多声源环境下的鲁棒说话人识别①

张凤仪, 夏秀渝, 冉国敬, 何 礼, 叶于林

(四川大学 电子信息学院, 成都 610065)

摘 要: 针对多声源干扰环境下说话人识别系统性能急剧下降的问题, 提出一种提取目标语音的前端处理方法, 该方法依据独立语音时频域的近似稀疏性, 基于目标语音方位信息采用非线性时频掩蔽方法提取目标语音. 建立了基于梅尔倒谱系数(MFCC)的高斯混合模型(GMM)说话人识别系统. 仿真实验证明, 该方法能有效提取目标语音, 提高说话人识别系统的鲁棒性. 该文多声源干扰仿真实验条件下, 说话人识别系统的识别率平均提高了25%左右.

关键词: 说话人识别; 语音增强; 方位信息; 时频掩蔽; MFCC 参数

Robust Speaker Recognition in Multi-Source Environment

ZHANG Feng-Yi, XIA Xiu-Yu, RAN Guo-Jing, HE Li, YE Yu-Lin

(College of Electronics and Information Engineering, Sichuan University, Chengdu 610064, China)

Abstract: The Speaker Recognition System is significantly affected by the Multi-Sound sources problem. In order to overcome this problem, a target sound extraction algorithm named time-frequency masking is proposed. The proposed algorithm is based on the sound source azimuth information and the approximate sparse nature of sound. A Mel-frequency cepstral coefficient (MFCC) based Gaussian mixture model (GMM) speaker recognition system is presented to improve the recognition robustness. The proposed algorithm has been tested on the simulated data through a number of experiments which shows the efficiency and robustness of the proposed algorithm. In the Multi-Sound sources environment, the recognition rate of the proposed algorithm can be improved by about 25%.

Key words: speaker recognition; speech enhancement; sound source azimuth information; time-frequency masking; MFCC

1 引言

随着计算机技术和信息化社会的发展, 说话人识别作为具有语音识别与理解功能的智能人 T 机接口, 其应用领域在不断扩大, 如保密安全、司法鉴定及军事领域等[1].

目前说话人识别的主要方法有动态时间规整[2]、GMM[3]、隐马尔科夫[4]、矢量量化[5]和神经网络[6]等. 当前大部分说话人识别系统在低噪声、低失真环境下性能已经有所改善, 实验证明, 这些传统方法的识别率可达 85%以上. 但系统在实际应用中不可避免的会受到干扰、信道等影响, 导致其性能恶化. 针对这种情况, 现今有提取鲁棒性参数和前段语音增强两种类型处理方法.鲁棒性参数一般有 MFCC、线性预测倒谱系数等. 前端语音增强方法有谱减法、维纳滤波法等. 当噪声近似为平稳宽带噪声时, 这两种方法是比较有效的.

然而, 实际环境是很复杂的.如当说话人识别系统应用于“鸡尾酒会”, 即多声源环境, 性能就急剧下降. 实际多声源环境中, 往往干扰源的类型、个数、方向、混合方式均未知, 要想从众多的干扰源中获取目标语音是非常困难的. 这是说话人识别系统实用化的关键, 也是当前的研究热点.

本文针对多声源环境下的语音识别, 提出了使用双通道提取目标语音的前端处理方法, 该方法依据语音的时频稀疏性, 仅利用目标语音的方位信息采用时频掩蔽的方法提取目标语音. 在建立的基于 MFCC 参

① 收稿时间:2014-08-11;收到修改稿时间:2014-09-05

数的 GMM 说话人识别系统上进行的对比实验证明, 在本实验条件下, 说话人识别系统性能有较大幅度提升, 目标语音的识别率平均提高 25%左右.

2 基于目标声源方位信息和时频掩蔽的语音增强算法

2.1 基于目标声源方位信息时频掩蔽算法基本原理[7]

在实际应用中, 目标说话人相对麦克风方位大致固定, 而干扰源的个数和方向均未知, 如公共场合打电话, 有主讲人的会议系统等. 因此, 提出了一种基于目标语音初始方位信息采用时频掩蔽提取目标语音的双通道系统方案.

一般来说, 目标信号与各路干扰信号来自不同方位, 双通道系统能提供一定的声音空间方位信息. 如图1所示, 点声源S发出的信号经过不同传输路径$h_1(t)$、$h_2(t)$到达主、从麦克风. 设$h_1(t)=h_2(t)*\Delta h(t)$, $\Delta h(t)$表示$h_1(t)$和$h_2(t)$的差异滤波器. 由图1清晰可见, 当声源方位不同时. 相应的$\Delta h(t)$也是不同的. 因此, 我们可以通过$\Delta h(t)$来辨识不同方位的信号. 这里将声源方位信息定义为$\Delta H(\omega)$, 为$\Delta h(t)$的频率响应.

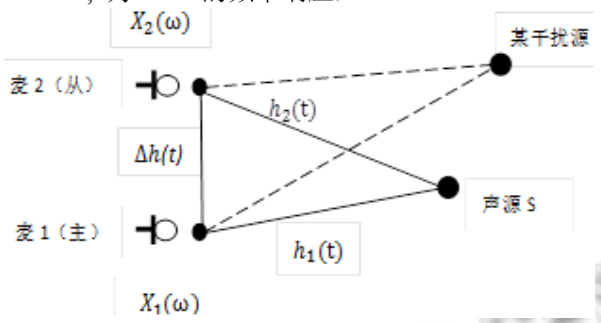

图1 方位信息图示

在单声源且无干扰源的情况下, 主、从麦克风接收的信号分别为$X_1(\omega)$、$X_2(\omega)$, 该声源的方位信息$\Delta H(\omega)$可由式(1)求得:

$$\Delta H(\omega)=\frac{X_2(\omega)}{X_1(\omega)} \tag{1}$$

实际可能是多声源情况, 这时麦克风收到的是混合信号, 这些信号无论时域或频域均是混杂的. 但研究表明[7], 独立信号在时频域上具有近似稀疏性, 就混合信号的某个时频单元来看, 它仅属于某个声源. 因此, 可用(1)式估计混合信号每个时频点所对应声源的方位信息, 然后将其和目标声源的方位信息比较, 采用时频掩蔽方法将目标语音分离出来.

设主、从麦克风接收的混合语音信号的某时频单元为$X_1(\omega_k,t)$和$X_2(\omega_k,t)$, 有:

$$\Delta H_t(\omega_k)=\frac{X_2(\omega_k,t)}{X_1(\omega_k,t)} \tag{2}$$

另设目标语音方位信息为$\Delta H_0(\omega)$, 某频点为$\Delta H_0(\omega_k)$. 由于针对的环境目标语音方位大致固定, 其方位信息可预先测定. 定义该时频单元估计的$\Delta H_t(\omega_k)$和目标语音$\Delta H_0(\omega_k)$的相对误差为参数α.

$$\alpha=\frac{||\Delta H_t(\omega_k)-\Delta H_0(\omega_k)||}{||\Delta H_t(\omega_k)||+||\Delta H_0(\omega_k)||} \tag{3}$$

然后, 采用(4)式求得非线性掩蔽函数$M_k(\omega_k,t)$:

$$M_k(\omega_k,t)=\frac{1}{1+(\alpha/\lambda)^p} \tag{4}$$

最后由(5)式得到时频掩蔽法提取的目标语音:

$$Y(\omega_k,t)=M_k(\omega_k,t)X_1(\omega_k,t) \tag{5}$$

以上时频掩蔽算法由于采用了非线性掩蔽函数, 在滤掉与目标语音不符频点时采用平滑截断, 因此保留了相对较好的听觉效果.

2.2 目标声源的方位信息实时跟踪处理

通过上文时频掩蔽方法初步提取的目标语音的信噪比已有很大提高. 但实际运用中, 目标声源的方位可能会有一些小变动, 如说话人移动头部或对身体坐姿进行一定调整等. 目标声源方位信息是该方法中最重要的一环, 若方位信息出现了偏差, 将直接影响到语音增强的效果. 因此, 提出了对目标声源的方位进行实时跟踪处理.

用上述时频掩蔽法从主、从麦克风中初步提取的目标语音不仅信噪比提高了, 而且依然保留了声源的方位信息, 进一步可利用这些信号采用批处理相关法辨识法来跟踪$\Delta H_t(\omega_k)$的变化. 公式如下:

$$\Delta H_t(\omega_k)'=\frac{G_{21}(\omega)}{G_{11}(\omega)} \tag{6}$$

其中, $G_{11}(\omega)$为初步时频掩蔽处理后主麦克风的信号功率谱, $G_{21}(\omega)$为初步时频掩蔽处理后主、从麦克风信号的互功率谱. 这个辨识过程可以不断进行以跟踪目标方位的变化. 也可用自适应方法对目标语音的方位进行实时更新, 也会有较好的效果.

将得到的准确的$\Delta H_t(\omega_k)'$用于时频掩蔽过程, 然

后通过短时傅里叶逆变换即可得到最终提取的目标声源信号. 当然, 由于本文接下来进行的是基于 GMM 的说话人识别, 因而可以直接将(5)式得到的频域信号提取 MFCC 参数备用, 无需在频域信号与时域信号之间反复变换.

3 双通道的前端处理的说话人识别系统

针对多声源环境下的说话人识别, 整个系统分为两个阶段: 目标语音增强阶段及说话人识别阶段.

系统原理图如图 2.

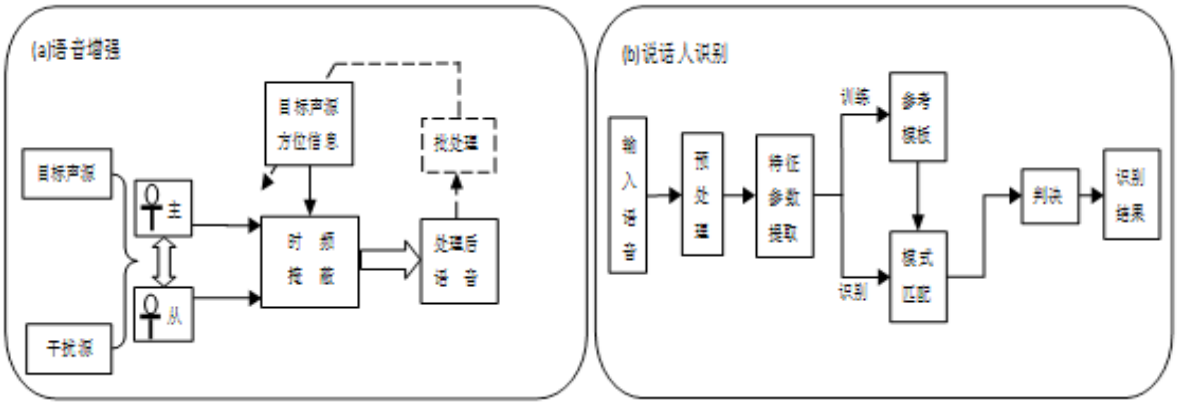

图 2 带双麦克风前端处理的说话人识别系统原理图

图 2(a)为前端语音增强部分, 由主、从麦克风接收到的混合信号, 利用已知的目标声源的方位信息进行时频掩蔽处理, 得到增强的语音.图 2(a)中虚线部分为利用批处理互相关法进行的方位实时更新的处理.

图 2(b)为说话人识别系统的原理框图, 分为两部分: 第一部分是说话人模型的训练, 将仅有目标语音而无其他干扰声源时, 主麦克风收到的语音作为训练语音, 提取特征参数训练模型参数.第二部分是说话人识别部分, 将测试语音进行预处理及特征参数提取, 输入模型进行匹配判决计算, 即为说话人识别.

语音的预处理包括: 预加重、加窗分帧、端点检测等.语音信号进行预加重可以减少口唇辐射的影响, 一般通过传递函数为 $H(z)=1-az^{-1}$ 的一阶 FIR 高通数字滤波器来实现, $0.9<a<1.0$. 在本文的实验中, 取 a=0.9375. 一般认为语音在 10ms~30ms 内是平稳的, 因此需对语音进行加窗分帧.端点检测是为了找出语音信号的起止点的位置, 区分开语音和各种非语音信号时段. 本文使用的是基于短时能量和短时过零率的端点检测方法.

在说话人识别中, MFCC 特征参数借鉴人耳的听觉机理, 相对于其他特征参数具有强抗噪性、高识别率的特点.对语音帧进行短时傅里叶变换并求出其短时能量谱, 用 M 个 Mel 带通滤波器进行滤波得到 Mel 谱, 然后对其取对数并进行反离散余弦变换就得到了 L 个 MFCC 系数, 一般 L 取 12~16 个左右.

说话人识别系统分为训练和识别. 为说话人建立GMM模型, 实际上就是通过训练, 估计其参数λ, 使其能最好的匹配训练特征矢量的分布. 在这里运用了最大似然的估计方法.

由于不易用通常办法找到似然函数的极大值点, 因此采用了EM算法[8][9][10](期望值最大)来估计GMM的参数λ. 它从一个初始模型开始. 每次迭代的估计出一个新的模型参数λ', 使$P(o|\lambda) \le P(o|\lambda')$, 然后再以λ'作为模型的参数开始下一次迭代, 直到收敛. 具体的参数迭代公式如(8)-(10):

$$c_i = \frac{1}{T}\sum_{t=1}^{T} P(q_t = i|o_t,\lambda) \tag{8}$$

$$\mu_i = \frac{\sum_{t=1}^{T}[P(q_t = i|o_t,\lambda)o_t]}{\sum_{t=1}^{T} P(q_t=i|o_t,\lambda)} \tag{9}$$

$$\sigma_{ik}^2 = \frac{\sum_{t=1}^{T}[P(q_t = i|o_t,\lambda)(o_{tk}-(\mu_{tk})^2]}{\sum_{t=1}^{T} P(q_t = i|o_t,\lambda)} \tag{10}$$

训练好模型后, 接下来是说话人的识别. 对于有N个人的说话人识别系统, 每个说话人用一个GMM模型来代表, 记为: $\lambda_1,\lambda_2,\ldots\ldots,\lambda_N$. 测试语音的识别结果由最大后验概率准则给出, 即:

$$n^* = argmax_{1\le n\le N}\, lnP(O|\lambda_n) = argmax_{1\le n\le N}\sum_{t=1}^{T} lnP(o_t|\lambda_n) \tag{11}$$

其中, 记测试语音的观测特征矢量序列为 O, $P(O|\lambda_n)$

为第 n 个人产生特征矢量集 O 的条件概率, 为了简化计算, 一般采用对数似然函数$\ln P(O|\lambda_n)$.

综上, 选取 MFCC 参数基于 GMM 方法, 由以上具体步骤便可以完成说话人模型的建立及说话人识别.

4 实验仿真

该实验选取模拟的会议环境进行仿真实验. 参考实际会议环境, 有 1 个固定方位的主讲人和其他 5 个方位的听众, 因此, 可以通过主讲人与各听众的方位信息的差异来增强主讲人方位的语音信息. 在该环境下, 作为目标声源的主讲人的方位相对固定, 主、从麦克风的位置固定, 而干扰源类型、个数、幅度和位置均未知. 利用专业软件“Room Impulse Respon-se”构造的实验环境如图 3 所示: 房间长 6 米, 宽 5 米, 高 3 米, 房间混响时间 T=360ms.两个全向麦克风的位置固定且间隔为 0.2m, 6 个声源处于不同空间方位, 声源 S1-S6 围绕会议桌布置,各声源到麦克风的冲击响应由专业软件生成.设目标语音为 S1, 与主、从麦呈三点一线. S2-S6 为干扰源, 为人声、敲门声及短信声等多种干扰源. 实验中, 麦克风的接收端均加入了信噪比为 30dB 的白噪声来模拟环境噪声.

在实验中, 采用了语音库“TIMIT Database”[11].该语音库中每人有 10 句录音, 每句时长为 2s~4s 不等. 在其中随机抽取 17 人、共 100 句语音进行实验.将每人 10 句话中的其中 4 句作为说话人模型训练, 其余 6 句作为说话人识别测试用.

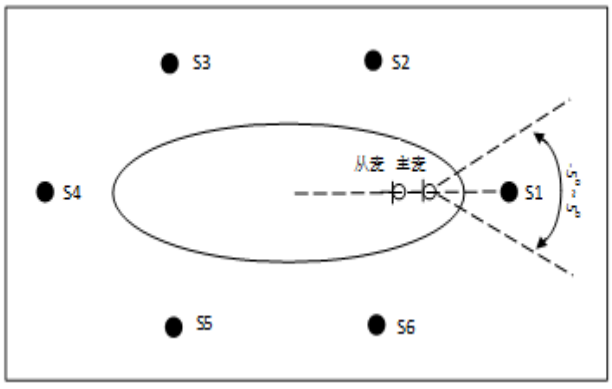

图 3 会议系统环境示意图

带双麦克风前端处理的说话人识别系统的实验步骤分为以下四步:

(1) 无干扰源时, 由主麦分别采集 N 个目标语音, 提取其 MFCC 特征参数, 分别进行说话人训练, 建立 N 个 GMM 模型.

(2)在无干扰源时, 由主、从麦克风收到的目标声源信号辨识目标源初始方位信息.

(3)多声源情况下分别采集主、从麦中的混合语音, 进行时频掩蔽提取目标语音.

(4)将混合语音/增强语音提取 MFCC 特征参数, 输入各模型进行匹配计算, 进行说话人识别对比实验.

在以上实验环境下, 进行了以下四个实验.

实验一: 目标语音提取实验

在本实验中, s1 为 TIMIT 语音库中随机抽取的 1 名说话人, s2-s6 为 5 个干扰源, 分别为 1 个男声、2 个女声、1 个敲门声和 1 个诺基亚短信音,实验结果如图 4 所示.

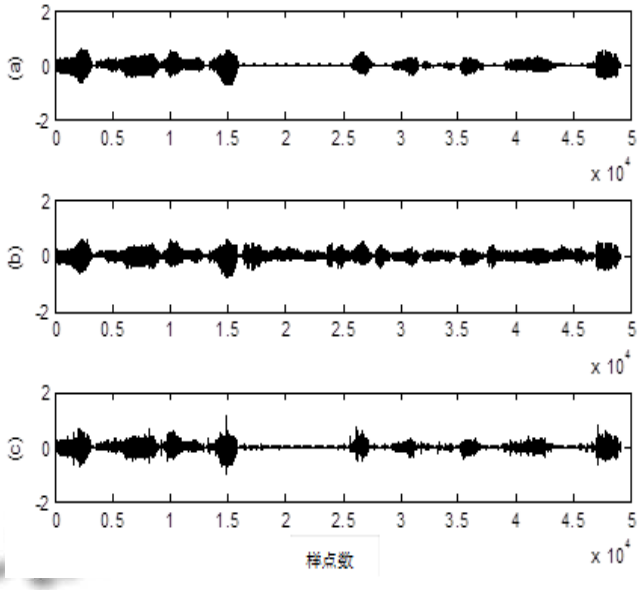

图 4 语音增强前后波形对比图;
(a)目标纯语音(b)主麦接收的混合声音(c)提取后语音

对比图 4 各波形可知, 从混合信号中较好地提取出了目标语音.对多次实验结果进行分析, 平均信噪比增益达 9dB.主观试听处理后的信号, 可清晰听见目标语音, 几乎听不见其他干扰声, 能明确辨别出说话人的词句, 但缺点是有一些音乐噪声.

实验二: 不同前端处理方法的说话人识别对比实验

用本文方法及传统的谱减法分别对混合语音进行处理.混合语音的组成分两种情况: (1)混合语音由目标语音与 5 个方向的干扰语音混合而成(后称“多”), (2)混合语音由目标语音与 5 个方向的白噪声混合而成(后称

“白”).实验中用改变干扰源幅度的方法来控制信噪比, 并且将纯净语音(后称“干净”)、混合语音(后称“污染”)、处理后语音分别输入由干净语音训练所得说话人系统中进行识别测试, 说话人的模型参数由干净语音训练所得, 测得干净语音的识别率为 93%.多次实验后取平均结果如表 1 所示.

表 1　不同前端处理的说话人识别结果

测试语音	信噪比					
	10dB (多)	5dB (多)	0dB (多)	5dB (多)	10dB (白)	0dB (白)
污染	32%	61%	67%	72%	31%	72%
谱减法	34%	62%	70%	77%	52%	81%
时频掩蔽法	74%	75%	78%	83%	72%	85%

由表 1 可得以下结论:

(1)说话人识别系统的性能在多声源环境下急剧下降, 目标说话人的识别率会因干扰源类型和信噪比大小有不同程度的降低, 识别率平均降低了 37%.

(2)传统谱减法处理平稳噪声时效果不错, 如干扰源仅为白噪声时, 混合信号经谱减法处理后识别率平均提高了 15%.但当干扰源为非平稳的语音或其他声源时, 其语音增强效果就很差了, 识别率仅提高 2-3%.

(3)本文方法消除多声源干扰效果明显.在各种情况下(不同干扰源和信噪比), 经过本文方法增强后的语音识别率都有显著提高, 平均增加 22%.尤其是处理干扰源为多种声源时效果明显, 如 SNR=-10dB 的多声源混合语音, 增强后识别率提高 42%.

实验三: 干扰源数目不同时的语音增强对比实验

用本文语音增强方法进行了干扰源数目不同时的说话人识别对比实验.实验中的干扰源在 1 个男声、2 个女声以及 1 个敲门声和 1 个诺基亚短信音中随机抽取.实验中各干扰源幅和目标语音的幅度大小基本一致.统计多次实验, 结果如表 2.

表 2　不同干扰源数量下目标语音增强的对比实验

测试语音	干扰源个数				
	1个	2个	3个	4个	5个
污染	68%	62%	49%	38%	32%
增强后	81%	79%	78%	75%	74%

由表 2 可以得出以下结论:

(1) 干扰源的数目越多, 污染语音的识别率就越低.

(2)经本文语音前端处理, 说话人识别率平均提高 27.6%.且处理后的说话人系统识别率能保持在 70%以上.

(3)由于该方法是建立在语音稀疏性和方位可区分性的基础上, 因此, 随着干扰源数目的进一步增加, 该方法性能会有所下降.

实验四: 目标语音方位跟踪实验

定义 s1、主麦连线和主、从麦连线的夹角为 θ, 设目标声源初始方位 θ 为 0O, 实验中设实际目标声源方位 θ 可在-5o~5o 范围内变动, 如图 3 虚线部分所示.

由公式(6)对目标语音方位信息进行跟踪, 并将根据初始方位信息的提取(后称“初步提取”)与更新方位信息后的提取(后称“跟踪提取”)得到的语音分别进行说话人识别, 多次实验取平均结果如表 3 所示.

表 3　目标语音方位跟踪的说话人识别对比实验

测试语音	信噪比			
	-10dB	-5dB	0dB	5dB
污染	32%	61%	67%	72%
初步提取	48%	65%	73%	78%
跟踪提取	73%	74%	78%	82%

由表 3 可得, 初始目标语音方位信息与未处理的污染语音相比, 其识别率平均只提高 8%.进行语音目标方位信息实时跟踪后, 目标语音得到了进一步的提取, 其识别率较跟踪前平均提高了 11%.

5　小结

针对多声源干扰环境下说话人识别系统性能急剧下降的问题, 本文提出了基于双通道利用目标语音方位信息进行时频掩蔽提取目标语音的前端处理方法.针对目标声源方位可能出现的小范围变动, 还提出了实时跟踪目标声源方位的处理.建立了基于 MFCC 和 GMM 说话人识别系统, 在一个多达六个声源的仿真会议系统环境下进行说话人识别实验.实验结果表明, 在多声源干扰情况下, 经本文的前端处理后目标语音的识别率平均提高了 25%左右.仿真实验所取得的良好效果显示了该算法的可行性.当然, 随着声源数目的增加以及信噪比的进一步降低, 本系统性能会有所下降.但以上实验已证明该算法对提高系统鲁棒性、增强说话人识别的效果是很好的.

参考文献

1 尹许梅.基于 MFCC 和矢量量化的说话人识别算法研究[学

位论文].长沙:湖南大学,2011.
2 檀蕊莲.动态事件规整算法与说话人识别技术研究.科技资讯,2010,8.
3 曹洁,潘鹏.基于 GMM 的说话人识别技术研究.计算机工程与应用,2011,47(11):114–117.
4 张学峰,王芳,夏萍.融合 LPC 与 MFCC 的特征参数.计算机工程,2011,37(4).
5 展领,景新幸.基于 VQ-MAP 和 SVM 融合的说话人识别系统.计算机工程与应用,2011,47(13).
6 陈仁林.基于神经网络的说话人识别算法研究[学位论文].银川:宁夏大学,2012.
7 夏秀渝,何培宇.基于声源方位信息和非线性时频掩蔽的语音盲提取算法.声学学报,2013,38(2).
8 Wang G, Yu HY, Shen ZX, et al. Fast convergence parameter estimation method based on Expectation-Maximum algorithm. Journal of Jilin University(Engineering and Technology Edition), 2013, 43(2): 532–537.
9 Ramezani A, Moshiri B, Khan AR, et al. Design of an adaptive maximum likelihood estimator for key parameters in macroscopic traffic flow model based on expectation maximum algorithm. IET Science, Measurement & Technology, 2011, 5(5): 189–197.
10 Wang G, Yu HY, Shen ZX. An Improved Symbol Detection Algorithm Based on Expectation-Maximum. Information Computing and Applications. Springer Berlin Heidelberg, 2013: 467–476.
11 Katsamanis A, Black MP, Georgiou PG, et al. SailAlign: Robust long speech-text alignment. Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research in Phonetic Science. 2011. 28–31.