基于长时信息的自适应话音激活检测

杨绪魁,屈 丹,张文林,闫红刚
(解放军信息工程大学,河南郑州 450001)

摘 要: 语音信号的长时信息应用于话音激活检测中表现优越.利用三种听觉滤波器组,对语音信号进行非线性的谱分解,本文提出了六种基于听觉滤波器组的长时信息,并提出了基于长时信息的自适应话音激活检测算法.该算法无需训练数据,根据多种长时信息,直接在待测信号中挑选出类别明确的信号,然后利用这些信号训练分类模型,对待测信号按帧进行语音-非语音分类.在TIMIT语音库和NOISEX-92噪声库上的实验表明,该算法在极低信噪比环境下,仍表现出更高的准确性和更强的稳健性.同时,在线实验表明,算法在实时处理中仍能取得优异的性能.

关键词: 话音激活检测;长时信息;听觉滤波器;自适应
中图分类号: TN912.34 **文献标识码:** A **文章编号:** 0372-2112 (2018)04-0878-08
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2018.04.016

Adaptive Voice Activity Detection Based on Long-Term Information

YANG Xu-kui, QU Dan, ZHANG Wen-lin, YAN Hong-gang
(*PLA Information Engineering University, Zhengzhou, Henan* 450001, *China*)

Abstract: The long-term information of speech signals shows excellent performances in the applications of voice activity detection. Six types of long-term information based on auditory filter banks are proposed through the non-linear spectral decomposition with three different auditory filters. Further, an adaptive voice activity detection algorithm based on these types of long-term information is proposed. Without additional training data, this algorithm use the data selecting from the test signals according to long-term information to train a speech/non-speech classifier, and classifies the current test signals using the speech/non-speech classifier frame by frame. Experiments on TIMIT dataset and NOISEX-92 dataset show that the algorithm improves the performance of VAD with higher accuracy and stronger robustness in low SNR noisy environments. The online experiments show that it can also obtain a good performance in real-time processing conditions.

Key words: voice activity detection; long-term information; auditory filter bank; adaptive

1 引言

话音激活检测(Voice Activity Detection, VAD)是语音信号处理相关应用必不可少的前端处理技术[1],如语音编码,语音增强,语音识别等.VAD的目的是检测当前音频信号中是否包含语音信号,即以帧为单位对输入信号进行分类,将其中语音信号标注出来.通过VAD可以有效地提升后端相应处理系统的效率和性能.

目前很多研究都关注于噪声环境下的话音激活检测[2~5].其中,基于长时信息的话音激活检测算法[6~9]并没有将各帧信号孤立开,而是充分利用相邻帧之间的相关信息进行检测,因此性能更为优越.但基于线性频谱的长时信息并没有充分利用语音信号的声学特性,因此性能受到一定的限制.

听觉滤波器组,如梅尔频率滤波器[10]、Gammatone滤波器[11]以及音高(Pitch)滤波器[12]等,是根据人耳耳蜗功能设计的滤波器,故听觉滤波器组的输出相当于对语音信号进行了近似对数形式的谱分解.这种非线性的谱分解可以更加显式的表示诸如共振峰之类的重要声学曲线[13].因此,基于听觉滤波器组的长时信息不仅利用了语音信号各帧之间的相关性,而且充分考虑了语音信号的声学特性和人耳听觉机理,因此与基于线性频谱的长时信息相比,其话音激活检测性能更优[9].但是这类算法在分类阶段需要根据不同噪声环境设置合适的阈值,若阈值设置不当将极大影响算法性能,因此不适用于复杂多变的噪声环境.

本文借鉴VQVAD算法[14]的思想,提出基于长时信息的自适应话音激活检测算法,该算法综合利用多种长时信息,直接在待测音频段中分别挑选出类别明确的语

收稿日期:2016-11-16;修回日期:2017-04-24;责任编辑:李勇锋
基金项目:国家自然科学基金(No.61673395, No.61403415);河南省自然科学基金(No.162300410331)

音信号和非语音信号;然后将挑选出的信号作为训练数据,分别训练出能够较好拟合当前音频段语音信号、非语音信号分布的高斯混合模型(Gaussian Mixture Model, GMM);接着利用 GMM 模型,计算各帧信号为语音帧或非语音帧的后验概率;最后根据后验概率进行语音非语音判决. 本文算法采用 GMM 模型进行分类,不仅判决更为准确,而且避免了基于长时信息的 VAD 算法中阈值设置不合理导致性能下降的缺点;同时,GMM 模型的训练数据直接从待测语音中产生,不存在训练环境与测试环境失配的问题. 因此,本文算法具有更高的准确性和更强稳健性. 基于 TIMIT 语音库和 NOISEX-92 噪声库的 VAD 实验进一步验证了本文算法的有效性.

2　基于听觉滤波器组的长时信息

听觉滤波器能够较好地模拟人耳对于声音频率的感知和处理过程,因此在音频处理相关应用中发挥了重要作用. 由于听觉滤波器设计遵循听觉心理学(Psycho-acoustical)的研究成果,因此基于此类滤波器而提出的声学特征也具有更好的区分性和鲁棒性. 本文基于这三种听觉滤波器,提出长时音高散度(long-term pitch divergence, LTPD)、长时梅尔频谱散度(long-term Mel spectral divergence, LTMD)、长时 Gammatone 频谱散度(long-term Gammatone spectral divergence, LTGD)、长时音高变化率(long-term pitch variability, LTPV)、长时梅尔频谱变化率(long-term Mel spectral variability, LTMV)、长时 Gammatone 频谱变化率(long-term Gammatone spectral variability, LTGV)六种长时信息特征参数.

2.1　基于听觉滤波器组的长时谱散度

对于基于音高、梅尔频率和 Gammatone 三种听觉滤波器的长时谱散度 LTPD、LTMD 以及 LTGD,首先定义 R_d 阶长时谱包络如式(1)所示:

$$E_*(k,l)=\max\{X_*(k,l+j)\mid j=-R_d,\cdots,R_d\} \quad (1)$$

其中,$\boldsymbol{X}_*$ 可以为音高特征 $\boldsymbol{X}_P$、梅尔频谱 $\boldsymbol{X}_M$ 或者是 Gammatone 频谱 $\boldsymbol{X}_G$,$X_*(k,l)$ 为其第 l 帧第 k 个频带上的幅值,$E_*(k,l)$ 为与之对应的长时谱包络.

对于噪声谱特征 $\boldsymbol{N}_*$ 可以通过基于 MMSE 的估计器[17]从 $\boldsymbol{X}_*$ 中估计得到. 则第 l 帧第 k 个子带上的平均噪声谱 $\overline{N}_*(k,l)$ 可以用式(2)定义.

$$\overline{N}_*(k,l)=\frac{1}{l}((1-\alpha)(l-1)\overline{N}_*(k,l-1)),l>1 \quad (2)$$

其中,$N_*(k,l)$ 为第 l 帧第 k 个频带上的噪声谱,且 $\overline{N}_*(k,1)=N_*(k,1)$.

根据文献[6]中长时谱散度的定义,则分别利用三种滤波器的谱包络和噪声谱(式(1)和式(2))可以分别得到 LTPD、LTMD 和 LTGD.

2.2　基于听觉滤波器的长时谱变化率

基于听觉滤波器的长时信号变化率首先采用 Bartlett-Welch 方法对 $\boldsymbol{X}_*$ 进行平滑,

$$S_*(k,m)=\frac{1}{2M_v+1}\sum_{i=m-M_v}^{m+M_v}|X_*(k,i)|^2 \quad (3)$$

其中,M_v 为移动窗长.

然后,根据 $S_*(k,m)$ 定义一个熵测度:

$$\xi_*(k,l)=-\sum_{m=l-R_v}^{l+R_v}\bar{S}_*(k,m)\log\bar{S}_*(k,m) \quad (4)$$

其中,$\bar{S}_*(k,m)$ 为 $S_*(k,m)$ 连续 $2R_v+1$ 帧的归一化谱,其定义如式(5)所示.

$$\bar{S}_*(k,l')=\frac{S_*(k,l')}{\sum_{i=l'-R_v}^{l'+R_v}S_*(k,i)} \quad (5)$$

最后,利用这个熵定义基于听觉滤波器的长时信号变化率:

$$V_*(l)=\frac{\sum_k(\xi_*(k,l)-\overline{\xi_*(l)})^2}{K} \quad (6)$$

其中,K 为频带数,$\overline{\xi_*(l)}$ 为第 l 帧熵测度的均值,定义如下:

$$\overline{\xi_*(l)}=\frac{\sum_k\xi_*(k,l)}{K} \quad (7)$$

利用三种滤波器可以得到 LTPV、LTMV 和 LTGV.

2.3　本文算法

上述的三种听觉滤波器都模拟了人类听觉系统的特性,但又各有侧重:音高滤波器模拟的是人耳对应声音频率的距离感知;梅尔频率滤波器依据是人类听觉系统所感知到的声音频率与该声音的物理频率的对应关系;Gammatone 滤波器仿真的是基底膜不同位置对声音的响应过程. 因此,基于这些滤波器提出的长时信息特征参数具有一定的互补性,综合利用它们能够提升 VAD 判决的准确性.

根据文献[6,7,9],基于长时信息的 VAD 算法仅仅通过设定阈值来进行语音-非语言的检测,虽然该阈值可以随着信号环境的变化而调整,但是其依赖于算法对信噪比的估计值,因此限制了算法在复杂信道环境下的应用.

基于分类器的 VAD 算法需要训练数据对语音-非语音分类器进行训练. 当应用环境与训练数据所处环境一致时,算法能够取得较好的性能,但是当应用环境与训练环境失配时,系统的性能会有较大的下降.

为了充分利用各种长时信息的互补性,同时保证 VAD 检测的鲁棒性,本文提出基于长时信息的自适应 VAD 算法. 该算法首先根据长时信息特征挑选出类别明确的帧信号;然后利用这些帧信号训练语音-非语音

模型;最后根据模型对类别不确定的帧进行分类. 该算法采用分类器进行语音-非语音判决,同时分类器的训练数据是直接从待测语音信号中挑选而得,因此算法的准确性、自适应性都将得到极大的提升. 算法的具体流程如下所示:

算法 1 基于长时信息的自适应 VAD 算法

输入:语音信号 $s(n)$,设定帧长和帧移;
1. 对原始信号 $s(n)$ 进行分帧,并提取 MFCC 特征及 GFCC 特征;
2. 计算各帧的长时信息;
3. 根据长时信息,挑选出能够明确给定类别的帧信号;
4. 利用这些标注帧信号的 MFCC 特征、GFCC 特征以及长时信息特征训练语音-非语音模型;
5. 根据训练得到的语音-非语音模型对各帧信号进行分类;
输出:各帧信号的 VAD 标注.

如算法 1 所示,基于长时信息的自适应 VAD 算法首先对音频信号进行分帧处理并提取各帧信号的梅尔频率倒谱系数(MFCC)特征 $\boldsymbol{M}(l)$、Gammatone 频率倒谱系数(GFCC)特征 $\boldsymbol{G}(l)$ 和长时信息特征 $\boldsymbol{L}(l)=[D_S(l),V_S(l),D_p(l)V_P(l),D_M(l),V_M(l),D_G(l),V_G(l)]$,其中 $\boldsymbol{L}(l)$ 的各维分别表示各帧信号的 LTSD、LTSV、LTPD、LTPV、LTMD、LTMV、LTGD 和 LTGV,其中前两个参数为经典长时谱散度和变化率,而后面六个参数分别为三种滤波器谱的谱散度和变化率. 令 $\boldsymbol{F}(l)=[\boldsymbol{M}(l),\boldsymbol{G}(l),\boldsymbol{L}(l)]$ 作为模型训练和测试的特征. 然后根据长时信息挑选出类别明确的帧信号,挑选方法如下:

(1)令 $\boldsymbol{L}_1=[D_S(1),D_S(2),\cdots,D_S(L)]$,依次类推,$\boldsymbol{L}_8=[V_G(1),V_G(2),\cdots,V_G(L)]$,其中 L 为信号的总帧数. 分别对 $\boldsymbol{L}_1,\cdots,\boldsymbol{L}_8$ 按照数值由大到小进行排序,得到各帧信号的排序号 $\boldsymbol{I}_1,\cdots,\boldsymbol{I}_8$,则第 l 帧信号属于语音帧的概率可以表示如下:

$$P(l)=\frac{\sum_{i=1}^{8}I_i(l)/L}{8} \tag{8}$$

其中,$I_i(l)$ 表示第 l 帧在第 i 个排序中的排序号.

(2)对概率 $P(l)$ 按数值大小进行排序,并将概率最大的 10% 对应的帧信号标记为语音帧,将概率最小的 10% 对应的帧信号标记为非语音帧;

挑选出类别明确的帧信号后,并利用这些帧信号特征 $\boldsymbol{F}(l)$ 训练语音-非语音模型;最后根据语音-非语音模型对各帧进行 VAD 检测,得到各帧的 VAD 标注.

采用 GMM 模型 $\lambda^S:(\omega_m^S,\mu_m^S,\Sigma_m^S)$ 和 $\lambda^N:(\omega_m^N,\mu_m^N,\Sigma_m^N)$ 作为语音-非语音模型,其中,λ^S 为语音相关的 GMM 模型,λ^N 为非语音相关的 GMM 模型,$\omega_m^*,\mu_m^*,\Sigma_m^*$ 分别为第 m 个混元的权重,均值和方差,此时 $*$ 分别代表语音和噪声两种情况. 为了简化,设定 λ^S 和 λ^N 的混元数相同且为 M. 则第 l 帧信号在语音-非语音模型上的似然度如式(9)和式(10)所示:

$$p(\boldsymbol{F}(l)\mid\lambda^S)=\sum_{m=1}^{M}\omega_m^S N(\boldsymbol{F}(l)\mid\mu_m^S,\Sigma_m^S) \tag{9}$$

$$p(\boldsymbol{F}(l)\mid\lambda^N)=\sum_{m=1}^{M}\omega_m^N N(\boldsymbol{F}(l)\mid\mu_m^N,\Sigma_m^N) \tag{10}$$

因此,最终可以通过比较对数似然度来进行 VAD 判定,即当 $\log(p(\boldsymbol{F}(l)\mid\lambda^S))\geqslant\log(p(\boldsymbol{F}(l)\mid\lambda^N))$ 时,第 l 帧判定为语音帧,反之判定为非语音帧.

为了实用化,可进一步降低算法的复杂度. 因此,训练语音-非语音模型时,可以用 k 均值(k-means)算法替代期望最大化(Expectation Maximization,EM)算法,同时假设语音-非语音的先验概率相同,则对数似然度判决准则可以简化为如式(11)所示的最近邻准则,

$$\theta_l=\min_m\|\boldsymbol{F}(l)-\mu_m^S\|-\min_n\|\boldsymbol{F}(l)-\mu_m^N\| \tag{11}$$

当 $\theta_l\leqslant 0$ 时,第 l 帧判定为语音帧,反之判定为非语音帧.

3 实验

3.1 实验数据

实验所用语音数据来源于 TIMIT 语料库[18]. 由于 TIMIT 语料库中单句语音过短,因此通过随机选取各子集(即 DR1 到 DR8)中一句语音拼接成一句长语音,并在句首、句末及各短句之间插入 2.5 秒的静音. 对各短句的波形进行规整,使之功率大致相当. 采用简单的基于能量的 VAD 算法生成标注,并进行人工修正. 最终得到的音频数据中,语音信号时长占总时长的 47.06%. 将这些音频数据分为开发集和测试集,各数据集约包含 600 秒音频数据. 开发集用于估计算法的参数,测试集用于性能评测.

实验所用噪声数据来源于 NOISEX-92 噪声库[19],各种噪声的标记及其对应录制环境如表 1 所示.

表 1 NOISEX-92 噪声标号、录制环境对应表

标号	录制环境
Factory1	板材切割及电器设备焊接工厂噪声
Factory2	汽车生产车间噪声
Leopard	军用车辆行驶噪声(速度 70km/h)
M109	M109 坦克内部噪声(速度 30km/h)
Opsroom	驱逐舰作战室背景噪声
Engine	驱逐舰引擎室噪声
F16	F-16 战斗机座舱噪声(航速 500 节)
Buccaneer1	喷气机座舱噪声(航速 190 节)
Buccaneer2	喷气机座舱噪声(航速 450 节)
Babble	100 个人在食堂同时说话
Hfchannel	高频无线电信道噪声(解调后)
Machinegun	.50 口径机枪射击声
Pink	粉红噪声
Volvo	Volvo 340 汽车内部噪声(速度 120km/h)
White	白噪声

3.2 评测指标

话音激活检测是一个典型的二分问题,即输入帧信号只能被标记为语音或者非语音. 对于这类问题,可用的评测指标有很多. 一般而言,评测指标的定义是:首先对分类结果进行统计,得到四种物理量的具体数值,然后根据这些物理量数值之间的比例进行定义. 四种物理量分别为:真阳、真阴、假阳和假阴. 基于这四种物理量定义四种评测指标:语音命中率(即召回率)TPR、非语音命中率 TNR、虚警率 FPR 和准确率 ACC. 综合考虑召回率和虚警率,还可以定义接收机工作特性(Receiver Operating Characteristic,ROC)曲线,用于权衡两者之间的矛盾. 进一步定义曲线下面积(Area Under Curve,AUC)用于综合衡量算法在各种阈值下的性能. 本文采用 TPR、TNR、ACC 以及 AUC 值对算法性能进行评价.

3.3 实验设置

TIMIT 中语音信号采样率为 16kHz,NOISEX-92 中噪声的采样率为 19.98kHz. 因此,对 NOISEX-92 中噪声降采样至 16kHz 后,采用 FaNT(Filtering and Noise Adding Tool)[①]工具生成指定信噪比的带噪语音信号,信噪比分别为 -10dB、-5dB、0dB、5dB、10dB、15dB、20dB. 采用 Chroma 工具箱[②]提取音高特征. 采用 CAaS(Cochleagram Analysis and Synthesis)工具箱[③]提取 Gammatone 频谱和 GFCC 特征. 采用 Voicebox 工具箱[④]提取梅尔频谱和 MFCC 特征. 基于 MMSE 的噪声估计器通过 Voicebox 工具箱中 estnoiseg 函数实现.

语音信号帧长为 30ms,帧移为 10ms. 在基于长时信息的 VAD 实验方法同文献[9](即算法标记为 LTSD、LTSV、LTPD 等).

3.4 实验结果

3.4.1 参数 R_d 对长时谱散度影响

为了研究参数 R_d 对长时谱散度的影响,令其以步长为 1,由 1 变化至 10,并在不同信噪比、不同噪声的环境下进行 VAD 检测实验. 表 2 列举了不同信噪比、噪声类型为 F16 和 Pink 时,使得基于四种长时谱散度的 VAD 算法性能最优(即准确率最大)的 R_d 的数值. 从表中可以看出,噪声类型不同时,参数 R_d 的最优值存在差异;信噪比不同时,参数 R_d 的最优值也不相同. 但是为不同噪声环境设置差异化的参数并不现实,因此应当综合考虑不同噪声环境下的系统性能,对参数 R_d 进行设置. 图 1 展示了参数 R_d 对基于长时谱散度的 VAD 算法在不同信噪比和噪声类型的条件下的平均准确率的影响. 由图 1 可知,LTPD、LTMD 和 LTGD 的性能优于 LTSD. 这是因为基于听觉滤波器组的长时信息能够更好地对语音信号进行谱分解,进而得到更具区分性的长时特征. 另外,当 $R_d = 6$ 时,四种基于长时谱散度的

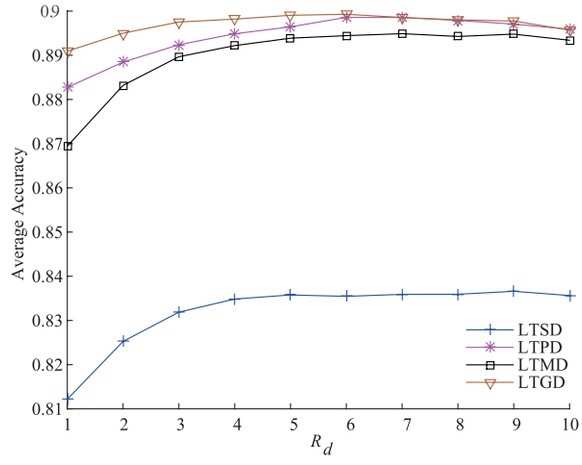

图1　参数 R_d 对长时谱散度影响
(综合考虑不同噪声类型、不同信噪比)

VAD 算法的平均准确率都达到了最大值,因此在接下来的实验中,令参数 $R_d = 6$.

表 2　信噪比不同、噪声不同的情况下参数 R_d 的最优值

SNR	LTSD		LTPD		LTMD		LTGD	
	F16	Pink	F16	Pink	F16	Pink	F16	Pink
-10dB	7	9	7	7	7	7	7	10
-5dB	8	9	8	9	8	8	8	8
0dB	5	8	7	4	8	4	5	5
5dB	3	3	3	1	6	2	3	4
10dB	1	1	2	1	2	1	4	2
15dB	1	1	2	1	1	1	1	1
20dB	1	1	1	1	1	1	1	1

3.4.2 参数 R_v 及 M_v 对长时谱变化率影响

为了研究 R_v 和 M_v 对长时谱变化率的影响,分别令 $R_v = 5,10,20,30,40,50$,$M_v = 1,5,10,20,30$ 进行实验. 与参数 R_d 类似,不同的信噪比、不同的噪声类型对 R_v 和 M_v 的选择都有影响,因此我们还是将算法在不同信噪比和噪声类型的条件下的平均准确率作为参数选择的依据. 图 2 展示了参数 R_v 和 M_v 的不同组合对算法平均正确率的影响. 可以看出,当参数选取不当时,算法的性能会有较为明显的下降,且当 $M_v = 10$,$R_v = 30$ 时,各算法都能取得较为不错的性能,因此在后续实验中令其为默认设置.

3.4.3 阈值对基于长时信息的 VAD 算法的影响

为了验证阈值选取对于 VAD 算法的影响,本文进

① http://dnt.kr.hsnr.de/download.html
② http://resources.mpi-inf.mpg.de/MIR/chromatoolbox/
③ http://web.cse.ohio-state.edu/pnl/software.html
④ http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

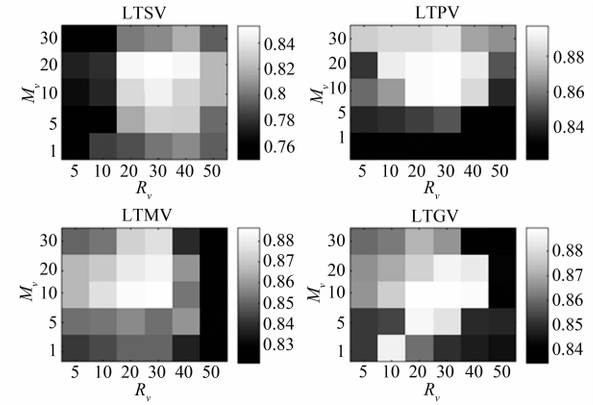

图2 参数R_d和M_v对长时谱散度的影响

行了如下实验. 遍历所有可行的阈值,并认为使得算法准确率最高的阈值为最优阈值. 将最优阈值得到的结果与阈值选取算法估计得到的结果进行比较,如表3 和表4 所示. 表3 和表4 分别展示了信噪比为 -5dB 时和信噪比为5dB 时,阈值选取对算法性能的影响. 从表中可以看出,估计得到的阈值与最优阈值之间都存在偏差,进而导致算法性能下降. 特别是在信噪比较低的情况下,性能下降更为明显.

表3 SNR = -5dB 时阈值选取对算法性能的影响

算法	阈值		ACC	
	最优	估值	最优	估值
LTSD	7.00	9.00	0.7231	0.7142
LTSV	-2.08	-1.88	0.8181	0.8086
LTPD	-5.00	-7.00	0.8445	0.8385
LTPV	-0.25	-0.40	0.8405	0.8270
LTMD	5.00	6.00	0.8426	0.8319
LTMV	-1.90	-1.70	0.8392	0.8344
LTGD	-1.00	-1.50	0.8462	0.8461
LTGV	-0.10	0.05	0.8362	0.8313

表4 SNR =5dB 时阈值选取对算法性能的影响

算法	阈值		ACC	
	最优	估值	最优	估值
LTSD	10.00	11.00	0.8894	0.8825
LTSV	-1.38	-1.48	0.8904	0.8895
LTPD	-9.00	-11.00	0.9366	0.9354
LTPV	0.15	0.05	0.8966	0.8920
LTMD	7.00	8	0.9205	0.9184
LTMV	-1.00	-1.10	0.9015	0.8990
LTGD	-4.00	-4.50	0.9341	0.9340
LTGV	0.50	0.55	0.9012	0.9003

3.4.4 自适应算法性能

由上述实验可知,参数或者阈值的不同选取对基于长时信息的 VAD 算法都有较大的影响,因此研究自适应的基于长时信息的 VAD 算法具有现实意义. 为了验证基于长时信息的自适应 VAD 算法的优越性,我们将其(Adapt)与十种算法进行比较,其中包括两种公认的性能较为优越的 VAD 算法(Sohn[20] 和 Harm[21]),和八种基于长时信息的 VAD 算法. 这八种长时信息 VAD 算法中,两种为经典长时信息 VAD 算法[6,9],六种为本文提出的长时特征构造的 VAD 算法. 表5 列举了这些算法在 -10dB 信噪比时,对受到不同噪声污染的语音的检测性能,其中加粗字体为多种方法中的最佳性能. 从表中数据可以看出,基于长时信息的 VAD 算法明显优于 Sohn 和 Harm 算法;尽管信噪比极低,自适应算法仍能取得不错的性能,且对绝大多数的噪声类型,其性能明显优于其他 VAD 算法. 对于 Babble 噪声,自适应算法与基于 LTSD 的 VAD 算法性能之间存在一定差距,这是由于该噪声内容为纯人声,自适应算法由于采用倒谱特征进行分类,并不能够很好的把语音与噪声区分开. 对于 Factory1 和 Machinegun 噪声,自适应算法的性能不及基于 LTPD 或 LTPV 的 VAD 算法,这是由于它们都是由包含了两种不同的噪声:对于 Factory1 噪声,其包含了相对平稳的机器轰鸣声和随机突发的金属撞击声或钢材切割声;对于 Machinegun 噪声,其包含了随机突发的机枪射击声和间隔的静音. 在信噪比较低的情况下,这样的突发噪声对算法检测性能有较大的影响,而基于音高的长时信息在这种情况下更具优势.

表6 给出了各 VAD 算法在不同信噪比下的语音命中率(TPR)、非语音命中率(TNR)、准确率(ACC)及其平均值. 从表中可以看出,在信噪比较低的情况下(10dB 以下),根据各种衡量指标进行评价,自适应算法的性能都明显优于其他算法. 信噪比较高时,自适应算法的 TNR 和 ACC 虽然不是最大值,但是与最大值十分接近. 而从各指标的平均值来看,自适应算法的性能的优势十分明显,对不同的信噪比都具有较好的适应性.

从表6 中可以看出,各种长时信息的性能随着信噪比的降低而降低,但是在相同信噪比下,长时信息的性能优于 Sohn 和 Harm 算法,基于听觉滤波器的长时信息的性能优于基于线性频谱的长时信息,综合利用多种长时信息的自适应算法性能最优. 因此,可以说明长时信息之间存在一定的互补性,且鲁棒性更强.

3.4.5 在线检测仿真实验

为了验证算法的在实时系统中的应用,我们进一步对在线话音激活检测进行了仿真实验. 对于各种噪声环境,从开发集中分别挑选 30 秒的语音数据和非语音数据,用于训练语音-非语音模型. 利用该语音-非语

音模型对测试集中数据进行话音激活检测. 实验结果如表 7 所示. 从实验结果中可以看出,根据 ACC 指标,在线检测相比于离线检测性能有所下降,但是仍然优于其他话音激活检测算法. 同时由于训练语音-非语音模型的数据与测试数据之间仍有一定失配,因此当判决阈值设置为 0 时,TPR 和 TNR 存在较大偏差,即 TPR 较大,而 TNR 较小.

表 5　信噪比为 −10dB 时各 VAD 算法的 AUC 值

Noise	Sohn	Harm	LTSD	LTSV	LTPD	LTPV	LTMD	LTMV	LTGD	LTGV	Adapt
Babble	0. 6082	0. 6080	0. 7112	0. 6318	0. 6134	0. 6060	0. 6024	0. 6364	0. 5792	0. 6372	0. 6418
Buccaneer1	0. 5291	0. 5280	0. 7038	0. 8576	0. 8241	0. 8732	0. 8214	0. 7699	0. 8672	0. 8752	0. 8819
Buccaneer2	0. 6036	0. 6034	0. 7177	0. 9138	0. 9160	0. 9234	0. 8790	0. 8473	0. 8810	0. 9257	0. 9320
Engine	0. 6915	0. 6916	0. 8392	0. 8954	0. 8075	0. 8746	0. 8805	0. 7622	0. 8069	0. 8447	0. 8975
Opsroom	0. 5569	0. 5575	0. 6439	0. 7832	0. 7576	0. 7787	0. 7436	0. 7592	0. 7486	0. 7751	0. 7868
F16	0. 6763	0. 6763	0. 7652	0. 8812	0. 8304	0. 8712	0. 8214	0. 7634	0. 8200	0. 8798	0. 8873
Factory1	0. 5446	0. 5442	0. 5456	0. 6394	0. 7089	0. 7353	0. 5950	0. 6765	0. 6477	0. 6798	0. 6631
Factory2	0. 7570	0. 7567	0. 7872	0. 9061	0. 8415	0. 8962	0. 8108	0. 8685	0. 8471	0. 9047	0. 9066
Hfchannel	0. 6919	0. 6918	0. 8427	0. 8957	0. 8247	0. 8484	0. 8573	0. 7858	0. 8497	0. 9143	0. 9254
M109	0. 7571	0. 7553	0. 8604	0. 9644	0. 9318	0. 9528	0. 9339	0. 9279	0. 9084	0. 9586	0. 9704
Leopard	0. 9433	0. 9429	0. 9128	0. 9663	0. 9453	0. 9700	0. 9774	0. 9729	0. 9164	0. 9665	0. 9836
Machinegun	0. 7457	0. 7311	0. 6742	0. 6177	0. 8455	0. 5953	0. 6226	0. 4952	0. 6577	0. 5999	0. 6927
Pink	0. 6058	0. 6057	0. 6500	0. 8962	0. 8883	0. 8847	0. 8473	0. 7737	0. 9257	0. 9151	0. 9439
Volvo	0. 9804	0. 9803	0. 9114	0. 9397	0. 9798	0. 9538	0. 9835	0. 9624	0. 9776	0. 9382	0. 9912
White	0. 6454	0. 6453	0. 6841	0. 9156	0. 9243	0. 9382	0. 9028	0. 8925	0. 9439	0. 9589	0. 9641
average	0. 6891	0. 6879	0. 7499	0. 8496	0. 8426	0. 8468	0. 8186	0. 7929	0. 8251	0. 8542	0. 8712

表 6　SNR 从 −10dB 变化到 20dB 时各算法 TPR、TNR、ACC 及其平均值

指标	SNR	Sohn	Harm	LTSD	LTSV	LTPD	LTPV	LTMD	LTMV	LTGD	LTGV	Adapt
TPR	−10	0. 6391	0. 6382	0. 5686	0. 7348	0. 7760	0. 7604	0. 8488	0. 7707	0. 7714	0. 7651	0. 8452
	−5	0. 7186	0. 7178	0. 7623	0. 8192	0. 8259	0. 8453	0. 8394	0. 8131	0. 8398	0. 8296	0. 8911
	0	0. 7656	0. 7649	0. 8425	0. 8682	0. 9143	0. 8759	0. 8843	0. 8679	0. 9039	0. 8670	0. 9258
	5	0. 8368	0. 8668	0. 8822	0. 8804	0. 9378	0. 8977	0. 8997	0. 8831	0. 9283	0. 8713	0. 9479
	10	0. 8906	0. 8863	0. 9061	0. 8920	0. 9476	0. 9074	0. 9229	0. 9093	0. 9430	0. 8961	0. 9366
	15	0. 9198	0. 9209	0. 9338	0. 9019	0. 9469	0. 9180	0. 9388	0. 9268	0. 9607	0. 9257	0. 9582
	20	0. 9362	0. 9373	0. 9566	0. 9260	0. 9554	0. 9383	0. 9496	0. 9428	0. 9610	0. 9363	0. 9737
	Ave	0. 8153	0. 8189	0. 8360	0. 8604	0. 9006	0. 8776	0. 8976	0. 8734	0. 9012	0. 8702	0. 9255
TNR	−10	0. 6336	0. 6323	0. 6644	0. 7556	0. 7187	0. 7748	0. 7177	0. 7706	0. 7372	0. 7533	0. 7846
	−5	0. 6855	0. 6838	0. 6804	0. 7992	0. 8640	0. 8278	0. 8315	0. 8302	0. 8534	0. 8328	0. 8576
	0	0. 7842	0. 7842	0. 8001	0. 8485	0. 8867	0. 8656	0. 8846	0. 8700	0. 9054	0. 8777	0. 9115
	5	0. 8433	0. 8027	0. 8830	0. 8793	0. 9337	0. 8782	0. 9176	0. 8849	0. 9252	0. 8853	0. 9399
	10	0. 8870	0. 8911	0. 9049	0. 9150	0. 9498	0. 9398	0. 9233	0. 9291	0. 9452	0. 9250	0. 9565
	15	0. 9174	0. 9157	0. 9393	0. 9349	0. 9540	0. 9478	0. 9295	0. 9488	0. 9594	0. 9434	0. 9557
	20	0. 9342	0. 9326	0. 9461	0. 9498	0. 9516	0. 9522	0. 9556	0. 9503	0. 9598	0. 9480	0. 9507
	Ave	0. 8122	0. 8060	0. 8312	0. 8689	0. 8941	0. 8837	0. 8800	0. 8834	0. 8980	0. 8808	0. 9081
ACC	−10	0. 6362	0. 6350	0. 6193	0. 7458	0. 7457	0. 7680	0. 7794	0. 7706	0. 7533	0. 7589	0. 8132
	−5	0. 7011	0. 6998	0. 7189	0. 8086	0. 8461	0. 8361	0. 8352	0. 8221	0. 8470	0. 8313	0. 8733
	0	0. 7754	0. 7751	0. 8201	0. 8578	0. 8997	0. 8704	0. 8845	0. 8691	0. 9047	0. 8727	0. 9182
	5	0. 8403	0. 8328	0. 8826	0. 8798	0. 9356	0. 8874	0. 9091	0. 8841	0. 9267	0. 8787	0. 9437
	10	0. 8887	0. 8888	0. 9055	0. 9036	0. 9488	0. 9081	0. 9231	0. 9092	0. 9442	0. 9055	0. 9471
	15	0. 9185	0. 9182	0. 9367	0. 9382	0. 9507	0. 9232	0. 9339	0. 9332	0. 9600	0. 9951	0. 9569
	20	0. 9351	0. 9348	0. 9511	0. 9445	0. 9534	0. 9492	0. 9528	0. 9468	0. 9604	0. 9466	0. 9615
	Ave	0. 8136	0. 8121	0. 8335	0. 8683	0. 8971	0. 8775	0. 8883	0. 8764	0. 8995	0. 8841	0. 9163

表 7 在线检测实验结果

SNR	TPR	TNR	ACC
-10dB	0.9345	0.5180	0.7140
-5dB	0.9456	0.7708	0.8531
0dB	0.9623	0.8889	0.9234
5dB	0.9767	0.9287	0.9513
10dB	0.9824	0.9314	0.9565
15dB	0.9893	0.9369	0.9631
20dB	0.9920	0.9373	0.9646
Ave	0.9693	0.8443	0.9068

4 结束语

本文利用三种听觉滤波器组对语音信号进行非线性的谱分解,进而结合长时谱散度和长时谱变化率,提出了 6 种基于听觉滤波器组的长时信息特征. 进一步根据这些长时特征,提出了自适应 VAD 算法. 实验表明,本文算法在各种噪声环境下都具有更高的准确性和更强的稳健性,尤其当信噪比较低时,本文算法的性能优势更为明显. 但是当信噪比极低的情况下,算法对受非平稳噪声干扰的语音进行检测时,性能下降明显. 因此,如何在极低信噪比环境下进行话音激活检测是下一步研究的重点.

同时对于在线检测,算法需要事先准备一定数量的满足当前信道环境特性的语音和非语音数据,算法的应用受到限制. 因此,下一步的工作在于如何提升算法在线检测的自适应性能.

参考文献

[1] RAMIREZ J, GORRIZ J-M, SEGURA J-C. Voice activity detection, fundamentals and speech recognition system robustness, robust speech recognition and understanding [OL]. https://www. intechopen. com/books/ robust_ speech_ ecognition_ and_ understanding/voice_ activ ety_ detection_ fundamentals_ and_ speech_ recognition_ sys tem_ robustness,2016-11-16.

[2] WISDOM S, OKOPAL G, ATLAS L, PITTON J. Voice activity detection using subband noncircularity[A]. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)[C]. Brisbane, Australia, 2015. 4505 - 4509.

[3] HEESE F, NIERMANN M, VARY P. Speech-codebook based soft voice activity detection[A]. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)[C]. Brisbane, Australia, 2015. 4335 - 4339.

[4] TAO F-J, HANSEN H-L, BUSSO C. An unsupervised visual-only voice activity detection approach using temporal orofacial features[A]. Proceedings of 16th Annual Conference of the International Speech Communication Association (INTERSPEECH)[C]. Dresden, Germany, 2015. 2302 - 2306.

[5] ZHAN G, HUANG Z-Q, et al. Spectrographic speech mask estimation using the time-frequency correlation of speech presence[A]. Proceedings of 16th Annual Conference of the International Speech Communication Association (INTERSPEECH)[C]. Dresden, Germany, 2015. 2287 - 2291.

[6] RAMIREZ J, SEGURA J-C, BENITEZ C, et al. Efficient voice activity detection algorithms using long-term speech information[J]. Speech Communication, 2004, 42(3): 271 - 287.

[7] GHOSH P-K, TSIARTAS A, NARAYANAN S. Robust voice activity detection using long-term signal variability [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(3): 600 - 613.

[8] MA Y, NISHIHARA A. Efficient voice activity detection algorithm using long-term spectral flatness measure[J]. EURASIP Journal on Audio, Speech and Music Processing, 2013: 87, DOI: 10. 1186/1687-4722 - 2013-21.

[9] YANG X-K, He L, QU D, ZHANG W-Q. Voice activity detection algorithm based on long-term pitch information [J]. EURASIP Journal on Audio, Speech, and Music Processing, 2016: 14, DOI: 10. 1186/s13636-016 - 0092-y.

[10] DAVIS S, MERMELSTEIN P. Comparison of parametric representations for monosyllabic word recognitions in continuously spoken sentences[J]. IEEE Transactions on Acoustics, Speech and Signal Processing, 1980, 28(4): 357 - 366.

[11] SCHLUTER R, BEZRUKOV I, WAGNER H, NEY H. Gammatone features and feature combination for large vocabulary speech recognition[A]. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)[C]. Hawaii, USA: IEEE, 2007. 649 - 652.

[12] MEINARD Muller. Information Retrieval for Music and Motion[M]. Berlin: Springer Verlag, 2007. 51 - 55.

[13] SEGBROECH M-V, TSIARTAS A, NARAYANAN S-S. A robust frontend for VAD: exploiting contextual, discriminative and spectral cues of human voice[A]. Proceedings of 14th Annual Conference of the International Speech Communication Association (INTERSPEECH)[C]. Lyon, France, 2013. 704 - 708.

[14] KINNUNEN T, RAJAN P. A practical, self-adaptive voice activity detector for speaker verification with noise telephone and microphone data[A]. Proceedings of IEEE In-

ternational Conference on Acoustics, Speech, and Signal Processing (ICASSP)[C]. Vancouver, Canada, 2013. 7229 – 7233.

[15] Georgiou T-T. Distances Between Power Spectral Densities [R]. Technique Report, arXiv:math/0607026v2, 2006.

[16] JOHANNESMA P-I-M. The pre-response stimulus ensemble of neurons in the cochlear nucleus[A]. Proceedings of IPO Symposium on Hearing Theory [C]. Eindhoven, Netherlands, 1972. 58 – 69.

[17] GERKMANN T, RICHARD C-H. 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.

[18] GAROFOLO J-S, LAMEL L-F, Fisher W-M, et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus[R]. NIST Interagency/Internal Report (NISTIR)-4930, 1993.

[19] VARGA A, STEENEKEN H-J-M. Assessment for automatic speech recognition: Ii. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems[J]. Speech Communication, 1993, 12(3): 247 – 251.

[20] SOHN J, KIM N-S, SUNG W. A statistical model-based voice activity detection[J]. IEEE Signal Processing Letters, 1999, 6(1): 1 – 3.

[21] TAN L-N, BORGSTROM B-J, ALWAN A. Voice activity detection using harmonic frequency components in likelihood ratio Test[A]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)[C]. Dallas, USA: IEEE, 2010. 4466 – 4469.

作者简介

杨绪魁　男,1988 年 10 月出生,福建光泽人. 现为解放军信息工程大学在读博士研究生,研究方向为语音信号处理与识别、机器学习等. E-mail:gzyangxk@163.com

屈　丹　女,1974 年 9 月出生,吉林九台人,现为解放军信息工程大学信息系统工程副教授、博士生导师. 主要研究方向为语音信号处理与识别、人工智能等.

张文林　男,1982 年 11 月出生,河北蔚春人. 现为解放军信息工程大学信息系统工程学院讲师. 主要研究方向语音信号处理与识别、人工智能等.

闫红刚　男,1975 年 10 月出生,河南驻马店人. 现为解放军信息工程大学信息系统工程学院副教授. 主要研究方向通信信号分析、语音处理与识别、机器学习等.