial Q(\Pi\mid\bar{\Pi})}{\partial \boldsymbol{W}_{fU_z}}=\frac{\partial Q_{\Pi}(\Theta,\bar{\Pi},\Pi)}{\partial \boldsymbol{W}_{fU_z}}=0 \tag{9}$$

得

$$\sum_{t=1}^{T}\sum_{j\in U_z}\sum_{k\in U_z}\gamma_{jk|f}(t)\boldsymbol{\Sigma}_{jfk}^{-1}\boldsymbol{y}_{ft}\tilde{\mu}'_{jfk}=\sum_{t=1}^{T}\sum_{j\in U_z}\sum_{k\in U_z}\gamma_{jk|f}(t)\boldsymbol{\Sigma}_{jfk}^{-1}\boldsymbol{W}_{fU_z}\tilde{\mu}_{jfk}\tilde{\mu}'_{jfk} \tag{10}$$

由式（10）两侧相等，通过求解一系列线性方程组可得到 $\{\boldsymbol{W}_f,\forall f\}$，根据 EM 算法，通过迭代，可获取参数优化解。

3　识别实验

识别实验是汉语普通话非特定人语音识别。为便于与 LMMLLR 算法比较，采用相同的实验环境，语音库由 19 个男性和 11 个女性近麦克风录制，采样率 8000，8 b 量化，语料为 0～9。语音参数使用 MFCC[13] 和 delta 参数，帧长 16 ms，帧移 8 ms，数据窗使用 Hamming 窗。含噪语音为不同噪声与纯净语音按不同信噪比进行混合。噪声选自 Noise92 噪声库，包括宽带噪声（white）和能量相对集中在某些频带的噪声（babble，destroyerengine）。其中，能量相对集中在某些频带的噪声环境，近似满足信号存在不受噪声影响的频带假设。子带定义 0～812 Hz，625～1750 Hz，1500～2812 Hz，2438～4000 Hz。实验结果见表 1 至表 3。

实验表明，对于宽带噪声（表 1），特别是低信噪比环境下，由于不满足 S 算法所需的信号存在不受噪声影响的频带假设，S 算法的性能较差。而 SMMLLR 算法，由于在最大似然估计的同时，采用

表 1　white 噪声下 S 算法，MLLR，LMMLLR，SMMLLR 误识率

Table 1　Error-rate of S algorithm, MLLR, LMMLLR, SMMLLR in white noise

SNR /dB	自适应数据数	误识率/%			
		S 算法	MLLR	LMMLLR	SMMLLR
0	5	74.7	53.7	46.3	47.6
	10	74.7	51.0	37.0	42.3
5	5	44.0	44.3	33.0	31.0
	10	44.0	43.3	28.3	25.7
10	5	34.6	40.3	24.3	21.0
	10	34.6	36.7	22.0	19.6

表 2　babble 噪声下 S 算法，MLLR，LMMLLR，SMMLLR 误识率

Table 2　Error-rate of S algorithm, MLLR, LMLLR, SMMLLR in babble noise

SNR /dB	自适应数据数	误识率/%			
		S 算法	MLLR	LMMLLR	SMMLLR
0	5	47.3	49.0	32.7	33.3
	10	47.3	43.3	31.3	29.6
5	5	40.3	45.3	26.0	28.0
	10	40.3	33.7	24.3	24.6
10	5	29.0	34.7	17.3	18.3
	10	29.0	28.7	16.7	16.0

表 3　destroyerengine 噪声下 S 算法，MLLR，LMMLLR，SMMLLR 误识率

Table 3　Error-rate of S algorithm, MLLR, LMMLLR, SMMLLR in destroyerengine noise

SNR /dB	自适应数据数	误识率/%			
		S 算法	MLLR	LMMLLR	SMMLLR
0	5	49.0	58.3	44.0	45.6
	10	49.0	50.7	39.0	38.6
5	5	40.0	40.3	26.0	27.0
	10	40.0	36.3	23.7	23.6
10	5	27.7	38.0	19.7	21.0
	10	27.7	35.7	15.7	15.4

同步多带模型和噪声污染假定，不仅不需要对信号做存在不受噪声污染的频带假设，而且克服了S算法子带分析中不同频带间信号相关性丢失的缺陷，因此SMMLLR不仅对于宽带噪声环境，而且对于能量相对集中在某些频带的噪声环境下都明显地降低了误识率。如在5自适应数据、white，babble和destroyerengine 5 dB环境下，SMMLLR相对于S算法，误识率分别下降29.5%，30.5%和32.5%。

由于SMMLLR通过隐变量使用噪声污染假定引入信息冗余，较好地捕捉了信号及噪声的频域特性，并根据信号的频谱特征进行最大似然补偿，而且同步特性又弥补了分频带导致的相关性丢失的缺陷，因此，SMMLLR比MLLR全带统一线性假设更好地补偿了噪声环境的恶化影响，所以SMMLLR性能上明显优于MLLR算法。如在white，babble，destroyerengine噪声0 dB环境、10自适应数据下，SMMLLR相对于MLLR，误识率分别下降17.1%，31.6%，23.9%。

LMMLLR算法中各频带信号处理是独立的，对信号处理来说较为简便。而SMMLLR算法中各频带信号处理利用了同步相关感知，有利于模型的简化，并且与全带分析是一致的。在不同的环境下，两种算法性能各有优势，这说明，人耳为了易于理解，总是有选择地跟踪特定频率的信号，独立与同步是交织进行的，这与人耳直接测试的感知效果是匹配的，这进一步表明，对SMMLLR和LMMLLR的研究是有效的。

需要指出：S算法需要计算 L 个HMM，此外还需要建立 2^L-1 个神经网络，并对其进行训练和计算；LMMLLR采用独立处理方式，不需要建立 2^L-1 个神经网络，需建立 2^L-1 个HMM；而SMMLLR采用多带同步模型，只需计算1个HMM。这说明SMMLLR算法简化了模型。

4 结论

人类听觉可以从带限信号中抽取音素信息。子带分离模型的一个重要缺陷是频带间相关信息的丢失。SMMLLR算法根据听觉感知特性，采用同步模型和噪声污染假定引入带间相关，一定程度上克服了子带分析所要求的独立感知假设导致的带间相关性丢失，在较少数据下就可获得较高的识别率，而且同步模型大大减少了模型数。同时，由于引入多带信息冗余，根据信号频谱特性进行模型补偿，一定程度上克服了MLLR算法全频带线性假设对环境间依存性描述不足的弱点，从而有效地改善了识别性能。人耳对不同频率的信号处理并不是绝对的独立处理，也不是绝对的同步处理，感知总是根据信号的特点变化，因此，从听觉实验入手将进一步研究特定频率信号的感知，改进现有识别方法。

参考文献

[1] Acero A. Acoustical and Environmental Robustness in Automatic Speech Recognition [M]. Kluwer Academic Press, Boston, MA, 1991

[2] Ephraim Y. Statistical-mode-based speech enhancement systems [J]. Proceedings of the IEEE, 1992, 80: 1526 ~ 1555

[3] Gales M J F. Model-based Techniques for Noise Robust Speech Recognition [D]. Engineering Department, Cambridge University, Cambridge, UK, 1995

[4] Gales M, Young S. Cepstral parameter compensation for HMM recognition in noise [J]. Computer Speech and Language, 1993, 12 (3): 231 ~ 239

[5] Sanker A, Lee C -H. Robust speech recognition based on stochastic matching [A]. ICASSP '95, Vol 1 [C]. Detroit, Michigan, USA, 1995. 121 ~ 124

[6] Leggetter C J, Woodland P C. Maximum likelihood linear regression for speaker adaption of continuous density hidden markov models [J]. Computer Speech and Language, 1995, 9: 171 ~ 185

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

[8] Hermansky H, Morgan N. RASTA processing of speech [J]. IEEE Trans On Speech Audio Processing, 1994, 2 (4): 578 ~ 589

[9] Tibrewala S, Hermansky H. Subband based recognition of noisy speech [A]. ICASSP '97, vol.2 [C]. Munich, Germany, 1997. 1255 ~ 1258

[10] 孙 暐 吴镇扬. 基于独立感知理论的鲁棒语音识别算法 [J]. 东南大学学报, 2005, 35(4): 506 ~ 509

[11] Bregman A S. Auditory Scene Analysis: the Perceptual Organization of Sound [M]. Cambridge, Massachusetts, The MIT Press, 1990

[12] Dempster A P, Laird N M, Rubin D B. Maximum likelihood estimation from incomplete data [J]. Journal Royal Statistical Society, Serials B, 1977, 39 (1): 1 ~ 38

[13] Mak B. A mathematical relationship between fullband and multiband mel-frequency cepstral coefficients [J]. IEEE Signal Processing Letters, 2002, 9: 241 ~ 244

(cont.on p.38)

可靠、简便，可以对综放沿空掘巷围岩稳定性进行科学分类。

3）随着综放沿空掘巷的推广和应用，其围岩稳定性影响因素的隶属函数应不断修正和完善。

参考文献

[1] 侯朝炯，郭励生，勾攀峰. 煤巷锚杆支护[M]. 徐州：中国矿业大学出版社，1999. 62～65

[2] 周保生，朱维申，李术才. 综放回采巷道围岩稳定性分类的研究[J]. 煤炭学报，2000，25(5)：469～472

[3] 马念杰，侯朝炯. 采准巷道矿压理论及应用[M]. 北京：煤炭工业出版社，1995. 144～147

[4] 周保生，朱维申，李术才. 神经网络在综放回采巷道锚杆支护设计中的应用研究[J]. 岩石力学与工程学报，2001，20(4)：497～501

[5] 朱川曲，缪协兴. 急倾斜煤层顶煤可放性评价模型及应用[J]. 煤炭学报，2002，27(2)：134～138

Classification Model and Its Application of Stability of Roadway Driving Along Next Goaf for Fully-mechanized Caving Face

Zhu Chuanqu, Wang Weijun, Shi Shiliang

(*School of Energy and Safety Engineering of Hunan University of Science and Technology, Xiangtan, Hunan* 411201, *China*)

[**Abstract**] The stability of roadway driving along next goaf for fully-mechanized caving face is synthetically influenced by many factors such as intensity of surrounding rock, intensity of coal, mining depth, joint and crack of surrounding rock, mining operation, top coal thickness, width of pillar and roadway section. On the basis of theoretical analysis, practical experience and observation data, the subordination functions of the factors influencing the stability of roadway driving along next goaf for fully-mechanized caving face are structured and the grey-fuzzy classification model is established. The application of examples shows that the model is accurate and reliable, and plays an important part in the support design, construction and management of roadway driving along next goaf for fully-mechanized caving face.

[**Key words**] roadway driving along next goaf for fully-mechanized caving face; stability of surrounding rock; classification model; subordination function

(cont. from p.34)

Multi-band Synchronization Model for Speech Recognition Under Noisy Condition

Sun Wei, Wu Zhenyang

(*Department of Radio Engineering, Southeast University, Nanjing* 210096, *China*)

[**Abstract**] Based on perception characteristic of human ear, this paper proposes synchronization multi-band maximum likelihood linear regression algorithm for robust speech recognition under noisy condition. The algorithm utilizes maximum likelihood as estimation criteria to compensate the effects of noisy condition with multi-band synchronization model and noise corruption assumption. The tests show that the proposed algorithm improves the performance of recognition system effectively.

[**Key words**] hidden Markov model; maximum likelihood; multi-band synchronization model; speech recognition