基于子带双特征的自适应保留似然比鲁棒语音检测算法

何伟俊 贺前华* 吴俊峰 杨继臣
(华南理工大学电子与信息学院 广州 510641)

摘 要：为了进一步提高低信噪比下语音激活检测(VAD)的准确率，该文提出一种基于子带双特征的自适应保留似然比鲁棒语音激活检测算法。算法采用子带归一化最大自相关函数与子带归一化平均过零率双重特征设置频率分量似然比的保留权值，同时利用已过去固定时长的 VAD 判决结果及对应的子带特征参数自适应地估计似然比的保留阈值。实验结果表明，此算法的 VAD 检测准确率相比原保留似然比算法在 10 dB, 0 dB 和−10 dB 平稳白噪声下分别提高了 1.2%, 7.2%和 8.1%，在 10 dB 和 0 dB 非平稳 Babble 噪声下分别提高了 1.6%和 3.4%。当其被用于 2.4 kbps 低速率声码器系统时，合成语音的感知语音质量评价(PESQ)比原声码器系统在白噪声下提高了 0.098~0.153，在 Babble 噪声下提高了 0.157~0.186。

关键词：语音激活检测；似然比；低信噪比；子带过零率
中图分类号：TN912.3 **文献标识码**： A **文章编号**：1009-5896(2016)11-2879-08
DOI: 10.11999/JEIT160157

Adaptively Reserved Likelihood Ratio-based Robust Voice Activity Detection with Sub-band Double Features

HE Weijun HE Qianhua WU Junfeng YANG Jichen
(*School of Electronic and Information Engineering*, *South China University of Technology*, *Guangzhou* 510641, *China*)

Abstract: In order to improve the correct rate of Voice Activity Detection (VAD) in low Signal Noise Ratio (SNR) environment, the paper presents an adaptive reserved likelihood ratio VAD method, which is based on sub-band double features. The method employs sub-band auto correlate function and sub-band zero crossing rate in the process of setting reserved weight. Reserved threshold is estimated adaptively according to the passed VAD results and their sub-band feature parameters. The experiment shows its promising performance in comparison with similar algorithms, the VAD correct rate is improved by 1.2%, 7.2%, and 8.1% respectively in 10 dB, 0 dB, and −10 dB stationary white noisy environment, 1.6% and 3.4% respectively in 10 dB and 0 dB non-stationary Babble noisy environment. The method is also applied to 2.4 kbps low bit rate vocoder and the Perceptual Evaluation of Speech Quality (PESQ) is improved by 0.098~0.153 in white noisy environment, 0.157~0.186 in Babble noisy environment.

Key words: Voice Activity Detection (VAD); Likelihood ratio; Low signal noise ratio; Sub-band zero crossing rate

1 引言

语音激活检测(Voice Activity Detection, VAD)目的在于从信号中区分出语音信号与非语音信号。在语音识别系统中，准确的VAD判决可提高识别率并节省处理时间[1]。在低速率语音编码系统(如ITU-T G.729, MELP)中，根据VAD判断当前信号帧是否有语音采用不同的编解码模式，从而在不影响合成语音质量的前提下降低编码速率[2]。传统的语音激活检测主要是基于短时能量、过零率、谱熵、LPC参数、倒谱特征、高阶统计量等语音特征参数的方法[3]，它们在高信噪比条件下具有令人满意的效果。

为了解决低信噪比下VAD检测问题，文献[4]提出了基于似然比检验(Likelihood Ratio Test, LRT)的VAD算法，此算法利用高斯统计模型对信号的傅里叶变换系数按语音与非语音两种假设进行建模，通过似然比检验法评估两种统计模型与当前观测数据的适配程度，从而作出VAD判决。文献[5]在2001年提出基于平滑统计似然比的改进算法，文献[6,7]在2005年和2007年相继提出基于多观测值的似然比VAD算法以及基于多假设多观测值的似然比VAD算法，它们主要利用长时语音信息，借助连续多个

收稿日期：2016-02-04；改回日期：2016-06-27；网络出版：2016-09-08
*通信作者：贺前华 eeqhhe@scut.edu.cn
基金项目：国家自然科学基金(61571192)，广东省公益项目(2015A010103003)，中央高校基本科研业务费项目华南理工大学(2015ZM143)
Foundation Items: The National Natural Science Foundation of China (61571192), The Science and Technology Foundation of Guangdong Province (2015A010103003), The Fundamental Research Funds for the Central Universities, SCUT (2015ZM143)

独立观测值提高检测性能。文献[8]在似然比计算中引入权值，提出加权似然比VAD算法；文献[9]在文献[8]的基础上使用声学模型对权值进行优化，提出基于多声学模型的加权似然比VAD算法，然而有限的权集合缺乏代表性，并且训练得到的权值并未体现分量似然比与语音特征间的联系。近两年，研究者们尝试使用机器学习类方法(例如深度神经网络[10,11]支持向量机[12]等)把似然比及计算似然比过程的相关参数作为特征融合起来提高算法鲁棒性，然而该类算法复杂度较高并且效果受似然比计算准确率影响。文献[13]在2015年提出了基于子带保留似然比的VAD算法，在似然比综合评估时通过保留权值保留具有明显语音特征的频率分量似然比，降低了非语音信号似然比虚高而导致的误检，提高了VAD的检测准确率。

鉴于原保留似然比算法采用单一时域特征设置保留权值，容易丢弃周期性相对不明显的语音信号(尤其是清辅音信号)的子带频率分量似然比，造成局部漏检率增加，本文提出了一种基于子带语音双特征的自适应保留似然比鲁棒语音激活检测算法，在设置保留权值过程中引入对检测清辅音信号更具鲁棒性的归一化子带平均过零率特征，并利用近期过去固定时长内的 VAD 判决结果及对应的子带特征参数自适应估计似然比的保留阈值，从特征和阈值双角度进一步提高了算法在低信噪比下 VAD 的检测准确率。

2 基于子带保留似然比的语音激活检测

在加性噪声环境下，对带噪声语音存在如式(1)两种假设：

$$\left.\begin{aligned} H_0 &: \boldsymbol{X} = \boldsymbol{N} \\ H_1 &: \boldsymbol{X} = \boldsymbol{S} + \boldsymbol{N} \end{aligned}\right\} \tag{1}$$

式中，H_0 与 H_1 分别表示无语音和有语音假设；$\boldsymbol{S}$，$\boldsymbol{N}$ 和 $\boldsymbol{X}$ 分别表示纯净语音、噪声和带噪声语音的傅里叶变换，其第 k 频率分量分别为 S_k，N_k 和 X_k。由于语音和噪声信号的离散傅里叶变换系数满足复高斯分布[14]，因此带噪声语音 X_k 在 H_0 与 H_1 两种情况下的条件概率密度函数分别定义为

$$p\left(X_k \left| H_0\right.\right) = \frac{1}{\pi\lambda_{N,k}} \exp\left\{-\frac{\left|X_k\right|^2}{\lambda_{N,k}}\right\} \tag{2}$$

$$p\left(X_k \left| H_1\right.\right) = \frac{1}{\pi\left[\lambda_{N,k} + \lambda_{S,k}\right]} \exp\left\{-\frac{\left|X_k\right|^2}{\lambda_{N,k} + \lambda_{S,k}}\right\} \tag{3}$$

其中，$\lambda_{N,k}$ 与 $\lambda_{S,k}$ 分别表示噪声和语音在第 k 分量上的方差。根据式(2)和式(3)，使用假设检验中的似然比检验方法设计判决准则，第 k 分量上的似然比 Λ_k 定义为

$$\Lambda_k = \frac{p\left(X_k \left| H_1\right.\right)}{p\left(X_k \left| H_0\right.\right)} = \frac{1}{1+\xi_k} \exp\left(\frac{\gamma_k \xi_k}{1+\xi_k}\right) \tag{4}$$

其中，$\xi_k = \lambda_{S,k} / \lambda_{N,k}$ 为先验信噪比，$\gamma_k = \left|X_k\right|^2 / \lambda_{N,k}$ 为后验信噪比。$\lambda_{N,k}$ 采用MMSE方法[15]进行估计，而先验信噪比 ξ_k 采用基于直接决策(Decision-Directed, DD)的方法[16]估计。

第 k 分量上的对数似然比 $\lg \Lambda_k$ 定义为

$$\ln \Lambda_k = \ln \frac{p\left(X_k \left| H_1\right.\right)}{p\left(X_k \left| H_0\right.\right)} = \ln\left(\frac{1}{1+\xi_k}\right) + \frac{\gamma_k \xi_k}{1+\xi_k} \tag{5}$$

最后，在判决准则中引入保留权值 θ_k，对频率分量似然比进行保留设置，只对保留下来的频率分量似然比进行综合评估，保留似然比VAD判决准则定义为

$$V_{\mathrm{VAD}} = \begin{cases} 0, & \varphi \le \eta_0 \\ 1, & \varphi > \eta_0 \end{cases} \tag{6}$$

其中，$\varphi = \sum_{k=0}^{L_r-1} \theta_k \ln \Lambda_k \Big/ \sum_{k=0}^{L_r-1} \theta_k$，$L_r$ 表示保留的分量似然比数量，保留权值 θ_k 根据其所对应子带 i 的语音特征 $R_{i,\max}$ 的大小进行设置，当 $R_{i,\max} \ge \sigma$ 时，对应第 i 子带的保留权值 θ_i 设置为“1”，其余情况设置为“0”。其中，σ 为保留阈值，$R_{i,\max}$ 表示第 i 子带信号的归一化最大自相关函数值，它是频率分量似然比的保留依据，而子带主要依据梅尔频率尺度进行划分，具体参考文献[13]。

3 基于子带双特征的自适应保留似然比语音激活检测

语音由元音和辅音两种音素组成，辅音根据声带是否振动分为浊辅音和清辅音。发元音和浊辅音时声带振动使信号具有周期性，原保留似然比算法利用归一化最大自相关函数保留了周期性较强的元音与浊辅音信号，而清辅音信号不具有周期性，其分量似然比因子带特征不明显而容易被丢弃，被判为非语音，造成漏检。若要进一步提高该类算法的检测性能，需找到与归一化最大自相关函数在表达清辅音信号上具有互补性质的鲁棒语音特征。根据文献[17]，清音相对浊音在高频部分具有更多能量分布，衰减相对较少，因此清音具有较高的平均过零率，并且平均过零率在背景噪声较大时对识别语音更有效。本文对一段信噪比为10 dB的带非平稳噪声(Babble噪声)语音及其子带归一化最大自相关函数值与子带归一化平均过零率进行分析，发现两类特征在语音前端或末端(虚线区域)对表示语音具有一定的互补关系，如图1所示。

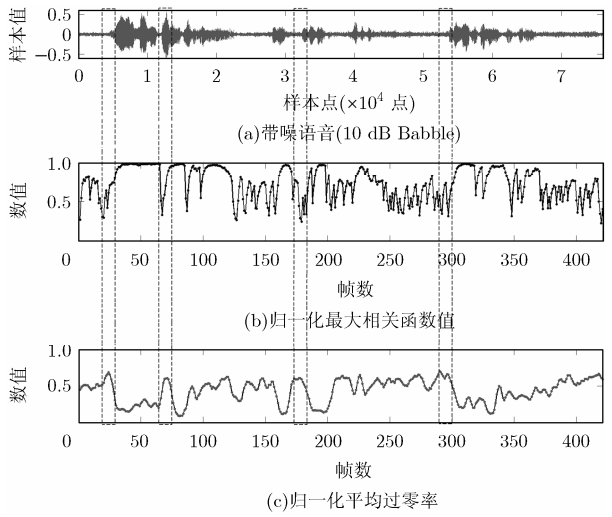

图 1 带噪语音信号的子带归一化自相关函数与子带归一化平均过零率分析

鉴于以上分析，本文采用子带归一化最大自相关函数和子带归一化平均过零率双重特征作为设置保留权值的依据，使具有任一特征的分量似然比都得以保留。考虑到说话人语音特征及背景环境噪声短时内一般不会突变，因此同时利用过去固定时长内的判决结果及相关子带特征参数自适应地估计似然比的保留阈值。所提出方法主要包括双特征设置保留权值、子带归一化平均过零率、自适应估计保留阈值3部分。

3.1 双特征设置保留权值

根据特征强度设置保留权值时，在子带归一化最大自相关函数值的基础上增加子带归一化平均过零率，即采用双重特征进行设置。由于语音在清辅音段的平均过零率高于非语音段，在浊辅音和元音段的平均过零率低于非语音段，因此采用上下双门限对子带平均过零率进行设置，具体定义为

$$\theta_i=\begin{cases}1, & R_{i,\max}\ge\sigma_i \ \text{ 或 } \ \left(Z_i-\alpha_i^u\right)\left(Z_i-\alpha_i^d\right)\ge 0,\\ & \alpha_i^u>\alpha_i^d\\ 0, & \text{其他}\end{cases} \tag{7}$$

其中 $R_{i,\max}$ 为第 i 子带归一化最大自相关函数，σ_i 为 $R_{i,\max}$ 门限下界。Z_i 为第 i 子带的归一化平均过零率，α_i^u 为 Z_i 的门限上界，α_i^d 为 Z_i 的门限下界。

子带归一化最大自相关函数值与子带归一化平均过零率的计算流程如图2所示。

通过带通滤波器滤波后得到子带信号 $s_i, i=1,2,\cdots,5$，利用子带信号计算子带归一化最大自相关函数值[13]和子带归一化平均过零率。

3.2 子带归一化平均过零率

短时过零率表示一帧语音信号时域波形穿过横轴(零电平)的次数，第 i 个子带的平均过零率 ZC_i 通过对子带语音信号计算获得，定义为

$$ZC_i=\frac{1}{2}\sum_{m=0}^{N-1}\left|\operatorname{sgn}\left[s_i(m)\right]-\operatorname{sgn}\left[s_i(m-1)\right]\right| \tag{8}$$

式中，$s_i(m)$ 为第 i 子带信号第 m 个采样点值，ZC_i 为第 i 个子带的短时平均过零率，N 为当前帧采样点数，$\operatorname{sgn}[\bullet]$ 为符号函数。然后，对子带平均过零率进行线性归一化，定义为

$$Z_i=\frac{ZC_i-ZC_{i,\min}}{ZC_{i,\max}-ZC_{i,\min}} \tag{9}$$

其中，$ZC_{i,\max}$ 与 $ZC_{i,\min}$ 分别表示第 i 子带过零率的最大值和最小值，可根据应用场景进行采样设置。

最后，对归一化子带平均过零率进行平滑：

$$Z_i(n)=\delta Z_i(n-1)+(1-\delta)Z_i(n+1) \tag{10}$$

其中，$Z_i(n)$ 表示第 n 帧第 i 子带的归一化平均过零率，δ 表示平滑系数。

3.3 自适应估计保留阈值

鉴于说话人语音特征及背景噪声在短时内是稳定的，且非语音段的归一化最大自相关函数低于元音或浊辅音段，而非语音段的平均过零率一般介于清辅音段与元音或浊辅音段之间。算法中以 Δt 为时间间隔，利用过去 Δt 时长内的 VAD 判决结果和对应的语音归一化特征参数对阈值 $\sigma_i,\alpha_i^u,\alpha_i^d$ 进行定时估计更新。首先，假设已过去最近一段时间 Δt 内包含 K 个 VAD 判决结果均为“0”的非语音段，第 j 个非语音段中包含 M 帧信号，则第 i 子带第 j 个非语音段中第 m 帧的归一化自相关函数值和归一化过零率分别定义为 $R_{V_{\text{VAD}}=0,i}^{j,m}$ 和 $Z_{V_{\text{VAD}}=0,i}^{j,m}$，$j=1,2,\cdots,K$，$m=1,2,\cdots,M$，利用非语音段中各子带特征参数的统计值估计当前分量似然比的保留阈值，即 $\hat{\sigma}_i=\frac{1}{K}\sum_{j=1}^{K}\max\left[R_{V_{\text{VAD}}=0,i}^{j,m}\right]$，$\hat{\alpha}_i^u=\frac{1}{K}\sum_{j=1}^{K}\max\left[Z_{V_{\text{VAD}}=0,i}^{j,m}\right]$，$\hat{\alpha}_i^d=\frac{1}{K}\sum_{j=1}^{K}\min\left[Z_{V_{\text{VAD}}=0,i}^{j,m}\right]$。其中，$\max\left[R_{V_{\text{VAD}}=0,i}^{j,m}\right]$，$\max\left[Z_{V_{\text{VAD}}=0,i}^{j,m}\right]$ 和 $\min\left[Z_{V_{\text{VAD}}=0,i}^{j,m}\right]$ 分别表示第 i 子带第 j 个非语音段中所有 M 帧的归一化自相关函数最大值，归一化平均过零率最大值和最小值。

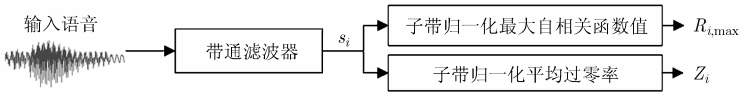

图 2 子带双特征计算流程图

4 实验与结果分析

本文把实验分为仿真实验与现场录音实验两部分。

4.1 仿真实验

4.1.1 实验设置 本文使用汉语普通话自然口语对话语料库(Chinese Annotated Dialogue and Conversation Corpus, CADCC)和NOISEX-92噪声数据库评价各算法的VAD检测性能。设置帧长为45 ms，帧移为22.5 ms。为模拟长时语音低信噪比的检测环境，本文采用如下方法构造仿真环境：选择多人对话样本，总时长约为20 min 37.526 s，共含528个语音段。首先，把样本从16 kHz降采样到8 kHz；然后，对样本进行人工标注，标注语音帧(包含元音和辅音)与非语音帧，其中语音帧约占75.03%，非语音帧约占24.97%。噪声样本包括高斯白噪声(平稳噪声)和Babble噪声(非平稳噪声)；最后，把语音与噪声合成低信噪比样本，信噪比分别为10 dB, 0 dB和−10 dB。平滑系数δ设置为0.5，保留阈值的更新时间间隔设置为18～33 s。将本文方法与文献[2]，文献[4]，文献[5]，文献[6]，文献[13]等方法相比较，通过判决检测参数[18]与接收机操作特性(Receiver Operating Characteristic, ROC)曲线图[19]评估方法的性能。

4.1.2 VAD检测性能 为每种方法选择相对最优的阈值，在这组固定的阈值下比较各种方法的判决检测性能[18]，其中包括检测准确率(CORRECT)、前端漏检率(FEC)、后端漏检率(BEC)、中段漏检率(MSC)、静音段误检率(NDS)、后端误检率(OVER)，具体如表1和表2所示。

从表1和表2可看出，一方面，本文方法在大多数情况下检测准确率(CORRECT)均高于其余方法，其中本文方法与文献[13]原保留似然比方法相比，在10 dB, 0 dB和−10 dB白噪声下分别提高了约1.2%,7.2%和8.1%，在10 dB和0 dB Babble噪声下分别提高了约1.6%和3.4%；另一方面，在两种背景噪声下，本文方法相比文献[13]的方法，在获得接近或更低总误检率(即NDS+OVER)的情况下，FEC, MSC和BEC均有不同程度的降低，从而证明算法的有效性。另外，表2的结果显示，在0 dB和−10 dB的Babble噪声下检测性能相比文献[6]和文献[13]方法略微下降，主要原因是部分语音段信号(例如语音与非语音段的始末过渡信号、较轻微或断续的语气信号)的双特征受噪声影响而变得不明显，部分子带似然比未得到保留而导致漏检率略有上升。

4.1.3 ROC曲线 图3为10 dB, 0 dB和−10 dB 白噪声 ((a), (b), (c)) 和 Babble 噪声 ((d), (e), (f)) 下的ROC曲线图。

从图 3(a)，图 3(b)和图 3(c)可看出，在 10 dB 白噪声下，本文方法当总误检率大于 15%时语音检测率均优于其余方法，即使总误检率小于 15%，本文方法也能获得接近于文献[13]的检测性能。在 0 dB和−10 dB 白噪声下，本文方法在总误检率大于 5%时，语音检测性能明显优于其余方法。从图 3(d)，图 3(e)和图 3(f)可看出，一方面，在 10 dB 的 Babble噪声下，当误检率大于 20%时，本文方法在语音检

表 1 白噪声条件下检测性能对比(%)

信噪比(dB)	方法	CORRECT	FEC	MSC	BEC	NDS	OVER
10	文献[4]	88.07	1.12	9.25	0.05	14.13	2.32
	文献[5]	90.56	1.32	5.13	0.01	10.87	7.53
	文献[6]	88.83	2.15	6.39	0.11	9.75	9.00
	文献[2]	75.61	4.54	21.17	0.39	13.29	5.97
	文献[13]	90.36	2.65	4.15	0	3.11	15.04
	本文方法	91.52	1.65	3.36	0.01	7.68	11.21
0	文献[4]	75.67	1.62	19.86	0.39	28.96	2.72
	文献[5]	79.10	1.94	15.28	0.16	25.34	6.15
	文献[6]	82.98	3.13	10.65	0.12	14.44	11.94
	文献[2]	66.56	3.03	31.22	0.62	25.68	3.45
	文献[13]	77.39	5.33	15.63	0.34	16.38	10.16
	本文方法	84.52	3.04	8.66	0.06	14.42	12.24
−10	文献[4]	56.51	3.02	42.82	0.80	30.56	3.44
	文献[5]	58.76	3.96	39.41	0.58	27.69	5.39
	文献[6]	64.07	8.70	28.42	0.59	17.92	12.65
	文献[2]	53.28	2.06	45.75	0.93	37.23	3.41
	文献[13]	56.54	8.78	36.99	0.61	23.31	11.38
	本文方法	64.72	8.47	28.09	0.32	21.05	9.47

表 2　Babble 噪声条件下检测性能对比(%)

信噪比(dB)	方法	CORRECT	FEC	MSC	BEC	NDS	OVER
10	文献[4]	76.58	0.98	18.40	0.27	31.95	2.78
	文献[5]	81.43	1.15	10.80	0.20	31.85	6.03
	文献[6]	83.07	1.13	7.98	0.22	26.92	12.86
	文献[2]	66.80	2.54	29.26	0.62	31.11	4.42
	文献[13]	83.42	3.00	7.67	0.08	18.70	15.42
	本文方法	85.07	1.91	6.93	0.05	19.20	13.86
0	文献[4]	66.96	1.22	23.23	0.27	46.52	11.54
	文献[5]	67.90	1.88	23.04	0.54	40.93	11.16
	文献[6]	71.27	2.45	18.18	0.29	32.73	19.50
	文献[2]	58.94	1.20	34.72	0.57	49.29	5.48
	文献[13]	67.42	4.53	21.44	0.67	31.10	19.31
	本文方法	70.80	4.50	16.59	0.30	29.97	22.68
−10	文献[4]	55.42	2.71	38.10	0.49	42.19	12.21
	文献[5]	53.99	4.27	39.87	0.71	37.07	12.47
	文献[6]	56.08	8.61	32.17	0.94	29.40	21.15
	文献[2]	51.95	1.26	44.26	0.70	48.53	5.00
	文献[13]	55.83	8.52	32.59	0.94	30.75	19.81
	本文方法	55.60	8.31	33.38	0.30	31.97	19.68

测性能上优于其余方法。另一方面，在 0 dB 的 Babble 噪声下，本文方法优于其余大部分方法，只相比文献[6]方法略微下降；在−10 dB Babble 噪声下，本文方法的优势虽逐渐变得不明显，但依然保持与其余方法相近的水平。

4.2 现场录音实验

4.2.1 实验设置　现场录制带噪语音样本，选择工作单位办公楼大堂(面积约为200~300 m^2)作为录制场地，场地中约有20～30人随意活动和讨论，以场地中人员活动和讨论的声音作为背景噪声，噪声类型近似于Babble噪声，随机录制两人对话语音，样本总时长为8 min25.6 s，共含179个语音段。其中语音帧占79.61%，非语音帧占20.39%。对现场录音样本的测试比较方式与仿真实验相同。

4.2.2 VAD检测性能　根据表3的检测性能可知，本文方法在相近总误检率(NDS+OVER)下的总漏检率(FEC+MSC+BEC)更低，因此检测准确率(CORRECT)均优于其余方法，检测准确率相比其余方法提高了1.11%~17.87%，其中，与文献[13]原保留似然比方法相比提高了1.11%。

图 4 的 ROC 检测曲线显示，在整体性能上，本文方法在总误检率(5%~45%)范围内均优于其余方法，其中略优于文献[13]的方法。

表3 检测性能对比(%)

样本类型	方法	CORRECT	FEC	MSC	BEC	NDS	OVER
现场录制带噪语音	文献[4]	78.83	1.00	19.48	0.07	19.45	4.19
	文献[5]	83.30	1.11	13.55	0.05	18.27	6.22
	文献[6]	81.60	2.10	14.83	0	18.01	6.13
	文献[2]	74.43	2.74	22.20	0.15	21.76	5.70
	文献[13]	91.19	0.87	3.54	0.41	11.37	13.01
	本文方法	92.30	0.51	2.92	0	13.23	11.13

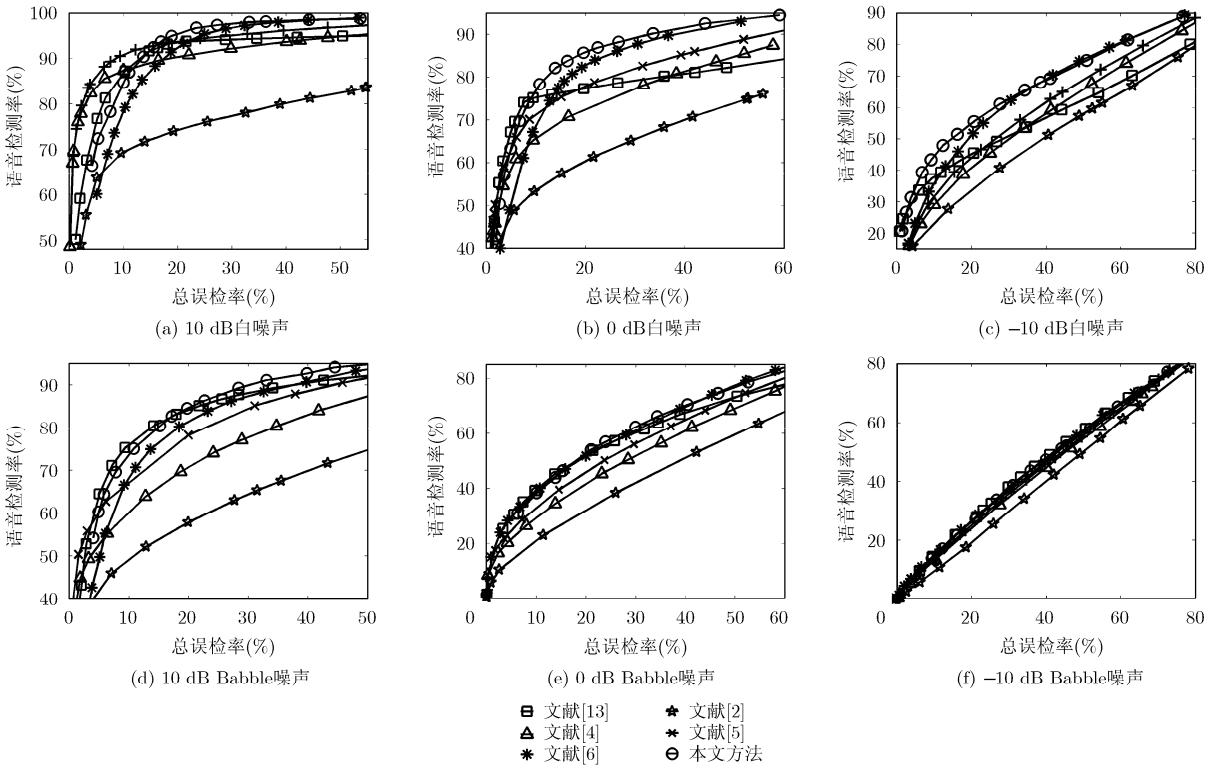

图 3 白噪声和 Babble 噪声条件下的 ROC 曲线(10 dB, 0 dB 和-10 dB)

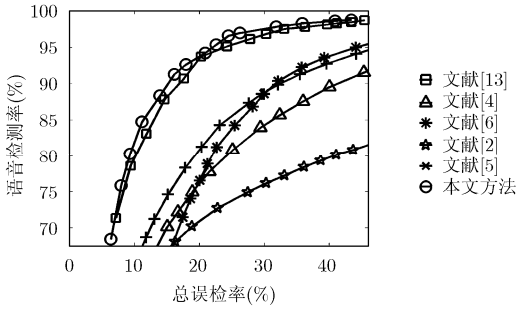

图 4 现场录制带噪语音的检测ROC曲线

另外，实验中发现本文方法在静音段误检率(NDS)有略微增加迹象，主要原因是实验中现场录制带噪语音以他人说话讨论的嘈杂声为背景，若算法检测出静音段噪声的他人讨论声音存在明显双特征，容易造成算法误检。

4.3 声码器性能测试

把本文方法应用于 2.4 kbps 低速率 MELP 声码器，测试声码器的编解码性能，使用感知语音质量评价 (Perceptual Evaluation of Speech Quality, PESQ)作为合成语音的质量评价标准。从表 4 可以看出，经由本文方法对 VAD 判决后再按原清浊模式进行编解码，合成语音质量在多数情况下优于其余方法。相比原 MELP 声码器(文献[2])中的方法，在 10 dB, 0 dB 和-10dB Babble 噪声下，PESQ 值分别提高了 0.159, 0.157 和 0.186；在 10 dB, 0 dB 和 -10 dB 白噪声下，PESQ 值分别提高了 0.153, 0.098 和 0.103；现场录制带噪语音的 PESQ 则提高了 0.142。

表 4 结果显示，本文方法从总体上提升了声码器性能，然而在白噪声环境下，相比文献[6]的方法，声码器性能提升较少甚至略微有所降低，原因是本文方法相比非保留似然比方法对部分样本(尤其是女声样本)的语音中段信号的漏检率略有增加，而此现象主要是由局部语音信号(例如词句始末端信号)的双特征不明显并受强噪声影响而造成的。结合此节与 4.1 节的分析可知，本文方法虽然相对原保留似然比算法保留更多清辅音信号，提高了 VAD 检测准确率，但对于极少数双特征均不突出的语音信号，检测性能仍略显不足。

5 结束语

在此类 VAD 算法中，“似然比”本质上是观测值与假设之间匹配度的一种评估，保留似然比算法的初衷是希望从时域角度把更准确或具明显语音特征的评估值保留下来。本文遵循此思路提出一种基于子带双特征的自适应保留似然比鲁棒语音激活检测算法，利用归一化平均过零率特征保留更多语音中清辅音的频率分量似然比，同时采用已判决结果

及特征信息自适应估计保留阈值，结果表明本文方法可在保持相近误检情况下减少漏检，提高了低信噪比下 VAD 的检测准确率。另外，本文方法被用于低速率声码器系统中，提升了系统在低信噪比环境中的鲁棒性，是基于统计模型的似然比检验方法结合鲁棒时域特征的再一次成功尝试。

表4 各种方法用于2.4 kbps声码器性能比较

信噪比(dB)	PESQ					
	文献[2]	文献[4]	文献[5]	文献[6]	文献[13]	本文方法
10 (Babble)	2.256	2.390	2.369	2.398	2.363	2.415
0 (Babble)	1.971	2.081	2.091	2.109	2.120	2.128
−10 (Babble)	1.843	1.970	2.020	2.017	2.031	2.029
10 (White)	2.540	2.582	2.592	2.700	2.693	2.693
0 (White)	2.355	2.443	2.448	2.449	2.441	2.453
−10 (White)	2.433	2.453	2.488	2.484	2.529	2.536
现场录制带噪语音	2.085	2.165	2.175	2.128	2.214	2.227

参 考 文 献

[1] SREEKUMAR K T, GEORGE K K, ARUNRAJ K, *et al*. Spectral matching based voice activity detector for improved speaker recognition[C]. 2014 International Conference on Power Signals Control and Computations (EPSCICON), Thrissur, 2014: 1–4. doi: 10.1109/EPSCICON.2014.6887507.

[2] DUTA C L, GHEORGHE L, and TAPUS N. Real time implementation of MELP speech compression algorithm using Blackfin processors[C]. 2015 9th International Symposium on Image and Signal Processing and Analysis (ISPA), Zagreb, 2015: 250–255. doi: 10.1109/ISPA.2015.7306067.

[3] CHUL Y I, HYEONTAEK L, and DONGSUK Y. Formant-based robust voice activity detection[J]. *IEEE/ACM Transactions on Audio*, *Speech*, *and Language Processing*, 2015, 23(12): 2238–2245. doi: 10.1109/TASLP.2015.2476762.

[4] JONGSEO S, NAM SOO K, and WONYONG S. A statistical model-based voice activity detection[J]. *IEEE Signal Processing Letters*, 1999, 6(1): 1–3. doi: 10.1109/97.736233.

[5] DUK C Y, AL-NAIMI K, and KONDOZ A. Improved voice activity detection based on a smoothed statistical likelihood ratio[C]. 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Salt Lake City, 2001: 737–740. doi: 10.1109/ICASSP.2001.941020.

[6] RAMIREZ J, SEGURA J, BENITEZ C, *et al*. Statistical voice activity detection using a multiple observation likelihood ratio test[J]. *IEEE Signal Process Letters*, 2005, 12(10): 689–692. doi: 10.1109/LSP.2005.855551.

[7] RAMIREZ J, SEGURA J C, GORRIZ J M, *et al*. Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition[J]. *IEEE Transactions on Audio*, *Speech*, *and Language Processing*, 2007, 15(8): 2177–2189. doi: 10.1109/TASL.2007.903937.

[8] ICK K S, HAING J Q, and HYUK C J. Discriminative weight training for a statistical model-based voice activity detection[J]. *IEEE Signal Processing Letters*, 2008, 15: 170–173. doi: 10.1109/LSP.2007.913595.

[9] YOUNGJOO S and HOIRIN K. Multiple acoustic model-based discriminative likelihood ratio weighting for voice activity detection[J]. *Signal Processing Letters*, 2012, 19(8): 507–510. doi: 10.1109/LSP.2012.2204978.

[10] FERRONI G, BONFIGLI R, PRINCIPI E, *et al*. A deep neural network approach for voice activity detection in multi-room domestic scenarios[C]. 2015 International Joint Conference on Neural Networks (IJCNN), Killarney, 2015: 1–8. doi: 10.1109/IJCNN.2015.7280510.

[11] INYOUNG H and JOON HYUK C. Voice activity detection based on statistical model employing deep neural network[C]. 2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2014: 582–585. doi: 10.1109/IIH-MSP.2014.150.

[12] TAN Yingwei, LIU Wenju, WEI J, *et al*. Hybrid SVM/HMM architectures for statistical model-based voice activity detection[C]. 2014 International Joint Conference on Neural Networks (IJCNN), Beijing, 2014: 2875–2878. doi: 10.1109/IJCNN.2014.6889403.

[13] 何伟俊, 贺前华, 刘杨. 基于子带保留似然比的鲁棒语音激活检测算法[J]. 华中科技大学学报(自然科学版), 2015, 43(11):

78–82. doi: 10.13245/j.hust.151115.

HE Weijun, HE Qianhua, and LIU Yang. Sub-band reserved likelihood ratio-based robust voice activity detection[J]. *Journal of Huazhong University of Science and Technology* (*Natural Science Edition*), 2015, 43(11): 78–82. doi: 10.13245/j.hust.151115.

[14] PEARLMAN W A and GRAY R M. Source coding of the discrete Fourier transform[J]. *IEEE Transactions on Information Theory*, 1978, 24(6): 683–692. doi: 10.1109/TIT.1978.1055950.

[15] GERKMANN T and HENDRIKS R C. Unbiased MMSE-based noise power estimation with low complexity and low tracking delay[J]. *IEEE Transactions on Audio*, *Speech*, *and Language Processing*, 2012, 20(4): 1383–1393. doi: 10.1109/TASL.2011.2180896.

[16] EPHRAIM Y and MALAH D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator[J]. *IEEE Transactions on Acoustics*, *Speech and Signal Processing*, 1984, 32(6): 1109–1121. doi: 10.1109/TASSP.1984.1164453.

[17] 赵力. 语音信号处理[M]. 第 2 版, 北京: 机械工业出版社, 2009: 38–39.

ZHAO Li. Speech Signal Processing[M]. Second edition, Beijing: China Machine Press, 2009: 38–39.

[18] MOUSAZADEH S and COHEN I. Voice activity detection in presence of transient noise using spectral clustering[J]. *IEEE Transactions on Audio*, *Speech*, *and Language Processing*, 2013, 21(6): 1261–1271. doi: 10.1109/TASL.2013.2248717.

[19] PETSATODIS T, BOUKIS C, and TALANTZIS F. Convex combination of multiple statistical models with application to VAD[J]. *IEEE Transactions on Audio*, *Speech*, *and Language Processing*, 2011, 19(8): 2314-2327. doi: 10.1109/TASL.2011.2131131.

何伟俊： 男，1982 年生，博士生，研究方向为语音信号处理与模式识别.

贺前华： 男，1965 年生，教授，博士生导师，研究方向为语音处理、数字音频处理、语音编码、音频事件分析及应用.

吴俊峰： 男，1992 年生，硕士生，研究方向为语音信号处理.

杨继臣： 男，1980 年生，博士，研究方向为多媒体检索.