基于单高斯模型集的汉语美子带特征重建算法

罗 宇, 杜利民
(中科院声学所语音交互技术研究中心, 北京 100080)

摘 要: 本文提出了基于单高斯模型集的汉语美子带特征重建(SGMDI)方法, 并通过试验研究了该算法对提高语音识别系统加性噪声鲁棒性的作用. 实验结果表明: SGMDI 方法能够明显提高语音识别系统对各类音子尤其是容易被加性噪声破坏的清辅音音子的识别正确率, 从而显著增强了语音识别系统的噪声鲁棒性.

关键词: 数据重建; 缺失特征方法; 语音识别
中图分类号: TN912134 文献标识码: A 文章编号: 0372-2112 (2004) 10-1654-04

Single Gauss Model Set Based MAP Data Imputation Method for Mel-Frequency Filter-Bank Vectors of Chinese Speech

LUO Yu, DU Li-min
(Center of speech interaction technology research, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100080, China)

Abstract: Single Gauss Model set based Data Imputation (SGMDI) method is developed to recover Mel-frequency-filter-bank vectors of Chinese speech. Experiments are carried out to study how SGMDI method improves Automatic Speech Recognition (ASR) system's robustness against additive noise. Experimental results show that SGMDI method can improve phoneme correction of all kind of phonemes. Especially for unvoiced phonemes, which are easily distorted by additive noise, phoneme correction will be significantly improved. Thus, ASR system's robustness against additive noise can be greatly improved by SGMDI method.

Key words: data imputation; missing data method; ASR

1 引言

缺失特征方法[1~7]是提高语音识别系统噪声鲁棒性的一种新方法. 该方法认为噪声和语音在时间-频率域上不同区域具有不同能量分布, 把局部信噪比较低的区域标记为"缺失", 而局部信噪比较高的区域标记为"可靠", 即进行缺失分量估计(Missing Component Estimation)或者掩蔽估计(Mask Estimation). 经过缺失分量估计后, 可以直接根据"可靠矢量"进行语音识别, 即模型边缘化方法; 也可以重建"缺失矢量", 得到完整矢量后进行语音识别, 即数据重建方法. 当噪声为不稳定噪声时, 缺失特征方法具有潜在的优越性.

本文提出了基于单高斯模型集的汉语美子带特征 MAP 重建(SGMDI)算法, 并在大词汇表非特定人连续语音识别这种复杂程度的条件下, 研究该方法对提高语音识别系统加性噪声鲁棒性的贡献以及对不同类型音子的影响.

论文的第 2 部分讨论汉语美子带特征单高斯模型集; 第 3 部分讨论基于单高斯模型集的汉语美子带特征数据重建算法; 第 4 部分实验分析了 SGMDI 算法提高语音识别系统噪声鲁棒性的作用; 第 5 部分给出了最后的结论.

2 单高斯模型集

本文采用在美频率(Mel-Frequency)域均匀分布的 26 个三角滤波器进行语音子带特征分析, 选择具有完整协方差矩阵的单高斯模型集来描述美子带特征, 并假设所有汉语纯净语音美子带特征都来自 N 个单高斯模型构成的单高斯模型集.

对于纯净语音特征矢量 S, 单高斯模型的概率密度函数如公式(1)所示:

$$P_{M_j}(S)=\frac{\exp\{-\frac{1}{2}(S-L_j)^{t}H_j^{-1}(S-L_j)\}}{(2P)^{\frac{n}{2}}|H_j|^{\frac{1}{2}}} \qquad (1)$$

公式(1)中, L_j, H_j 是第 j 个单高斯模型的均值矢量和协方差矩阵($1\le j\le N$), N 是单高斯模型集合中单高斯模型数目.

在估计单高斯模型参数之前, 首先需要对美子带特征矢量进行聚类, 聚类算法可以选择常用的 K-均值聚类算法[9]. 假设美子带特征分布符合高斯分布, 每个聚类对应一个单高斯模型. 估计单高斯模型 i 的先验概率

收稿日期: 2003-09-28; 修回日期: 2003-12-10
基金项目: 国家 973 重点基础研究发展项目"图像、语音、自然语言理解和知识挖掘"; 汉语自然口语对话的理论和实验平台研究(G1998030505)

$$P(M_i)=\frac{N(SIM_i)}{\sum_{i=1}^{L}N(SIM_i)},\quad 1\le i\le N \tag{2}$$

其中, $N(SIM_i)$ 表示在数据库中, 语音特征矢量属于单高斯模型 M_i 的次数.

估计单高斯模型 i 均值和协方差[9]:

$$\mu_i=\frac{1}{SIcluster(i)}\sum_{SIcluster(i)}(S),\quad 1\le i\le N$$
$$\Sigma_i=\frac{1}{SIcluster(i)}\sum_{SIcluster(i)}[(S-\mu_i)(S-\mu_i)^t],\quad 1\le i\le N \tag{3}$$

3　汉语语音美子带特征数据重建算法

3.1　理想缺失分量估计方法

本文数据重建算法所处理的对象是汉语语音美子带特征矢量. 缺失分量估计和数据重建都在美子带特征矢量空间内进行. 假设纯净语音美子带特征矢量为 S, 噪声美子带特征矢量为 N, 理想缺失分量估计按如下公式进行:

$$MSK_i(k)=\begin{cases}1 & \text{if } SNR_i(k)=10\log_{10}\left(\frac{S_i(k)}{N_i(k)}\right)>D\\ 0 & \text{if } SNR_i(k)=10\log_{10}\left(\frac{S_i(k)}{N_i(k)}\right)<=D\end{cases} \tag{4}$$

其中, $S_i(k)$ 是纯净语音第 i 帧美子带特征的第 k 个分量(对应第 k 个美三角子带内的纯净语音能量); $N_i(k)$ 是噪声第 i 帧美子带特征的第 k 个分量(对应第 k 个美三角子带内的噪声能量); D 是判断该子带是否可靠的门限, 根据人耳掩蔽效应, 选择 D 范围为 -5~5dB. $MSK_i(k)=1$ 表示第 i 帧语音第 k 个美子带信噪比较高, 是/可靠0子带; $MSK_i(k)=0$ 表示第 i 帧语音第 k 个美子带信噪比较低, 是/缺失0子带. 理想缺失分量估计的条件在实际环境中很难得到满足, 但是, 理想缺失分量估计可以用于评价数据重建算法的性能.

3.2　美子带特征数据重建

经过缺失分量估计, 语音特征 S 分为两个矢量: 不可靠分量构成的/缺失矢量0 S^m 和可靠分量构成的/可靠矢量0 S^o, 表示为: $S=[S^oS^m]$. SGMDI 算法原理如图 1 所示:

(1) 计算、比较/可靠0矢量对每个高斯模型的边缘化概率, 选择边缘化概率最大的高斯模型, 作为语音特征在高斯模型集中所属高斯模型的估计(参见 B 图); (2) 根据/可靠矢量0、高斯模型参数, 按最大后验概率准则(MAP)估计出/缺失矢量0(参见 C 图).

基于单高斯模型集的美子带特征数据 MAP 重建算法的步骤如下:

首先, 根据/可靠矢量0 S^o, 估计语音美子带特征 S 在单高斯模型集中所属单高斯模型:

$$j^*=\arg\max_j(P_{M_j}(S^o)\cdot P(M_j)) \tag{5}$$

其中, $P(M_j)$ 是出现第 j 个单高斯模型的先验概率, $P_{M_j}(S^o)$ 表示语音特征矢量 S 属于模型 M_j 的情况下, 观测到 S^o 的概率, 即是第 j 个单高斯模型对缺失特征 S^o 的边缘化概率

$$P_{M_j}(S^o)=\int P_{M_j}(S)\,dS^m=\int P_{M_j}(S^mS^o)\,dS^m \tag{6}$$

根据第 j^* 个单高斯模型参数, 按最大后验概率(Maximum A Posterior)准则来进行缺失特征重建[5,7]:

$$S^m=\mu_{j^*m}+\Sigma_{j^*mo}\Sigma_{j^*oo}^{-1}(S^o-\mu_{j^*o}) \tag{7}$$

其中, μ_{j^*m} 表示第 j^* 个单高斯模型中/缺失0子带所对应的均值矢量; μ_{j^*o} 表示第 j^* 个单高斯模型中/可靠0子带所对应的均值矢量; Σ_{j^*oo} 表示第 j^* 个单高斯模型中/可靠0子带所对应的协方差矩阵; Σ_{j^*mo} 表示第 j^* 个单高斯模型中/可靠0子带和/缺失0子带间的协方差矩阵.

4　试验分析

下面, 本文将在大词汇表非特定人汉语连续语音识别任务下, 通过试验分析 SGMDI 算法的重建效果, 以及对提高语音识别系统噪声鲁棒性的作用.

4.1　实验条件

大词汇表非特定人汉语连续语音识别系统的训练和测试数据来自 863 语音数据库, 其中 166 人的数据用于训练, 8 人的数据用于测试. 噪声数据来自 NoiseX92 噪声数据库. 试验选用 2 种噪声(高斯白噪声、babble 噪声), 按照 SNR= 25, 20, 15, 10, 5, 0 dB 加入到纯净语音中, 同时, 保存纯净语音和噪声用于理想缺失分量估计.

语音信号使用 25 ms 哈明窗对连续语音进行分帧, 并用 0.97 的预加重滤波器提升高频分量, 相邻帧重叠 15 ms. 语音特征矢量选择 MFCC_E_D_A. 子带分析采用在 0~8000 Hz 范围内按照美刻度均匀分布的 26 通道三角滤波器组. 语音识别系统隐马尔可夫模型的结构为: 停顿模型使用 3 个状态; 静音和子音模型使用 5 个状态, 首尾两个状态没有输出, 仅用来连接模型. 解码使用文法无关, 困惑度为 406 的汉语拼音音节网络.

受高斯白噪声破坏的语音首先转换为美子带特征, 经过理想缺失分量估计、数据重建后, 转换为 MFCC 特征, 并作为识别器的输入. 试验中, 根据人耳掩蔽效应, 选择掩蔽门限 R= -5 dB, 单高斯模型数 N= 256.

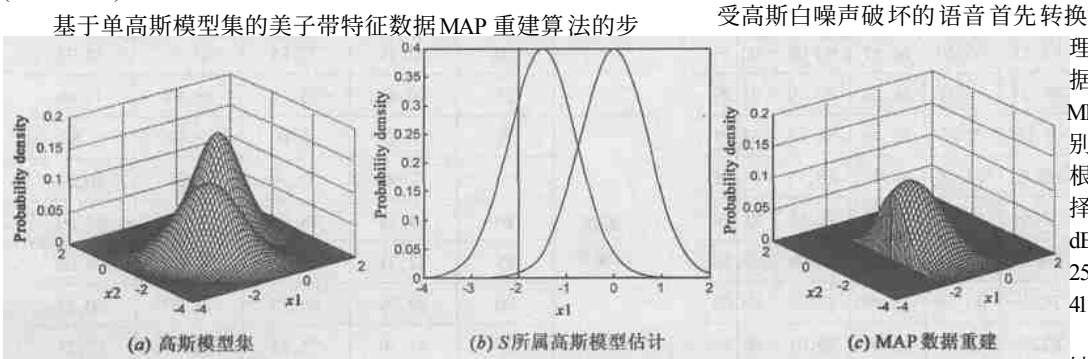

图 1　MAP 数据重建原理

4.2　实验结果及分析

图 2 给出了理想缺失分量估计后, 利用 SGMDI 算法对受噪声

破坏的语音特征矢量序列进行数据重建, 得到的结果. 从图 2 可以看出, 加性噪声破坏了纯净语音特征矢量的形态和分布. SGMDI 算法能够有效地重建出受加性噪声破坏的美子带特征, 重建后的美子带特征较好的重现了原始纯净语音美子带特征的形态和分布(参见图 2).

另一方面, 即使是进行理想缺失特征估计, 用单高斯模型进行缺失特征数据重建的结果仍然会丢失语音特征矢量的部分重要信息. 语音中具有较低能量的音子可能在所有子带中都被标记为缺失, 从而难与静音部分区分开来, 因此重建为/ 静音0(参见图 2).

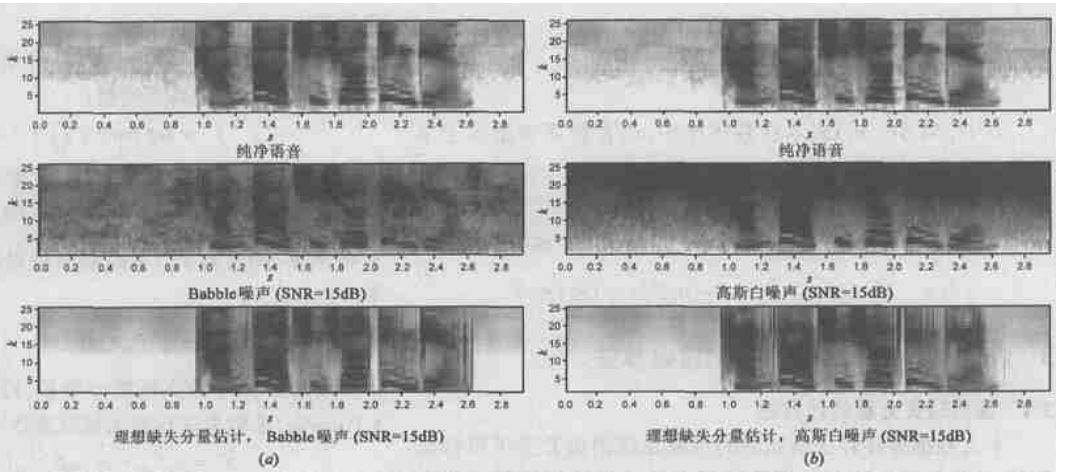

图 2 美特征重建试验结果: (a) 受高斯白噪声破坏的语音的实验结果, (b) 受 babble 噪声破坏的语音的实验结果. (图中的汉语语音是:/ 谈到汽车定点(tan2 dao4 qi4 che1 ding4 dian3)0)

在语音识别实验中, 定义识别正确率(Correct) 和准确率(Accuracy)[10]为:

$$\text{Correct}\% = \frac{N - D - S}{N} \times 100\% \quad (8)$$

$$\text{Accuracy}\% = \frac{N - D - S - I}{N} \times 100\%$$

其中, N 为识别单元(单词、音子、音节等) 出现的总次数, D 为删除错误次数, S 为替换错误次数, I 为插入错误次数.

表 1 给出了对受噪声破坏的语音进行实验, 得到的各类音子平均识别正确率变化表. 从表 1 可以看出, 加性噪声的存在将导致语音识别系统对各类音子的识别正确率发生明显下降: 能量较低, 持续时间较短的清辅音音子容易受到噪声的破坏, 音子识别正确率大幅下降; 能量较高, 持续时间较长的浊辅音音子和元音音子抵抗噪声能力较强, 音子识别正确率下降幅度较小.

经过基于单高斯模型集的美子带特征重建算法之后, 各类音子的音子正识率得到了不同程度的提高. 受到噪声严重破坏的清辅音音子正识率获得了大幅度的提高; 受噪声影响较小的浊辅音音子和元音音子平均音子正识率提高幅度相对较小(参见表 1).

表 2 是语音识别实验得到的音节正确率和音节准确率结果. 实验结果说明加性噪声破坏了语音特征矢量的形态和分布, 造成语音识别系统性能大幅下降. 经过基于单高斯模型集的美子带特征重建后, 由于重建后的美子带特征图较好的重

表 1 音子正识率分类统计表

噪声环境		Noisy, 音子平均识别准确率(%)			SMRGDI, 音子平均识别准确率(%)		
噪声类型	信噪比	清辅音	浊辅音	元音	清辅音	浊辅音	元音
Babble 噪声	0 dB	01 54	261 00	20121	251 53	401 71	561 59
	5 dB	219115	61162	31134	541 35	671 58	751 26
	10 dB	25114	76179	62186	711 02	821 42	841 56
	15 dB	57115	84134	79130	791 58	881 20	881 62
	20 dB	72163	87122	86140	841 57	901 08	901 61
	25 dB	80153	89123	89141	861 26	901 79	911 29
高斯白噪声	0 dB	01 41	0143	1147	121 46	481 87	531 21
	5 dB	288193	248151	24126	341 78	691 77	711 95
	10 dB	11 15	33161	47183	551 47	791 55	811 59
	15 dB	29159	60197	68177	671 23	831 99	851 68
	20 dB	52162	74136	80156	751 95	871 27	891 06
	25 dB	68169	82178	86165	811 91	891 01	901 79
纯净语音		88127	91101	91188	881 27	911 01	911 88

表 2 含噪语音美子带特征重建, 语音识别实验结果比较(高斯/ babble 噪声, SNR= 15 dB)

含噪语音		音节正确率(%)		音节准确率(%)	
噪声类型	SNR(dB)	NOISY	SGMDI	NOISY	SGMDI
Babble 噪声	0	3109	231 78	231 07	131 64
	5	9172	411 75	251 81	321 03
	10	27153	581 41	8115	511 23
	15	48162	681 20	291 74	621 36
	20	62111	731 88	451 97	681 75
	25	70148	761 13	581 91	711 49
高斯白噪声	0	2180	151 80	2149	31 56
	5	7106	321 19	2134	201 24
	10	17119	481 07	1131	391 49
	15	32131	571 84	121 09	511 02
	20	48128	661 35	281 00	601 17
	25	61119	721 33	561 14	671 25
纯净语音		78162		741 71	

现了原始纯净语音段美子带特征的形态和分布,因此系统识别率有了较大的提高(参见表 2).

高斯白噪声是典型的平稳噪声信号,而 Babble 噪声是典型的非平稳信号.经过基于单高斯模型集的美子带特征重建后,识别系统的性能都得到了明显的提高.相对于谱减法(SS2Spectrum Subtraction)、矢量台劳级数(VTS2Vector Taylor Series)等算法,对非平稳噪声信号,该方法具有更好的鲁棒性.

5　结论

SGMDI 算法假设所有汉语纯净语音美子带特征都来自 N 个单高斯模型构成的单高斯模型集,并按最大后验概率准则(MAP)估计出语音美子带特征矢量中受噪声破坏的0缺失0分量.试验结果表明,SGMDI 算法能够增强语音识别系统对平稳高斯白噪声和非平稳 Babble 噪声的鲁棒性.本文的进一步工作是解决缺失分量估计问题.

参考文献:

[1] A Vizinho, P Green, M Cooke and L. Josifovski. Missing data theory, spectral subtraction and signal2to2noise estimation for robust ASR: An integrated study[A]. Eurospeech. 99[C]. Budapest, 1999.

[2] Martin Cooke, Phil Green, Ljubomir Josifovski, Ascension Vizinho. Robust ASR with unreliable data and minimal assumptions[A]. Robust 99[C]. Tamper, Finland.

[3] Morris A C, Cooke M & Green P. Some solutions to the missing feature problem in data classification, with application to noise robust ASR[A]. Proc. ICASSP. 98[C]. 1998. 737- 740.

[4] B Raj, M L Seltzer, R M Stern. Robust speech recognition: the case for restoring missing features[A]. Workshop on Consistent and Reliable Acoustic Cues for Sound Analysis (CRAC) 2001[C]. September, 2001, Aalborg, Denmark.

[5] Bhiksha Raj, Michael L. Seltzer, Richard M. Stern. Reconstruction of damaged spectrographic features for robust speech recognition[A]. International Conference on Spoken Language Processing[C]. October, 2000, Beijing, China.

[6] Philippe Renevey, Rolf Vetter, Jens Kraus. Robust speech recognition using missing feature theory and vector quantization[A]. Eurospeech 2001[C]. Scandinavia, pp1107.

[7] B Raj. Reconstruction of Incomplete Spectrograms for Robust Speech Recognition[D]. Ph. D dissertation, ECE Department, CMU, April, 2000.

[8] Lawrence Rabiner, Biing2Hwang Juang. Fundamentals of speech Recognition, 语音识别基本原理[M]. 清华大学出版社;

[9] 边肇祺等. 模式识别[M]. 清华大学出版社.

[10] Steve Young, Dan Kershaw, Julian Odell, Dave Ollason, Valtcho Valtchev, Phil Woodland. The HTK Book (for HTK Version 310)[M]. Microsoft.

作者简介:

罗　宇　男,1974 年出生于四川,1998 年毕业于清华大学,中国科学院声学研究所博士. E2mail: luoyu@tsinghua.org.cn

杜利民　男,1983 年、1987 年、1991 年分别于北京大学、中国科大研究生院、中国科学院声学研究所获理学学士、工学硕士、理学博士学位;1996 年美国麻省理工学院(MIT)访问科学家;1999 年美国 AT&T 访问研究员.现为中国科学院声学研究所主任研究员;语言、语音和交互信息技术部主任;青海省人民政府科技顾问;中国移动互联网应用协议 CLV 专家组成员;中国电子学会理事;中国声学学会会士;中国通信学会通信理论与信号处理专业委员会副主任;中国人工智能学会人工神经网络专业委员会副主任;中国人工智能学会自然语言理解专业委员会委员;国际电子电气工程师协会(IEEE)高级会员;国际语音通信协会(ISCA)会员;5 电子学报6 编委.多年从事语音信号与信息处理技术的研究,主持研究的主要项目包括汉语话者无关的连续语音鲁棒识别、连续语音关键语检测、语音交互助理、自然口语识别与对话、语音同声翻译、噪声环境下语音增强和语音提取、低速率语音压缩.