·开发研究与工程应用· 文章编号：1000-3428(2018)02-0316-06 文献标志码：A 中图分类号：TP391

基于计算听觉场景分析的说话人转换检测

杨登舟[1,2]，刘 加[3]，夏善红[1]

（1. 中国科学院电子学研究所，北京 100190；2. 中国科学院大学，北京 100049；
3. 清华大学 电子工程系，北京 100084）

摘 要： 在短时语音说话人快速转变的说话人转换检测中，用于训练说话人模型的连续语音较短导致模型不稳健，致使说话人转换检测的性能较差。为此，提出一种新的说话人转换检测方法。借鉴人耳听觉处理机制将语音信号分解为多个子带，可以得到准确的浊、清音边界，实现对零散清、浊音子段的拼接。利用贝叶斯信息准则判决语音子段间的疑似转换点，并运用音高特征做区间验证。实验结果表明，该方法在平均语音子段时长为1.34 s的极短语音条件下，可使说话人转换检测的等错率降至 23.2%，$F1$ 值达到 70%。

关键词： 说话人转换检测；计算听觉场景分析；伽马通能量倒谱系数；音高；贝叶斯信息准则

中文引用格式： 杨登舟，刘 加，夏善红. 基于计算听觉场景分析的说话人转换检测[J]. 计算机工程，2018，44(2)：316-321.

英文引用格式： YANG Dengzhou，LIU Jia，XIA Shanhong. Speaker Change Detection Based on Computational Auditory Scene Analysis[J]. Computer Engineering，2018，44(2)：316-321.

Speaker Change Detection Based on Computational Auditory Scene Analysis

YANG Dengzhou[1,2]，LIU Jia[3]，XIA Shanhong[1]

（1. Institute of Electronics，Chinese Academy of Sciences，Beijing 100190，China；
2. University of Chinese Academy of Sciences，Beijing 100049，China；
3 Department of Electronic Engineering，Tsinghua University，Beijing 100084，China）

【Abstract】 In Speaker Change Detection (SCD) of rapid conversion condition with short speech segment，speaker models training from deficient speech frames of a speaker are not rubust enough，and SCD performance is less satisfied. Therefore，a new SCD method based on Computational Auditory Scene Analysis (CASA) is proposed. The speech signal is decomposed into a number of narrow sub-band signals owing to the auditory processing mechamism of human ears. Accurate voiced speech and unvoiced speech boundaries are obtained，voice sub-segments is spliced from scattered voice and unvoiced sub-segments. Speaker change points are determined between the speaker voice sub-segments by Bayesian Information Criterion (BIC)，pitch features extracted from voiced portion are used to verify region. Experimental results show that Equal Error Rate (EER) of SCD can be reduced to 23.2%，which corresponding to 70% of the $F1$-value，in the rapid conversion situation of average 1.34 s speech sub-segment.

【Key words】 Speaker Change Detection (SCD)；Computational Auditory Scene Analysis (CASA)；Gammatone Energy Cepstral Coefficients (GECC)；pitch；Bayesian Information Criterion (BCI)

DOI： 10.3969/j.issn.1000-3428.2018.02.054

0 概述

随着电子通信和计算机技术的快速发展，大量的语音数据被存储，如何快速地建立语音检索是亟需解决的问题。说话人转换检测（Speaker Change Detection，SCD），也称说话人分割[1]，是语音信号处理中的一项实用技术，从一段语音中将不同说话人说话的时刻检测出来，将语音分割出满足要求的片段，可以很方便地建立索引，为信息的进一步处理提供便利[2-3]。

语音切分类问题可以抽象成模型判别问题，用特定长度的窗（定长窗或者变长窗）扫描整段语音，当窗内左右两部分的语音之间的差异达到某个阈值，认为在窗左半部分和右半部分发生了明显的改变，有理由怀疑此处语音的声源发生了变化[4]。在说话人转换检测的研究中，窗左右两侧语音之间的差异度量方式主要有模型差异、参数差异以及模型和参数相结合[5-6]这三大类。在基于模型的方法中，

基金项目： 国家自然科学基金“噪声和短语音条件下的说话人识别”(61370034)。

作者简介： 杨登舟(1986—)，男，博士研究生，主研方向为说话人识别；刘 加，教授、博士；夏善红，研究员、博士。

收稿日期： 2017-02-14 **修回日期：** 2017-03-20 **E-mail：** yangdengzhou@sina.com

从训练数据中挑选出不同的发声源,训练出各自的模型,同时训练出所有声源的全局模型,通过分析全局模型和个体模型的不同之处,得到模型间转化关系或者找到可区分的模型差异,常用的模型包括通用背景模型(Universal Background Model,UBM)、样本说话人模型(Sample Speaker Model,SSM)、隐马尔科夫模型(Hidden Markov Model,HMM)。基于参数的方法,使用较多的特征主要包括时域短时能量、过零率、频域的子带能量、倒谱特征、线性预测系数等。通常使用差异度量准则有贝叶斯信息准则(Bayesian Information Criterion,BIC)、广义似然比(Generalized Likelihood Ratio,GLR)、KL 散度(Kullback-Leibler divergence)、归一化交叉似然比(Normalized Cross Likelihood Ratio,NCLR)等。

在说话人识别问题中,由于事先可以获取训练数据,可以事先训练出多个不同的说话人模型,在判决阶段只要将一段语音的特征和所有参考模型做比较,和哪一个更近就判别成哪个,在闭集测试中,性能较好[7]。而说话人转换检测比说话人识别难度大,主要难点在于对一段语音做切分任务,并不会提供该语音中所包含的说话人的训练语料,因此不能准确获取到说话人的模型,特别是在短时说话人迅速转变的对话口语语音中完成稳定说话人建模难度更大,需要挖掘短时说话人差异区分性大、能全面描述说话人发声特性的特征。计算听觉场景分析(Computational Auditory Scene Analysis,CASA)[8]根据听觉生理学和听觉心理学的研究成果,利用计算机模拟人耳耳蜗的听觉处理机制来处理接收到的语音信息,该理论能够较好地解决诸如同信道语音分离问题,充分利用语音的周期性和短时连续性2 个重要的线索来区分不同的声源。

本文提出一种基于听觉场景分析的说话人转换检测方法,将语音分割成相邻的若干语音子段,提取伽马音能量倒谱系数特征,在贝叶斯信息准则的判决下得到初始说话人转换点,最后利用浊音的基频特征对漏检和错检的转换点进行后处理,以达到较好的检测结果。

1　计算听觉场景分析

人每天在各种复杂的声学环境中倾听语音,提取需要的信息,可以从周围嘈杂的多人说话环境中锁定自己感兴趣的声源对象,只要信噪比合适,人耳可以将目标声源的声音从背景语音中完全分离出来,并且做得非常出色,取决于人类具有听觉场景分析(Auditory Scene Analysis,ASA)[9]的能力。

人耳的耳蜗基底膜就好像是一个初级的频率分析器,可以将声音中的各种频率在基底膜上的位置进行编码。当基底膜上下振动,其柯蒂氏器(Corti)也随之产生相同的振动模式,并促使毛细胞纤毛发生弯曲形变,毛细胞去极化并在其顶部产生耳蜗电位,该电位会引起毛细胞底部神经纤维的应激反应,释放出化学物质,引导神经末梢兴奋,传输至中枢神经。人耳除了具有频率分析特性,对声波强度的编码也非常高效,通过神经单元兴奋后发放神经冲动的数量来确定强度。

1.1　Gammatone 滤波器组模型模拟耳蜗的频率分析

听觉场景分析中将原始语音信号拆分成多个子带信号的过程是通过 Gammatone 滤波器组[10]来实现的。Gammatone 滤波器组是由一系列不同带宽不同中心频率的带通滤波器组成,Gammatone 滤波器的冲激响应为:

$$g_c(t)=\begin{cases}t^{\tau-1}\exp[-2\pi t\cdot B(f_c)]\cos(2\pi f_c t+\varphi), & t>0\\ 0, & \text{其他}\end{cases} \tag{1}$$

其中,τ 是滤波器的阶数,φ 是初始相位,$B(f_c)$是滤波器组的带宽,f_c 是中心频率。当 $\tau=4$ 时和人耳听觉滤波器非常吻合。滤波器的带宽由中心频率对应的等价直角带宽(Equivalent Rectangular Bandwidth,ERB)确定:

$$ERB(f)=24.7\times(4.37f/1\,000+1) \tag{2}$$

$$B(f)=1.019\times ERB(f) \tag{3}$$

线性频率 f 和“ERB-rate”尺度频率 F_{ERB} 的换算关系为:

$$F_{\text{ERB}}(f)=21.4\times\lg(0.004\,37f+1) \tag{4}$$

将线性频率 80 Hz ~5 000 Hz 转化为“ERB-rate”尺度频率,并在“ERB-rate”尺度下均匀取出 128 个,生成子带数 $C=128$ 的 Gammatone 滤波器组。将原始语音信号 $s(t)$通过滤波器组滤波,输出 C 个子带信号$u_c(t)$:

$$u_c(t)=s(t)\times g_c(t),c=1,2,\cdots,C \tag{5}$$

1.2　毛细胞触发模型模拟耳蜗的强度分析

原始语音信号 $s(t)$经过 Gammatone 滤波器滤波后得到 $u_c(t),c=1,2,\cdots,C$(为表述方便,下文将省略子带下标 c,并不影响理解)。将 $u(t)$经过 Meddis 毛细胞模型[11],可以得到描述听觉神经触发概率的信号 $v(t)$。毛细胞触发概率的计算过程通过以下 4 个方程完成:

$$k(t)=\begin{cases}g\cdot\dfrac{v(t)+A}{v(t)+A+B}, & v(t)+A>0\\ 0, & \text{其他}\end{cases} \tag{6}$$

$$\frac{\mathrm{d}q(t)}{\mathrm{d}t}=y\cdot(1-q(t))+x\cdot w(t)-k(t)q(t) \tag{7}$$

$$\frac{\mathrm{d}c(t)}{\mathrm{d}t}=k(t)q(t)-l\cdot c(t)-r\cdot c(t) \tag{8}$$

$$\frac{\mathrm{d}w(t)}{\mathrm{d}t}=rc(t)-x\cdot w(t) \tag{9}$$

在式(6)~式(9)中,g、r、l、h、A、B、x、y 是模型常数,$q(t)$、$c(t)$、$w(t)$是中间变量,在毛细胞传导模型中有具体意义,听觉末梢发放概率 $v(t)=h\cdot c(t)$。

2 区分性特征提取

2.1 伽马通能量倒谱系数

在语音识别、说话人识别和语种识别中都可以见到梅尔频率倒谱系数(Mel-frequency Cepstral Coefficients,MFCC)[12]发挥的重要作用。梅尔频率倒谱系数是将语音帧的快速傅里叶变换(Fast Fourier Transformation,FFT)频谱通过相互交叠且中心频率沿梅尔频率线性分布的 24 个三角滤波器组,对三角频窗内的能量计算对数,对数谱计算离散余弦变换(Discrete Cosine Transform,DCT)得到梅尔频率倒谱系数。伽马通频率倒谱系数[13]借鉴了梅尔频率倒谱系数特征提取的原理。MFCC 中对能量求对数得到倒谱,在 GFCC 中变成了计算响度压缩,本文建立了一个介于 GFCC 和 MFCC 之间的特征,伽马通能量倒谱系数(Gammatone Energy Cepstral Coefficients,GECC),它和 GFCC 的提取不同之处如图 1 所示,GECC 仅在于利用响度和能量的差异。

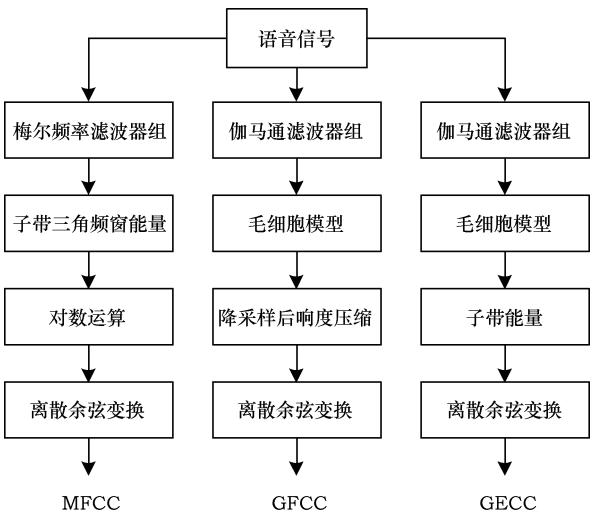

图 1 特征提取流程

对毛细胞触发模型的输出 $v(t)$进行 100 Hz 降采样,得到分帧信号 $w(m)$,$m=1,2,\cdots,M$,M 是帧数。各帧能量记为 $G_c(m)$,对 $G_c(m)$,$c=1,2,\cdots,C$ 计算 M 阶的离散余弦变换来降低 M 个子带间的数据相关性,取前 D 维的数据,得到 GECC 特征:

$$GECC_{m,k}=\sqrt{\frac{2}{C}}\sum_{c=0}^{C-1}G_c(m)\cdot\cos\left(\frac{k\pi}{2C}\cdot(2c+1)\right)$$
$$m=1,2,\cdots,2M,k=0,1,\cdots,D-1 \tag{10}$$

2.2 音高

从人的发音结构和语音的形成过程,可以把语音信号等效成激励-滤波器模型,声门产生激励,声门激励满足准周期性就可以产生有固定谐波结构的语音信号,这类语音称之为浊音[14];将不具有周期性且与噪声类似的声门激励生成的语音信号称为清音。声带、嘴唇、口腔的作用可以等效成声道滤波器响应。声道滤波器反映的主要是语义信息(音素,词汇),说话人的特性主要取决于声门激励。浊音的基频在听觉的感受就表现在音高上,每个人的音高略有不同,分布在 50 Hz~500 Hz 的范围内,男性的音高比女性要低,成人的音高比小孩的要低。音高的差异可以作为说话人区分的一个重要特征。

对应某个特定子带 c、时间帧 m 内的毛细胞触发输出 $v(t)$的自相关:

$$A(m,c,\tau)=\sum_{k=-N/2}^{N/2-1}v_c(mN/2-k)\times v_c(mN/2-k-\tau)\times h^2(k+N/2) \tag{11}$$

其中,N 是窗长。对第 m 帧的自相关矩阵 $\boldsymbol{A}_m(c,\tau)$沿着频率子带计算累加相关 $\hat{\boldsymbol{A}}_m(\tau)$:

$$\hat{\boldsymbol{A}}_m(\tau)=\sum_{c=1}^{C}\boldsymbol{A}_m(c,\tau) \tag{12}$$

通常人类的基音范围在 80 Hz ~500 Hz,对应的延时区间是 $\tau\in[2\text{ ms},12.5\text{ ms}]$,通过搜索最大值得到音高 P_m:

$$P_m=\frac{1\ 000}{\underset{2<\tau<12.5}{\operatorname{argmax}}\hat{A}_m(\tau)} \tag{13}$$

对检测的音高序列做平滑处理,得到连续的基音轨迹。

3 说话人转换检测系统

本文基于听觉场景分析的说话人转换检测由听觉外围处理、特征提取、转换点判决 3 个部分组成,如图 2 所示。听觉外围处理将语音信号经由伽马通滤波器组滤波,再用 Meddis 毛细胞触发模型得到听觉神经末梢的发放概率。对发放概率按帧能量检测对应帧是浊音、清音还是静音,各帧的属性标记以后,得到浊音的连续片段,称为子段,记为 S。对所有相邻的子段对(S_i,S_{i+1})进行贝叶斯信息准则判决,得到分割初步判决结果。经过贝叶斯信息准则判决后,已经得到一定数量的说话人转换点,区间验证的作用是试图利用音高信息,对可疑的转化点进行剔除,并尝试找回已经被遗漏的转化点。

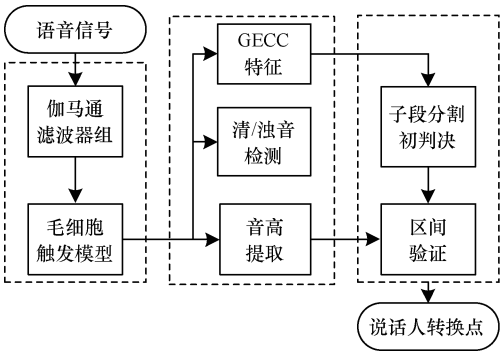

图 2　基于听觉场景分析说话人转换检测系统

3.1　清浊音检测

对毛细胞触发模型的输出 $v(t)$ 进行短时分帧,并计算在各子带内每帧的能量图 $E(c,m)$。沿时间轴方向对子带能量进行能量规整:

$$\hat{E}(c,m)=\frac{M\cdot E(c,m)}{\sum_{m=1}^{M}E(c,m)},c=1,2,\cdots,C,m=1,2,\cdots,M \tag{14}$$

对规整后的能量图 $\hat{E}(c,m)$ 进行二值化处理:

$$E_b(c,m)=\begin{cases}1,\hat{E}(c,m)>th_1\\0,\hat{E}(c,m)<th_0\end{cases}$$
$$c=1,2,\cdots,C,m=1,2,\cdots,M \tag{15}$$

其中,th_0 为低能量判决门限,th_1 为高能量判决门限。

首先检测浊音,在频率小于 950 Hz 的低频区(中心频率离 950 Hz 最近的子带记为 C_s),浊音一定会有能量中心,而清音或者背景噪声在此区域内的能量与浊音的谐波能量相比,几乎可以忽略不计[15]。按以下约束对各帧进行标记:

$$UVS(m)=\begin{cases}V,(\exists c,E_b(c,m)=E_b(c+1,m)=1)\\X,(\forall c,E_b(c,m)=E_b(c+1,m)=1)\end{cases}$$
$$c=1,2,\cdots,C_s,m=1,2,\cdots,M-1 \tag{16}$$

其中,V 表示浊音,X 表示未定。标记为 V 的所有帧记为集合 $setV$,标记为 X 的所有帧记为集合 $setX$。

清音在高频区(频率大于 950 Hz)虽然没有能量中心,但和静音相比有明显的能量分布,在 $setX$ 中各帧按照以下约束进行标记:

$$UVS(m)=\begin{cases}U,\sum_{c=C_s}^{C}E_b(c,m)>(C-C_s)/4\\S,\sum_{c=C_s}^{C}E_b(c,m)\leqslant(C-C_s)/4\end{cases}$$
$$c=C_s\cdots C,m\in setX \tag{17}$$

其中,U 表示清音,S 表示静音,清音帧的集合记为 $setU$,静音帧的集合记为 $setS$。

3.2　分割初判决

对分帧信号标记浊音、清音、静音以后,可以得到语音的连续片段,称为子段,记为 S,$S_i=\{x_1,x_2,\cdots,x_{M_i}\}$,$\boldsymbol{x}_j$ 是第 j 帧的特征矢量,M_i 是第 i 段的帧数。说话人 A 说了一串语音,该段语音中包含若干 A 的子段,然后转变成 B 的若干子段。属于同一说话人的子段之间相似度较高,而不同说话人之间的相似度较低。对完整语音按照说话人不同进行分割,就可以通过检验相邻的子段对 (S_i,S_{i+1}),对以下 2 种假设做出判决:

$$\begin{cases}H_0:S_i,S_{i+1}\text{是来自于同一个人}\\H_1:S_i,S_{i+1}\text{来自于不同的人}\end{cases} \tag{18}$$

这是模型选择问题,如果 $p(H_0)>p(H_1)$,则假设 H_0 成立,反之亦然。对子段的特征训练单高斯模型,$S_i\sim N(\mu_i,\Sigma_i)$,$S_{i+1}\sim N(\mu_{i+1},\Sigma_{i+1})$,$S_i\cup S_{i+1}\sim N(\mu,\Sigma)$,单高斯模型对特征进行似然度打分:

$$L_i=\sum_{m=1}^{M_i}\lg(x_m|N(\mu_i,\Sigma_i)) \tag{19}$$

$$L_{i+1}=\sum_{m=1}^{M_{i+1}}\lg(x_{i+1,m}|N(\mu_{i+1},\Sigma_{i+1})) \tag{20}$$

$$L=\sum_{m=1}^{M_i}\lg(x_{i,m}|N(\mu,\Sigma))+\sum_{m=1}^{M_{i+1}}\lg(x_{i+1,m}|N(\mu,\Sigma)) \tag{21}$$

此时判决结果可以表示为:

$$H=\begin{cases}H_0,L-(L_i+L_{i+1})\geqslant th\\H_1,L-(L_i+L_{i+1})<th\end{cases} \tag{22}$$

贝叶斯信息准则(BIC)在模型选择问题上具有较好的性能,并有广泛的应用[16],贝叶斯信息准则满足:

$$th=\frac{\lambda}{2}\left(D+\frac{1}{2}D(D+1)\right)\lg(M_i+M_{i+1}) \tag{23}$$

其中,D 是 GECC 特征维度,λ 是调节因子,一般设为 1 即可。

对所有相邻的子段对 (S_i,S_{i+1}) 进行贝叶斯信息准则判决,得到分割初步判决结果。

3.3　区间验证

经过贝叶斯信息准则判决后,已经得到一定数量的说话人转换点,区间验证的作用是试图利用音高信息,对可疑的转化点进行剔除,并尝试找回已经被遗漏的转化点。

根据贝叶斯信息准则判决产生的相邻转换点之间的时间帧区间内存在的子段个数 N,采用不同的处理策略。

当 $N=1$ 时,两相邻转换点之间有一个孤立子段,此时判断孤立子段两侧转换点之间的时间间隔是否足够小,如果小于 1 s 且孤立子段的音高和左右两侧有一边比较吻合,就剔除掉吻合度较低的那一侧的转化点。当 $1<N<5$ 时,不做处理。当 $N>5$ 时,从距离两侧转换点最近的子段开始,逐步向中间的子段逼近描绘基音轨迹,哪一侧轨迹吻合度高就先向中间过渡一个子段,继续逼近,直到两边汇合,如果最后汇合处两侧的基音轨迹存在明显跳变,就在汇合处补充一个转化点。

4 实验设置与数据分析

测试数据库选用 conTIMIT 数据集[17],一共包含 55 条语音波形文件,统计语音时长 3 675 s,有效分割点数 1 071 个,平均每个说话人段长 3.29 s,最短 1.14 s,最长 11.75 s,标准差 1.75 s。语音采样频率为 16 000 Hz,实验中语音分帧帧长 20 ms,帧移10 ms,GFCC 特征选择 23 维基本特征加一阶差分特征,MFCC 特征选择 13 维基本特征加一阶差分特征。

对说话人转化检测的性能评价,用等错率和 $F1$ 值。当虚警率(False Alarm Rate, FAR)和漏报率(Miss Detection Rate, MDR)相等时,得到等错率(Equal Error Rate, EER):

$$FAR = \frac{FA}{FA + GT} \tag{24}$$

$$MDR = \frac{MD}{GT} \tag{25}$$

$$EER = FAR(\arg\min_{\lambda} | FAR(\lambda) - MDR(\lambda) |) \tag{26}$$

用召回率($Recall$)和准确率($Precision$)计算 $F1$ 值:

$$Recall = \frac{GD}{GT} \tag{27}$$

$$Precision = \frac{GD}{GD + FA} \tag{28}$$

$$F1 = \frac{2Recall \times Precision}{Recall + Precision} \tag{29}$$

其中,FA 是转换点虚报个数,MD 是未检测出的转换点个数,GT 是实际的转换点个数,GD 是正确检测出的转换点个数。

在数据集上用贝叶斯信息准则作为距离准则得到说话人转换点,并和加权距离度量(Weighted Distance Measure, WDM)[18]准则检测的性能做对比。表 1 给出浊音子段、清音子段、语音子段(包含浊音和清音)的段长统计信息。分别计算分割边界转换点的漏报率-虚警率曲线,如图 3 ~ 图 5 所示,对应的等错率结果如表 2 所示。单独计算浊音子段,BIC 和 WDM 两种方法的转换点与检测点都是非常差的,80% 的子段段长落在 0.1 s ~ 0.5 s 范围内,造成 BIC 失效。在同样极短时间的条件下,清音子段的表现比浊音好得多。把相邻浊音和清音连接成语音子段,段长平均达到 1.34 s,与说话人识别的最低 2 s 的要求已经比较接近,GECC 特征在 BIC 准则下达到最好检测效果,等错率降到 26.8% 。

表 1 浊音、清音、语音段长统计 s

子段类别	平均值	最小值	最大值	标准差
浊音	0.281	0.1	1.32	0.172
清音	0.145	0.1	0.90	0.061

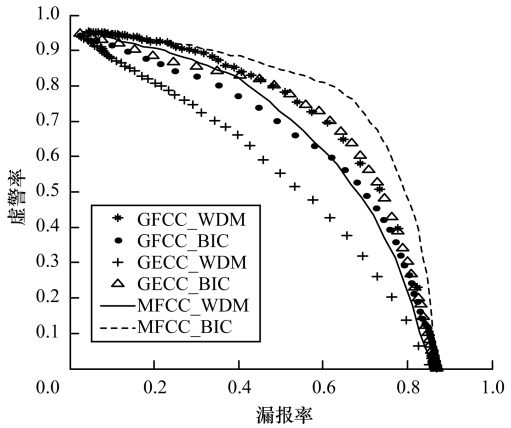

图 3 浊音子段(V-S)虚警率和漏报率曲线

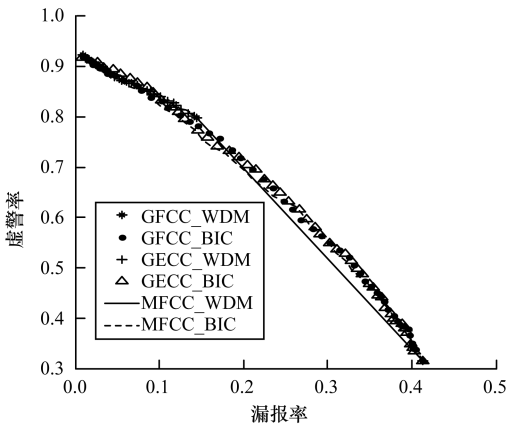

图 4 清音子段(U-S)虚警率和漏报率曲线

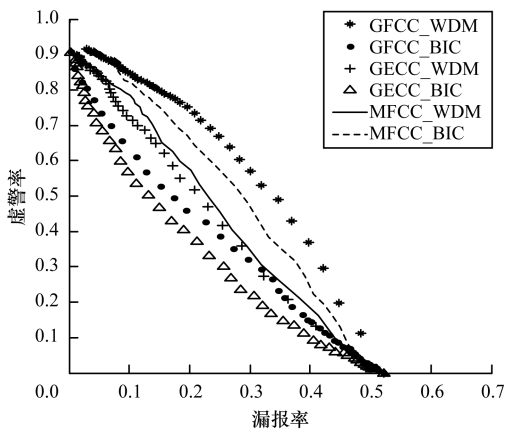

图 5 语音子段(UV-S)虚警率和漏报率曲线

表 2 不同特征、不同子段类别切分等错率 %

准则	V-S			U-S			UV-S		
	GFCC	GECC	MFCC	GFCC	GECC	MFCC	GFCC	GECC	MFCC
WDM	64.8	52.5	60.9	36.4	36.4	36.4	38.4	29.8	31.3
BIC	60.9	65.7	70.5	38.6	38.8	38.7	30.9	26.8	35.1

浊音子段的音高特征在说话刚开始时会出现跳高音陡降至稳态基频区的过程,在说话结束时几乎都会发生从稳态基频降频的收尾效应,但在同一个说话人语音内跳跃幅度比较平稳,在区间验证的过程中利用这一信息,既可以剔除掉一些虚警转换点,也可以找回一些已经漏掉的转换点,从表 3 可以看到最终的等错率可以下降到 23.2%,相应的 $F1$ 值为 70.0%。

表 3　结合音高补偿后的检测性能　%

准则	EER			$F1$		
	GFCC	GECC	MFCC	GFCC	GECC	MFCC
WDM	34.5	26.1	26.9	55.7	66.4	65.5
BIC	27.7	23.2	30.9	64.4	70.0	60.4

5　结束语

在基于听觉场景分析的说话人转变检测中,由于伽马通滤波器和毛细胞模型对人耳听觉系统的模拟,可以将语音信号按照人的听觉感知对各个频带进行精细划分,得到准确的清音和浊音信息以及稳健的基频轨迹。基于此,本文一种提出基于听觉场景分析的说话人转换检测方法。将语音分割成相邻的若干语音子段(包含清音、浊音、极短静音),提取伽马通能量倒谱系数特征,在贝叶斯信息准则的判决下得到初始说话人转换点,最后利用浊音的基频特征对漏检和错检的转换点进行后处理,最终得到较好的检测结果。在 conTIMIT 数据集上的测试结果表明,不做音高检测,最优性能是选用 GECC 特征在 BIC 准则下等错率达到 26.8%,利用音高信息,得到 GFCC 特征在 BIC 准则下性能提高到 23.2%,GECC 和 GECC 特征的性能优于 MFCC,BIC 准则优于 WDM 准则,在短时语音说话人快速转变的口语对话环境中,即使无法训练说话人模型,也可以达到一定的检测准确性。

参考文献

[1] BAZYAR M, SUDIRMAN R. A New Speaker Change Detection Method in a Speaker Identification System for Two-speakers Segmentation[C]//Proceedings of 2014 ACM Symposium on Computer Applications and Industrial Electronics. New York, USA: ACM Press, 2014:141-145.

[2] MALEQAONKAR A S, ARIYAEEINIA A M. Efficient Speaker Change Detection Using Adapted Gaussian Mixture Models[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(6):1859-1869.

[3] ZAHID S, HUSSAIN F, RASHID M, et al. Optimized Audio Classification and Segmentation Algorithm by Using Ensemble Methods[J]. Mathematical Problems in Engineering, 2015(11):209-214.

[4] 郑继明,张　萍. 改进的 BIC 说话人分割算法[J]. 计算机工程,2010,36(17):240-242.

[5] KOTTI M, BENETOS E, KOTROPOULOS C. Computationally Efficient and Robust BIC-based Speaker Segmentation[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2008, 16(5):920-933.

[6] YANG J, HE Q, LI Y, et al. Speaker Change Detection Based on Mean Shift[J]. Journal of Computers, 2013, 8(3):638-644.

[7] WU Z, EVANS N, KINNUNEN T, et al. Spoofing and Countermeasures for Speaker Verification: A Survey[J]. Speech Communication, 2015, 66(1):130-153.

[8] 张学良,刘文举,李　鹏,等. 改进谐波组织规则的单通道浊语音分离系统[J]. 声学学报,2011,36(1):88-96.

[9] CUSACK R, DECKS J, AIKMAN G, et al. Effects of Location, Frequency Region, and Time Course of Selective Attention on Auditory Scene Analysis[J]. Journal of Experimental Psychology: Human Perception and Performance, 2004, 30(4):643-656.

[10] MAKA T. Change Point Determination in Audio Data Using Auditory Features[J]. International Journal of Electronics and Telecommunications, 2015, 61(2):185-190.

[11] MEDDIS R. Simulation of Mechanical to Neural Transduction in the Auditory Receptor[J]. The Journal of the Acoustical Society of America, 1986, 79(3):702-711.

[12] LI L. Performance Analysis of Objective Speech Quality Measures in Mel Domain[J]. Journal of Software Engineering, 2015, 9(2):350-361.

[13] KAUR G, SINGH D, RANI P. Robust Speaker Recognition Biometric System a Detailed Review[J]. Emerging Research in Management & Technology, 2015, 4(5):281-288.

[14] 王　民,任雪妮,孙　洁. 一种高效的基音检测与评估算法[J]. 计算机工程与应用,2014,50(14):126-132.

[15] 胡　瑛,陈　宁. 基于小波变换的清浊音分类及基音周期检测算法[J]. 电子与信息学报,2008,30(2):353-356.

[16] CHEN S, GOPALAKRISHNAN P. Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion[C]//Proceedings of Broadcast News Transcription and Understanding Workshop. San Francisco, USA: Morgan Kaufmann Publishers, 1998:127-132.

[17] SEO J S. Speaker Change Detection Based on a Graph-partitioning Criterion[J]. The Journal of the Acoustical Society of Korea, 2011, 30(2):80-85.

[18] KWON S, NARAYANAN S S. Speaker Change Detection Using a New Weighted Distance Measure[C]//Proceedings of the 7th International Conference on Spoken Language Processing. Washington D. C., USA: IEEE Press, 2002:2537-2540.

编辑　索书志