doi:10.14132/j.cnki.1673-5439.2021.01.004

噪声环境下应用于语音标注的端点检测算法研究

俞景彦,赵晓群

(同济大学 电子与信息工程学院,上海 201804)

摘要:端点检测是语音标注的重要前序技术,针对语音标注,设计了一种基于信噪比分类的自适应端点检测算法。该算法首先对标注语音的信噪比分布范围进行分析,将信噪比分类,在每类信噪比范围内选择对应较优的算法。在高信噪比范围选择子带谱熵法,在中等信噪比范围内选择均匀子带频带方差法,而在低信噪比环境下先对带噪语音进行谱减法去噪处理,再采用基于均匀子带频带方差的端点检测算法。仿真实验表明,对语音标注采用的音频信号进行端点检测,该算法能达到较高的检测正确率,证明了算法的有效性。

关键词:端点检测;信噪比分类;子带谱熵;子带频带方差;语音标注

中图分类号:TN912.3 **文献标志码**:A **文章编号**:1673-5439(2021)01-0025-09

Endpoint detection algorithm for speech annotation in noisy environment

YU Jingyan, ZHAO Xiaoqun

(College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China)

Abstract: The endpoint detection is an important preamble technology for speech annotation. For speech annotation, an adaptive endpoint detection algorithm based on signal-to-noise ratio (SNR) classification is designed. Firstly, the distribution range of the labeled speech SNR is analyzed and the SNR is classified by the algorithm. Then, the corresponding better algorithm is selected in the range of each type of the SNR. In the high signal-to-noise ratio range, the sub-band spectral entropy method is selected; in the medium signal-to-noise ratio range, the uniform sub-band frequency band variance method is selected. In a low signal-to-noise ratio environment, the noisy speech is firstly subjected to spectral subtraction denoising processing, and then an endpoint detection algorithm based on uniform sub-band frequency band variance is used. Simulation experiments show that the endpoint detection algorithm for audio signals in the speech annotation can achieve a high detection accuracy rate, thus proving the effectiveness of the algorithm.

Keywords: endpoint detection; signal-to-noise ratio (SNR) classification; sub-band spectral entropy; sub-band frequency band variance; speech annotation

随着大数据时代的到来,数字语音处理、语音识别等技术在不断发展的同时,对语音语料数据的需求也越来越大。语音标注通过一定的形式对原始语音及其对应的发音文本进行配对,能够为语音识别等技术提供前端语料数据[1]。在语音标注中,端点检测技术起着至关重要的作用,对长语音语料进行处理时,需要提取有声片段并分割成短语或词汇,进而标注对应文本信息,生成训练数据。对语音端点检测技术的研究不仅能够提高语音标注的准确率,还能在加快语音识别速度的同时,一定程度上保障

收稿日期:2020-05-07;**修回日期**:2020-08-29 **本刊网址**:http://nyzr.njupt.edu.cn

作者简介:俞景彦,女,硕士研究生;赵晓群(通信作者),男,博士,教授,zhao_xiaoqun@tongji.edu.cn

引用本文:俞景彦,赵晓群.噪声环境下应用于语音标注的端点检测算法研究[J].南京邮电大学学报(自然科学版),2021,41(1):25-33.

识别的准确率。

从包含语音的一段信号中检测出语音的起始点和终止点,准确分离有效语音信号和无效噪声信号的技术被称为语音端点检测(Voice Activity Detection,VAD)。语音端点检测方法大致可分成无监督方法和有监督方法两大类。无监督的语音端点检测方法大多是基于阈值的,此类方法中最经典的是采用短时能量和过零率作为特征参数的双门限端点检测法。常用特征参数还包括频谱熵[2]、频带方差[3]、谱距离[4]和频谱平坦度[5]等。基于统计模型的方法常利用统计似然比检验辅助端点检测[6]。无监督的端点检测技术算法简单、计算快速,能够满足实时系统的要求;但存在鲁棒性差、抗噪声性能差等弊端,当信噪比急速下降时,简单的特征参数不再有效,端点检测效果变差。有监督语音端点检测算法中基于模式匹配的方法常采用支持向量机[7]、高斯混合模型[8]和隐马尔可夫法[9]等。随着人工智能的飞速发展,包括深度置信网络[10]、卷积神经网络[11]、循环神经网络[12-13]在内的多种深度学习框架都应用于语音端点检测技术中。有监督的端点检测技术需要获取与实际应用环境相匹配的特定训练数据,随着数据的逐渐完善,检测准确率会不断提高;但算法复杂度高,且需要大量的训练数据。

语音标注的主要目的是生产语音语料数据,且标注语音一般质量较好,综合考虑语音标注的特点和算法的复杂程度,采用无监督的端点检测算法。本文通过分析工程条件下语音标注中语音信号的信噪比范围和分布,设计了一种噪声环境下基于信噪比分类的自适应端点检测方法。实验结果表明,对于标注场景下的语音片段,在不提高算法复杂度的前提下,本文算法能达到较高的端点检测正确率。

1 语音信号的信噪比

1.1 信噪比范围与分类

现实生活中,存在各种各样的噪声,人们处于被噪声包围的环境内,发出的语音是带噪语音。带噪语音可以认为是纯净语音和噪声混叠形成的。信噪比(Signal to Noise Ratio,SNR)是信号的平均功率与噪声的平均功率之比,它反映了带噪语音信号在时域中呈现的总体特点,被定义为

$$\mathrm{SNR} = 10\lg \frac{\sum_{n=0}^{N-1} s^2(n)}{\sum_{n=0}^{N-1} d^2(n)} \tag{1}$$

其中,$\sum_{n=0}^{N-1} s^2(n)$ 代表信号的能量;$\sum_{n=0}^{N-1} d^2(n)$ 代表噪声的能量。

在实际应用中,存在许多典型的噪声环境,嘈杂语噪声和音乐噪声通常选择 20 dB 的信噪比[14],车载噪声的信噪比一般为 5~10 dB[15]。不同场所环境下的噪声强度也不同,教室、医院、住所和百货店内的噪声强度是最低的,实际信噪比范围大约为 5~15 dB;而火车和飞机上的噪声强度很高,信噪比在 0 dB 左右[16]。此外,另一种特殊的噪声是谈话者干扰,此类噪声可以通过数字混叠的方式模拟,信噪比大约为 10 dB[14]。低信噪比环境下的语音端点检测算法研究,通常将纯净语音信号与噪声信号叠加,产生一定信噪比的带噪语音,此时语音信噪比已知,语音信号的质量可控。而现实环境下的语音信号由于信噪比未知,语音质量不可控,对于质量过差的语音会选择弃用或采用语音增强技术进行预处理。语音增强算法通常在 0~15 dB 信噪比的环境下工作[16],15 dB 以上的语音无需处理,可以直接使用;所以,追求过低信噪比条件下端点检测性能的提高在实际应用中意义不大。同时,端点检测技术是语音标注的前序技术,工程应用中为保证准确率会采用人工标注,人工语音标注是专业人员通过观察语音波形和收听语音内容进行文本或韵律的标注。用于人工标注的语音信号一般质量较好,信噪比不会过低。

通过阅读参考文献,结合语音标注中的信噪比要求,总结得到:在语音信号处理中,根据人类感觉极限设定的信噪比极大值 $\mathrm{SNR_{MAX}}$ 的典型值为 30 dB,信噪比极小值 $\mathrm{SNR_{MIN}}$ 设置为 -10 dB 较为合适。当信噪比过高时,噪声功率对于语音信号功率而言可忽略不计,可近似认为是纯净语音;当信噪比过低时,语音信号功率远小于噪声功率,再进一步进行处理,对人体听觉器官造成的影响已经可以忽略不计,并且过低的信噪比在语音标注研究中意义不大。同时根据带噪语音的信噪比,可以将带噪语音分成高信噪比语音(信噪比高于 15 dB)、中等信噪比语音(信噪比在 5 dB 和 15 dB 之间)和低信噪比语音(信噪比低于 5 dB)。其中临界信噪比值 5 dB 和 15 dB 分别向上归类,属于中等信噪比和高信噪比。

为验证信噪比范围和分类的准确性,对一段纯净语音叠加不同信噪比的粉红噪声进行实验。粉红噪声是自然界最常见的噪声,也是最常用于声学测试的声音。在 -10 dB 到 30 dB 范围内平均间隔 5 dB 选择一个测试点,图 1 为纯语音信号与信噪比分

别为 30、25、20、15、10、5、0、−5 和−10 dB 的带噪语音信号的时域波形。通过波形观察可知,当带噪语音的信噪比低于5 dB时,随着信噪比的降低,语音信号逐渐淹没在噪声中,通过肉眼观测很难对语音段中每个汉字音节进行区分,人耳分辨音节的能力逐渐丧失,甚至无法分辨。当信噪比高于 15 dB 时,噪声对语音信号的影响极小,验证了信噪比分类的合理性。

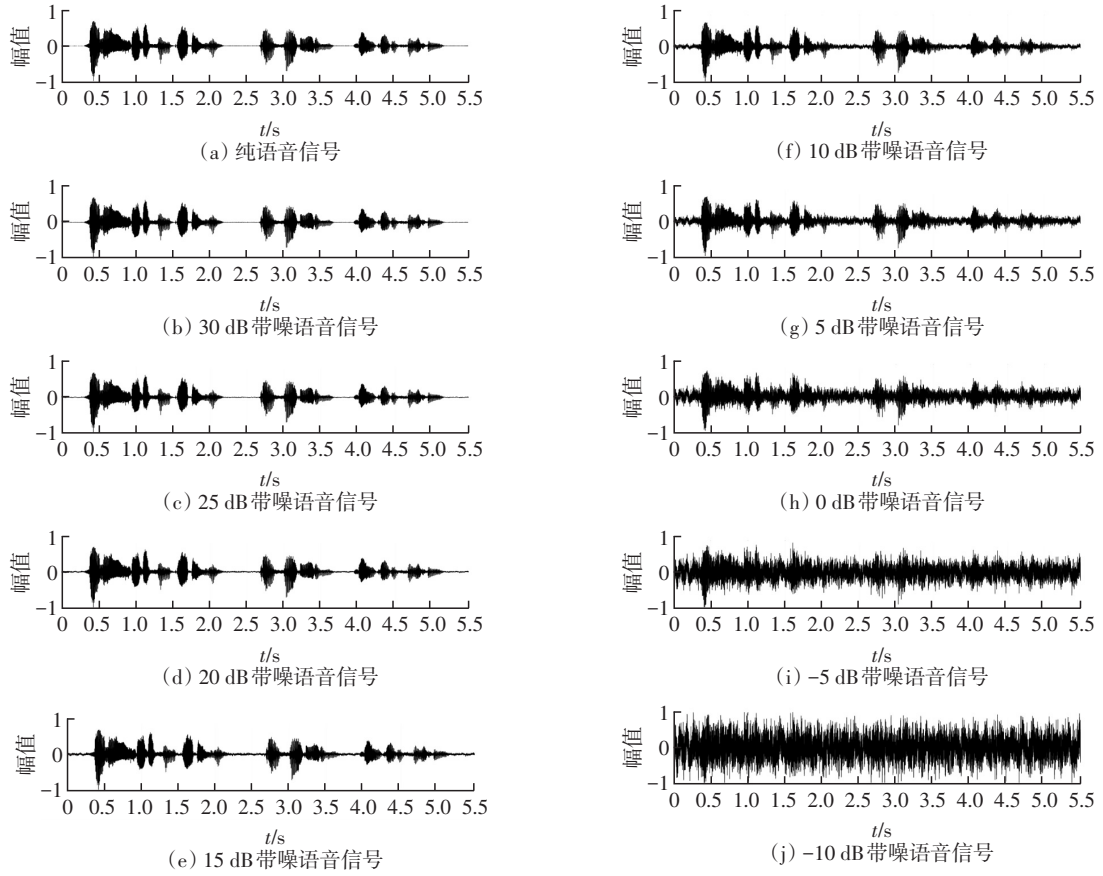


图 1　纯语音信号和信噪比为−10~30 dB 的带噪语音信号时域波形

1.2　噪声叠加与信噪比计算

根据信噪比的定义公式(1)可知,单有带噪语音很难求得其信噪比,所以在研究工作中,大多数学者会将纯净语音信号与噪声信号进行叠加,产生一定信噪比的带噪语音进行实验研究。在生成叠加了任意噪声的带噪语音的过程中,首先要判断纯净语音信号的采样频率与噪声的采样频率是否相等,若不相等需要对噪声重采样,使两者频率相同。同时,若噪声数据的长度与纯语音数据的长度不相等,要对噪声数据进行截断或补足。根据纯语音信号和去除直流分量的噪声信号分别求得两者的能量,根据设定的信噪比计算噪声的设定方差值,并以此调整噪声的幅值,最后通过叠加构成带噪语音。

在纯语音信号能量已知的条件下,带噪语音信号与纯语音信号能量之差就是噪声的能量,根据信噪比计算公式(1)可直接求得带噪语音的信噪比。但在语音标注应用中,通过带噪语音往往无法直接获得纯净语音信号的能量,此时需要通过噪声估计计算噪声的能量,再利用公式(1)计算信噪比。语音在活动期间,带噪语音信号在单个频带的功率通常会衰减到噪声的功率水平。因此,可以通过追踪短时窗内带噪语音谱每个频带的最小值,得到各频带内噪声水平的估计值[16]。由噪声功率谱计算整段信号的噪声能量,同时由于带噪语音信号已知,可直接求得带噪语音能量,利用信噪比公式(1)可计算得到语音信号的信噪比。

2　基于信噪比分类的自适应端点检测算法

复杂噪声环境下的语音端点检测是语音信号处理研究领域的热点之一。传统方法侧重于对语音信号进行特征提取,从而忽略了背景噪声的影响,导致端点检测结果不够理想。对于高信噪比的语音信号,选择传统的端点检测算法,就能够达到较高的准确率;而当语音信号的信噪比急剧下

降时,传统方法的准确率就直线下降,此时需要寻求更为鲁棒的端点检测算法。通过分析计算语音信号的信噪比,根据信噪比选择对应高性能的端点检测算法,显得至关重要。本文在语音信号的预处理环节增加了信噪比计算和语音分类,输入语音信号根据计算得到的信噪比大小分成高信噪比、中等信噪比和低信噪比 3 类,并自动选择对应较优的端点检测算法进行处理,得到端点检测结果。基于信噪比分类的自适应端点检测算法流程如图 2 所示。

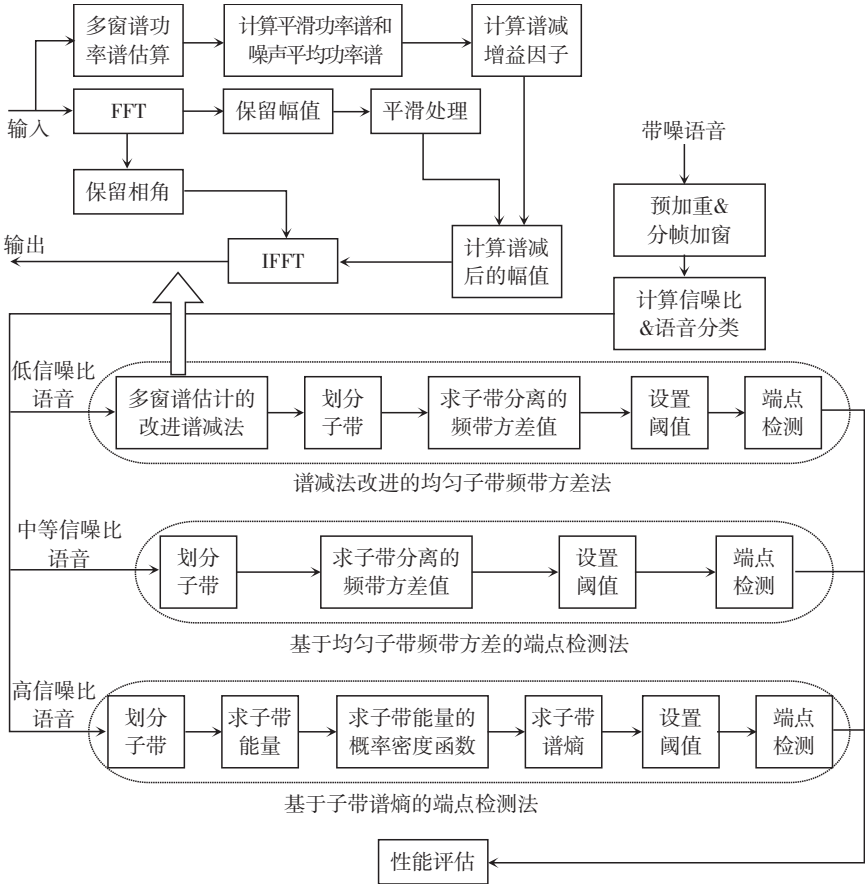


图 2 基于信噪比分类的自适应端点检测算法流程图

根据信噪比对语音信号进行分类,信噪比处于 15 dB 到 30 dB 的带噪语音定义为高信噪比语音,采用基于子带谱熵的端点检测方法(下文统称为高 SNR 算法)。谱熵作为一种常用的频域特征参数,在信噪比较高的环境下,应用于端点检测能得到较为理想的结果。但随着信噪比的降低,语音信号与噪声信号的谱熵值差别逐渐减小[17]。当信噪比低于 5 dB 时,谱熵曲线几乎没有明显的上升沿和下降沿,此时无法再用此算法进行准确的端点检测。信噪比处于 5 dB 到 15 dB 的带噪语音定义为中等信噪比语音,采用基于均匀子带频带方差的端点检测方法(下文统称为中等 SNR 算法)。与子带谱熵特征相比,子带方差在语音段和非语音段有更好的区分度,后续的阈值更好设置,在信噪比不太理想的情况下,端点检测更为准确[3]。信噪比处于−10 dB 到 5 dB 的带噪语音定义为低信噪比语音,由于此情况下的语音质量不佳,需要先对带噪语音进行语音增强,采用多窗谱估计的改进谱减法进行降噪,随后再采用基于均匀子带频带方差法进行端点检测(下文统称为低 SNR 算法)。

2.1 基于子带谱熵的端点检测算法

语音的谱熵一般低于噪声的谱熵,根据这一特性可以很好地区分语音段和噪声段[17]。当语音信号受到噪声干扰时,语音信号的谱熵值会随着信噪比的降低而降低,影响对语音和非语音的区分[17]。此时可以通过将多子带分析技术与谱熵计算相结合,采用子带谱熵作为特征参数进行端点检测。基于子带谱熵的端点检测方法将每一帧语音信号分成若干个子带,求每一个子带的谱熵,进而消除每一条谱线的幅值受噪声影响的问题[17]。下面是子带谱熵的计算方法:

(1) 设带噪语音信号 $x(n)$ 加窗分帧后得到的第

i 帧语音信号为 $x_i(m)$,对其进行离散傅里叶变换得到频谱 $X_i(k)$,每个分量的能量 $Y_i(k)=|X_i(k)|^2$。

(2) 对 $X_i(k)$ 的每一帧信号划分子带。假设每个子带由 s 条谱线组成,共有 N_b 个子带,此时第 i 帧语音信号中的第 m 个子带的子带能量为[17]

$$E_b(m,i)=\sum_{k=(m-1)*s}^{(m-1)*s+s-1}Y_i(k),1\leqslant m\leqslant N_b \tag{2}$$

(3) 根据文献[18]提出的算法引入一个常数 K 到子带谱熵法分布式中,用于提高语音和噪声的区分度,得到新的子带能量的概率分布密度为

$$p'_b(m,i)=\frac{E_b(m,i)+K}{\sum_{k=1}^{N_b}(E_b(k,i)+K)},K>0 \tag{3}$$

(4) 得到子带谱熵的定义为

$$H'_b(i)=-\sum_{m=1}^{N_b}p'_b(m,i)\log p'_b(m,i) \tag{4}$$

根据子带谱熵的计算式(4),可以计算得到前导无话段的平均子带谱熵值,并以此设置双门限法中的两个阈值。此算法只采用了子带谱熵这一个特征参数,并且由于语音的谱熵值小于噪声的谱熵值,所以要采用单参数的双门限反向检测法。若某帧的子带谱熵值低于较小阈值,则此帧一定是语音帧,再根据子带谱熵值何时低于较高阈值来判断语音的端点。

2.2　基于均匀子带频带方差的端点检测算法

频带方差检测法计算某一帧语音信号各频带能量的方差,以此作为特征参数检测语音的起止端点[19]。由于频带方差的检测算法是对每条谱线求解方差,得到的方差波动较大,稳定性较差。在此基础上,可以利用均匀子带频带方差的方法进行改进。下面是均匀子带频带方差的计算步骤:

(1) 设带噪语音信号 $x(n)$ 加窗分帧后得到的第 i 帧语音信号为 $x_i(m)$,对其进行离散傅里叶变换得到频谱 $X_i(k)$。

(2) 在频谱的正频域内存在 $(N/2+1)$ 条谱线,将这 $(N/2+1)$ 条幅度谱线分割成 q 个子带,每个子带含有 $s=\mathrm{fix}\left[\left(\frac{N}{2}+1\right)/q\right]$ 条谱线(其中 fix[·] 表示取其整数部分),则构成的子带有[20]

$$XX_i(m)=\sum_{k=1+(m-1)s}^{1+(m-1)s+(s-1)}|X_i(k)| \tag{5}$$

(3) 令 $XX_i=\{XX_i(1),XX_i(2),\cdots,XX_i(q)\}$,计算得到的均值为

$$E_{i,1}=\frac{1}{q}\sum_{k=1}^{q}XX_i(k) \tag{6}$$

(4) 得到方差的定义式为

$$D_{i,1}=\frac{1}{q-1}\sum_{k=1}^{q}[XX_i(k)-E_{i,1}]^2 \tag{7}$$

根据式(7)求出每帧均匀子带频带方差值后,再根据已知前导无话段的帧数,求出相应阈值,并采用单参数的双门限判决法来确定语音的起止位置[21]。由于语音的子带方差值大于噪声的子带方差值;所以,当某帧的子带方差值高于较大阈值时,肯定是语音帧,再根据子带方差值何时高于较低阈值来判断语音信号的端点。

2.3　谱减法改进的均匀子带频带方差的端点检测算法

低信噪比语音中的语音信号被各种噪声干扰、甚至淹没,需要通过语音增强技术从噪声背景中提取有用的语音信号。语音增强有许多常用方法,其中谱减法由于算法简单和便于计算,是应用最广泛的语音增强方法[21]。常见谱减法有基本谱减法[22]、多窗谱估计的改进谱减法[21]和调制域谱减法[23-24]。在-5 dB 高斯白噪声环境下,基本谱减法的去噪效果已经不是很明显[24]。在 0 dB、f16 噪声环境下,多窗谱估计谱减法和调制域谱减法噪声残留较少,效果较好[24]。综合考虑去噪效果和算法实现的复杂程度,选择多窗谱估计的改进谱减法。

多窗谱估计是对同一数据序列采用多个正交的数据窗分别求直接谱,然后求平均,得到谱估计,因此能得到较小的估计方差。利用多窗谱估计实现谱减的步骤如下[21]:

(1) 带噪语音信号 $x(n)$ 进行加窗分帧预处理后得到的第 i 帧语音信号为 $x_i(m)$,对分帧后的信号做 FFT,分别求出幅度谱 $|X_i(k)|$ 和相位谱 $\theta_i(k)$。在相邻帧间做平滑处理,以 i 帧为中心前后各取 M 帧,共 $(2M+1)$ 帧进行平均,计算得到平均幅度谱

$$|\bar{X}_i(k)|=\frac{1}{2M+1}\sum_{j=-M}^{M}|X_{i+j}(k)| \tag{8}$$

(2) 把分帧后的信号进行多窗谱估计,得到多窗谱功率谱密度

$$P(k,i)=\mathrm{PMTM}[x_i(m)] \tag{9}$$

(3) 对多窗谱功率谱密度估计值进行相邻帧之间的平滑处理,计算得到平滑功率谱密度

$$P_y(k,i)=\frac{1}{2M+1}\sum_{j=-M}^{M}P(k,i+j) \tag{10}$$

(4) 已知前导无话段共 NIS 帧,计算得到噪声

的平均功率谱密度值为

$$P_n(k)=\frac{1}{\mathrm{NIS}}\sum_{i=1}^{\mathrm{NIS}}P_y(k,i) \tag{11}$$

(5) 利用谱减关系计算增益因子 $g(k,i)$,其中 α 为过减因子,β 为增益补偿因子。

$$g(k,i)=\begin{cases}\dfrac{P_y(k,i)-\alpha P_n(k)}{P_y(k,i)} & P_y(k,i)-\alpha P_n(k)\geqslant 0\\ \dfrac{\beta P_n(k)}{P_y(k,i)} & P_y(k,i)-\alpha P_n(k)<0\end{cases} \tag{12}$$

(6) 根据增益因子 $g(k,i)$ 和平均幅度谱 $|\overline{X}_i(k)|$ 可以求得谱减后的幅度谱 $|\hat{X}_i(k)|=g(k,i)*|\overline{X}_i(k)|$。将谱减后的幅度谱与步骤(1)中的相位谱结合进行 IFFT,可得到降噪后的语音信号

$$\hat{x}_i(m)=\mathrm{IDFT}[\,|\hat{X}_i(k)|\exp(\mathrm{j}\theta_i(k))\,] \tag{13}$$

降噪后的语音信号采用基于均匀子带频带方差的端点检测方法,根据单参数双门限判决法确定语音端点的位置。

3 仿真实验与结果分析

3.1 实验环境

实验音频采用一段内容为“海轮随着海波荡漾,在海港里,静静地航行”的单声道男声纯净语音信号 seaboat.wav,音频长度为 5.5 s,采样频率为 16 kHz,16 bits 量化。同时采用清华大学 30 h 中文语音库 thchs30 中的 100 条纯净语音,采样频率为 16 kHz,采样大小为 16 bits。噪声数据来自 NOISEX-92 噪声语料库,选取白噪声(下文用 white 表示)、餐厅内嘈杂噪声(下文用 babble 表示)和车内噪声(下文用 volvo 表示)。实验在 Win10 操作系统、MATLAB R2018b 软件上进行。输入语音信号在分帧加窗预处理中使用 Hamming 窗,帧长为400 帧,帧移为 160 帧。在基于子带谱熵和基于均匀子带频带方差法中,每个子带由 4 条谱线组成;子带谱熵法中 $K=0.5$。

本文用正确率作为算法的评价指标。通过对 seaboat.wav 语音和 thchs30 语料库中的样本语音进行收听和目测,人工标记出语音的起始点和结束点,以此作为样本语音的参考端点,与算法得到的结果进行比较。端点检测正确率的定义为

正确率=(总帧数-错误帧数)/总帧数 (14)

其中,错误帧数是噪声误判为语音的帧数和语音误判为噪声的帧数的总和。

3.2 不同端点检测算法效果比较

本文针对不同信噪比环境选择了 3 种端点检测算法,为检验 3 种算法在对应信噪比范围内的端点检测效果,将纯净语音信号 seaboat.wav 叠加 babble 噪声进行实验。由于实际生活中常见噪声多具有随机性和非平稳性,故选择 babble 噪声作为测试噪声。将 babble 噪声信号与纯净语音信号混合成 9 条-10 dB 到 30 dB 范围内间隔 5 dB 的带噪语音信号,并用 3 种算法进行端点检测。信噪比为 20 dB、10 dB 和 0 dB 的 3 种噪声环境下的检测结果分别如图 3 至图 5 所示。

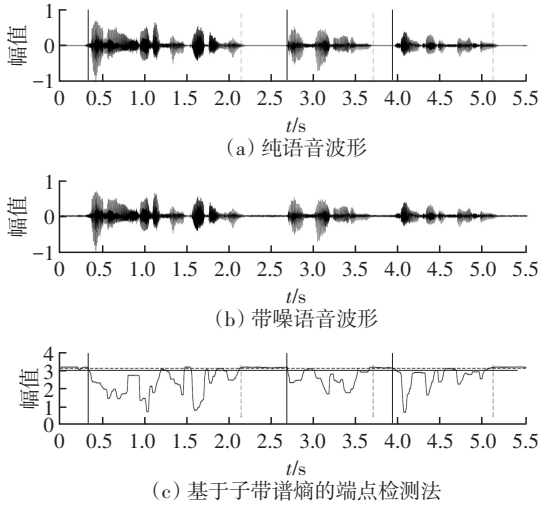


图 3 信噪比为 20 dB 的 babble 环境下端点检测结果

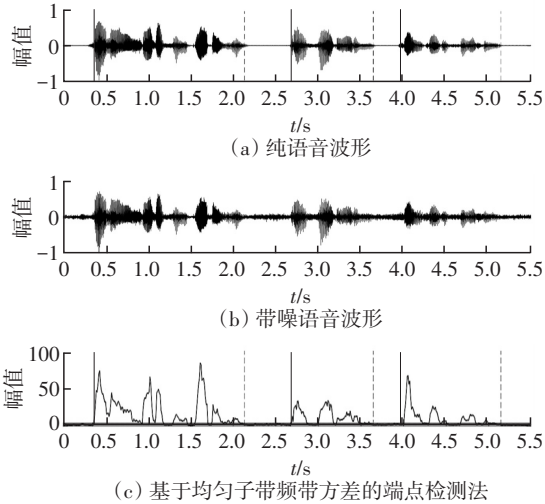
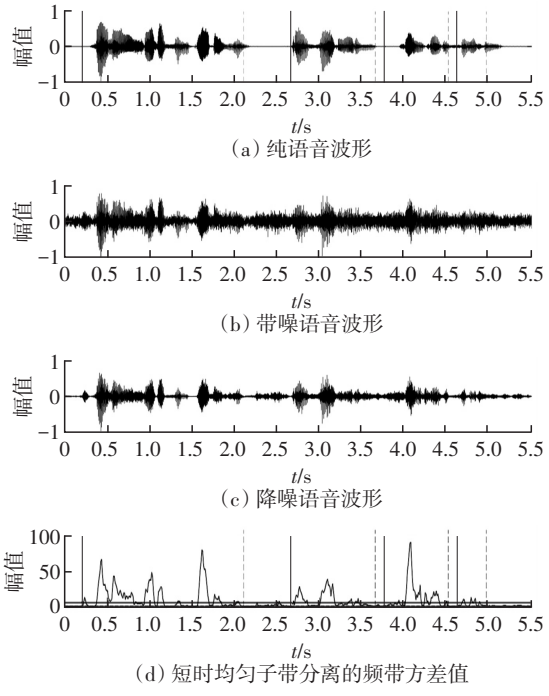


图 4 信噪比为 10 dB 的 babble 环境下端点检测结果

图 3 至图 5 中,图 3(a)、图 4(a)、图 5(a)给出了原始语音信号的时域波形以及端点检测结果,实线代表语音段的起始位置,虚线代表语音段的终止

位置。图 3(c)、图 4(c)、图 5(d)展示了端点检测算法中特征参数值随时间变化的情况,其中水平实线和虚线分别代表双门限法的两个阈值,垂直实线和虚线对应语音段的起止位置。由图 3 至图 5 可知,当信噪比为 20 dB 和 10 dB 时,高 SNR 算法和中等 SNR 算法都能准确检测出语音段。但当信噪比为 0 dB 时,低 SNR 算法虽然能够检测出每段语音的位置,但由于降噪过程中产生的部分音乐噪声,导致算法将部分静音帧误判为语音帧,存在一定误差。



(a) 纯语音波形
(b) 带噪语音波形
(c) 降噪语音波形
(d) 短时均匀子带分离的频带方差值

图 5　信噪比为 0 dB 的 babble 环境下端点检测结果

3 种算法的检测正确率见表 1。从表 1 中可知,高 SNR 算法在 15~30 dB 范围内的正确率较高,尤其当信噪比高于 20 dB 时,正确率能达到 97%以上。当信噪比低于 15 dB 时,正确率就大幅下降,尤其是在 0 dB 以下的高噪声环境中,正确率很低。中等 SNR 算法的检测正确率在 5~15 dB 范围内较高,信噪比低于 0 dB 的环境中大幅下降,因此该方法在高噪声环境下并不适用。但对于语音质量较高的带噪语音,检测正确率也有所下降。多次实验并分析算法原理后可知,基于均匀子带频带方差法的阈值是根据前导无话段的子带频带方差平均值设置的,此均值与信噪比密切相关,且随信噪比变化剧烈,导致不同信噪比语音的最优阈值各不相同。阈值设定需要根据不同的信噪比进行适当调整,才能达到较优的检测正确率。在中等 SNR 算法中,是以使 5~15 dB 信噪比范围内的检测正确率较高为目标设定的阈值。低 SNR 算法在-10~5 dB 范围内保持了一个较好的检测正确率,比同噪声环境下另外两种算法效果都好,适用于低信噪比环境。

不难看出,高信噪比范围(15~30 dB)内 3 种方法的正确率都较高,而低信噪比范围(-10~5 dB)内检测正确率较低,甚至部分方法并不适用,间接验证了信噪比分类的合理性。同时,通过比较相同信噪比条件下不同算法的正确率可知,15~30 dB 信噪比范围内高 SNR 算法正确率最高,5~15 dB 信噪比范围内中等 SNR 算法较为合适,-10~5 dB 范围内应采用低 SNR 算法,虽然低 SNR 算法在较高质量语音中检测效果也很优秀,但考虑到算法复杂性,可以采用其他简单算法替代。此结论也验证了算法选择的合理性。

表 1　不同信噪比条件下端点检测算法正确率比较　%

信噪比/dB	高 SNR 算法	中等 SNR 算法	低 SNR 算法
-10	44.7	31.6	72.8
-5	55.3	36.0	80.8
0	54.0	69.7	88.1
5	71.7	90.7	93.8
10	81.0	97.6	97.3
15	91.1	96.4	96.2
20	97.8	92.5	96.5
25	98.4	91.6	96.2
30	98.5	90.0	94.7

3.3　基于信噪比分类的端点检测算法验证

为验证算法的整体检测效果,随机选取 thchs30 语音库中的 100 条语音,分别叠加 white 噪声、babble 噪声和 volvo 噪声,生成随机信噪比的带噪语音。由于此算法的应用场景是语音标注技术,处理的音频数据质量较优,所以将 100 条语音按照低信噪比、中等信噪比和高信噪比 3 类比例为 15∶40∶45 进行噪声叠加,即高信噪比范围内随机抽取 45 条纯净语音,中等信噪比范围随机抽取 40 条语音,低信噪比范围随机抽取 15 条语音。同时,纯净语音信号采用 rand 函数叠加随机信噪比值的噪声,计算带噪语音的信噪比后选择对应算法进行端点检测,并计算检测正确率。表 2 给出了低、中、高信噪比范围内的平均检测正确率和 100 条语音的整体检测正确率。

表 2　基于信噪比分类的端点检测算法正确率比较　%

噪声环境	低 SNR	中等 SNR	高 SNR	整体
white	88.8	96.0	97.1	95.4
babble	81.5	94.1	91.3	91.0
volvo	86.5	77.2	97.0	87.5

在计算带噪语音信噪比时,部分信噪比在临界值 5 dB 和 15 dB 附近的语音会存在分类错误的情况,但绝大部分语音都能准确分类。从表 2 中可知,本文提出的基于信噪比分类的端点检测算法在 white、babble 和 volvo 噪声环境下的总体正确率均能达到 85%以上,尤其是在 white 噪声环境下,整体正确率高达 95%以上。通过观察各信噪比范围内的平均正确率,发现即使在语音质量不好的低信噪比环境下,正确率也能达到 80%以上。

为验证本文算法的有效性,将本文算法与基于子带谱熵的端点检测算法(表 3 中简称为谱熵法)、基于均匀子带频带方差的端点检测算法(表 3 中简称为方差法)和文献[22]中提出的算法(表 3 中简称为文献[22]方法)进行了检测正确率的比较,具体结果见表 3。文献[22]提出的算法采用改进谱减法和频谱方差相结合的语音端点检测算法,是现阶段常见的端点检测算法。由表 3 分析可知,在 white 和 babble 噪声环境下,本文算法的检测正确率均高于对比方法。在 volvo 噪声环境下,使用基于子带谱熵的端点检测算法正确率更高,但本文算法的检测正确率仍高于方差法和文献[22]方法。可见本文算法具有更好的检测正确率,整体检测效果符合预期,应用于语音标注能达到较高的检测准确性。

为进一步验证本文算法的复杂度,以端点检测速度为指标,在 white 噪声条件下采用本文算法和文献[22]方法对 100 条语音进行端点检测,并计算检测时间的平均值。本文算法的端点检测时间为 12.969 s,文献[22]方法的端点检测时间为13.305 s。对比可知,两种算法的复杂度相当,所用时间相近。综合比较端点检测正确率和算法复杂度可知,本文算法能在不提高算法复杂度的前提下,达到较高的端点检测正确率。

表 3 不同端点检测算法的正确率比较 %

噪声环境	本文算法	谱熵法	方差法	文献[22]方法
white	95.4	93.5	90.4	90.9
babble	91.0	83.7	86.8	89.0
volvo	87.5	92.9	84.6	86.9

4 结束语

本文提出了一种基于信噪比分类的端点检测算法。该算法首先根据预先设定的标准对信噪比分类,在每类范围内选择对应较优的算法。实验结果表明,针对语音标注中较高质量的语音信号,本文算法的整体正确率高,适用于实际标注应用。同时,对语音信号进行预先分类,采用适当的端点检测算法,能在一定程度上提高效率。但是,本算法依旧存在不足,低信噪比环境下的检测正确率仍有待提高,对于部分噪声信号适用性不强。如何进一步改进算法,使其适应更复杂的噪声环境是今后研究的方向。

参考文献:

[1] 周学文,呼和.语音声学参数自动标注/提取系统简介[J].中文信息学报,2014,28(3):123-128.
ZHOU Xuewen, HU He. Introduction to automatic labeling/retrieving system for acoustic parameters [J]. Journal of Chinese Information Processing, 2014, 28(3): 123-128. (in Chinese)

[2] GHOSH D, MURALISHANKAR R, GURUGOPINATH S. Robust voice activity detection using frequency domain long-term differential entropy [C]// International Speech Communication Association. 2018: 1220-1224.

[3] KYRIAKIDES A, PITRIS C, FINK A, et al. Isolated word endpoint detection using time-frequency variance kernels [C]// Conference Record of the 45th Asilomar Conference on Signals, Systems and Computers. 2011: 585-589.

[4] TAN Z H, LINDBERG B. Low-complexity variable frame rate analysis for speech recognition and voice activity detection[J]. IEEE Journal of Selected Topics in Signal Processing, 2010, 4(5): 798-807.

[5] MA Y N, NISHIHARA A. Erratum to: efficient voice activity detection algorithm using long-term spectral flatness measure [J]. EURASIP Journal on Audio, Speech, and Music Processing, 2013, 2013: 87.

[6] LI X, HORAUD R, GIRIN L, et al. Voice activity detection based on statistical likelihood ratio with adaptive thresholding [C]// IEEE International Workshop on Acoustic Signal Enhancement. 2016: 1-5.

[7] WU J, ZHANG X L. Efficient multiple kernel support vector machine based voice activity detection [J]. IEEE Signal Processing Letters, 2011, 18(8): 466-469.

[8] WU X, ZHU M, WU R, et al. A self-adapting GMM based voice activity detection[C]// IEEE 23rd International Conference on Digital Signal Processing (DSP). 2018: 1-5.

[9] DENG S, HAN J, ZHENG T, et al. A modified MAP criterion based on hidden Markov model for voice activity detection [C]// IEEE International Conference on Acoustics, Speech and Signal Processing. 2011: 5220-5223.

[10] ZHANG X L, WU J. Deep belief networks based voice activity detection[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21(4): 697-710.

[11] THOMAS S, GANAPATHY S, SAON G, et al. Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions [C]// IEEE

International Conference on Acoustics, Speech and Signal Processing.2014:2519-2523.

[12] HUGHES T, MIERLE K. Recurrent neural networks for voice activity detection [C] // IEEE International Conference on Acoustics, Speech and Signal Processing. 2013:7378-7382.

[13] EYBEN F, WENINGER F, SQUARTINI S, et al. Real-life voice activity detection with LSTM recurrent neural networks and an application to Hollywood movies[C]//IEEE International Conference on Acoustics, Speech and Signal Processing.2013:483-487.

[14] 李晔.数字语音编码技术[M].北京:电子工业出版社,2013:44-46.

[15] 姚黎.车载语音识别系统的语音增强方法研究[D].武汉:武汉理工大学,2012.
YAO Li.Research of the speech enhancement methods in car speech recognition system[D].Wuhan: Wuhan University of Technology,2012.(in Chinese)

[16] PHILIPOS C L.Speech Enhancement[M].Chengdu: University of Electronic Science and Technology of China Press,2012:337-339.

[17] ZHANG Y, WANG K, YAN B.Speech endpoint detection algorithm with low signal-to-noise based on improved conventional spectral entropy[C]//12th World Congress on Intelligent Control and Automation.2016:3307-3311.

[18] 王琳,李成荣.一种基于自适应谱熵的端点检测改进方法[J].计算机仿真,2010,27(12):373-375,395.
WANG Lin, LI Chengrong. An improved speech endpoint detection method based on adaptive band-partition spectral entropy[J].Computer Simulation,2010,27(12): 373-375,395.(in Chinese)

[19] 杨晓海.基于语音端点检测的移动设备无障碍出行服务助手的研究与实现[D].杭州:浙江大学,2017.
YANG Xiaohai. An accessible travel service assistant for mobile device based on speech activity detection [D]. Hangzhou:Zhejiang University,2017.(in Chinese)

[20] 朱春利,李昕.基于多特征融合与动态阈值的语音端点检测方法[J].计算机工程,2019,45(2):250-257.
ZHU Chunli, LI Xin. Speech endpoint detection method based on multi-feature fusion and dynamic threshold[J]. Computer Engineering,2019,45(2):250-257.(in Chinese)

[21] 韩芳,靳宗信.低信噪比下的端点检测算法研究[J].西北师范大学学报(自然科学版),2016,52(5):55-59.
HAN Fang,JIN Zongxin.Study of endpoint detection algorithm in low SNR[J].Journal of Northwest Normal University (Natural Science),2016,52(5):55-59.(in Chinese)

[22] 王路露,夏旭,冯璐,等.基于频谱方差和谱减法的语音端点检测新算法[J].计算机工程与应用,2014,50(8):194-197.
WANG Lulu, XIA Xu, FENG Lu, et al. New speech endpoint detection algorithm based on spectrum variance and spectral subtraction[J]. Computer Engineering and Applications,2014,50(8):194-197.(in Chinese)

[23] 王瑶,曾庆宁,龙超,等.低信噪比环境下语音端点检测改进方法[J].声学技术,2018,37(5):457-467.
WANG Yao, ZENG Qingning, LONG Chao, et al. An improved speech endpoint detection method under low SNR [J]. Technical Acoustics, 2018, 37(5): 457-467.(in Chinese)

[24] 王群,曾庆宁,郑展恒.低信噪比下语音端点检测算法的改进研究[J].科学技术与工程,2017,17(21):50-56.
WANG Qun, ZENG Qingning, ZHENG Zhanheng. Research of speech endpoint detection in low SNR environment[J].Science Technology and Engineering,2017, 17(21):50-56.(in Chinese)