doi:10.3969/j.issn.1001 - 893x.2013.08.013

# 改进的基于长时谱能量差异和基音比例的语音检测方法*

孟一鸣[1,**],欧智坚[2]

(清华大学 电子工程系,北京 100084)

**摘 要**:语音检测是语音信号处理的前端,利用长时谱能量差异特征的语音检测无法区分突发噪声和语音,掺杂着突发噪声的语音信号会对语音处理系统带来不良影响。提出了一种基于长时谱能量差异特征和基音比例特征相结合的语音检测方法,该方法的优点是,在利用长时谱能量差异特征基础上引入基音比例特征,从而有效减少了将信号中突发噪声误判为语音的错误。实验显示,该算法能够在多种信噪比环境下取得很好的检测结果。

**关键词**:语音信号处理;语音检测;长时谱能量差异;基音比例;突发噪声

**中图分类号**:TN912.3 **文献标志码**:A **文章编号**:1001 - 893X(2013)08 - 1039 - 05

# Improved Voice Activity Detection Based on Long-term Spectral Divergence and Pitch Ratio Features

MENG Yi-ming[1], OU Zhi-jian[2]

(Department of Electronic Engineering, Tsinghua University, Beijing 100084, China)

**Abstract**: Voice Activity Detection (VAD) is the front-end of speech processing and the VAD algorithm which uses long-term spectral divergence (LTSD) feature can't discriminate abrupt noise from speech. The speech signal with abrupt noise will adversely affect the speech processing system. This paper proposes a VAD algorithm which combines LTSD feature and pitch ratio feature. The advantage of the algorithm is that by introducing pitch ratio feature, it can effectively reduce the false alarms of taking abrupt noise as speech. Experimental results show that the algorithm achieves good performance for VAD under various signal-to-noise ratios.

**Key words**: speech processing; voice activity detection; long-term spectral divergence; pitch ratio; burst noise

## 1 引 言

语音检测,顾名思义就是指从一段信号中检测出是否含有语音,并确定语音的起始位置。语音检测是语音信号处理系统中的一个重要组成部分,准确的语音检测能够有效提高后期语音信号处理的精确度,并减少系统的运算量。

目前常用的语音检测方法主要分为基于统计模型的方法[1]和基于特征门限的方法。基于统计模型的方法具有较高的准确率,但是其依赖于有效的训练数据,且运算量较大。相对于前者,基于特征门限的方法实时性较高,但要求所选择的特征具有稳健性[2-3]。

在实际的应用过程中,我们通常需要面对种类繁多的背景噪声以及不同的信噪比环境,好的语音检测算法应具有较强的自适应能力;同时在语音处理系统中多存在实时性要求,故算法应尽量简单以降低运算量。

Ramlrez J 等人提出的基于语音长时谱能量差异特征[4]是一种能有效区分语音和背景噪声的特征,但是该特征却无法分辨出突发噪声和语音,这就使得在后期处理过程中会引入许多突发噪声的特性。语音中一个重要的特征就是基音特征,在发声过程中,基

* 收稿日期:2013 - 04 - 03;修回日期:2013 - 05 - 09 Received date:2013 - 04 - 03;Revised date:2013 - 05 - 09

** 通讯作者:camelmengnudt@163.com **Corresponding author**: camelmengnudt@163.com

音是连续存在且占有一定比例的。而突发噪声大部分不存在基音频率,即使出现,其比例也非常少。

本文提出了将语音的长时谱能量差异和语音的基音比例作为特征,通过自适应噪声调整判决门限的语音端点检测方法,该方法在多种噪声的不同信噪比条件下具有较强的稳定性。经过后期试验验证,该算法应用在说话人识别系统中能够有效提高系统的实际性能。

## 2 算法描述

### 2.1 语音检测模型

对于语音检测器,输入是实时的数字语音信号,输出是具有一定延迟的语音检测判决序列(二进制 0 – 1 序列)。因为语音是具有一定持续时间的统计信号,我们无法对每个采样点单独进行语音/非语音的判决,所以语音检测算法采取分帧判决的方式。语音检测器对输入语音进行分帧、加窗,然后对帧信号进行分析,给出对每帧信号的语音/非语音判决。

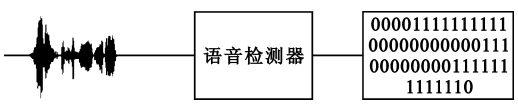


图 1 语音检测模型

Fig.1 VAD model

语音检测的核心部分主要包括特征选择、噪声估计以及门限判别。

### 2.2 长时谱能量差异

人类耳蜗的特殊构造使人类对语音的频谱信息十分敏感,使得语音的理解十分依赖信号的频谱分布[6]。而且语音作为非平稳的随机过程,在时域上与噪声的幅度分布相近,难以区分。而语音存在着上下文联系,综合前后多帧数据联合起来判别会增加判别的准确性[5]。

基于以上两点考虑,本文采用了长时谱能量差异(Long-Term Spectral Divergence, LTSD[4])作为信号处理的主要特征。

令 $x$ 是一段含有噪声的语音信号,其被分为相互交叠的帧 $\tilde{x}_1 \cdots \tilde{x}_n$,第 $i$ 帧 $\tilde{x}_i$ 经过 FFT 变换后得到幅度谱 $X_i$,$X_i$ 第 $k$ 个频带的值为 $X_i(k)$。定义第 $i$ 帧的 $N$ 阶长时谱包络(Long-Term Spectral Envelope, LTSE[4])为

$$LTSE_i^N(k) = \max\{X_j(k)\}_{j=i-N}^{j=i+N} \tag{1}$$

定义 $N$ 阶长时谱能量差异(LTSD)为

$$LTSD_i^N = \frac{1}{NFFT}\sum_{k=0}^{NFFT}\frac{[LTSE_i^N(k)]^2}{E^2(k)} \tag{2}$$

其中,$E(k)$代表当前(第 $i$ 帧处)对噪声幅度谱第 $k$ 个频带的估计值。

LTSD 特征结合了 $2N+1$ 帧的信息,最大化了当前帧与噪声的差异,从图 2 中可以看出,该特征能够较好地区分出噪声和语音。

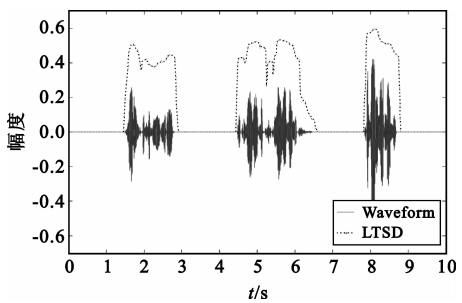


图 2 一段语音的 LTSD 特征

Fig.2 LTSD characteristic of a segment of voice

### 2.3 噪声估计

在计算 LTSD 特征以及更新判别门限时,需要对噪声幅度谱 $E(k)$进行估计。通常情况下,假定一段音频开始时前几帧数据为非语音,利用此段语音可以对噪声 $E(k)$进行初始化,得到 $E_0(k)$。之后在语音检测的过程中,可以利用判定为非语音帧的统计信息对噪声估计进行更新,依次得到 $E_1(k)$,$E_2(k)$,$E_3(k)$,…(下标表示噪声更新序号)。

在本文中,当连续被判为非语音的帧数达到 $L$ 时,开始对噪声信息进行更新。假设当前为第 $h$ 次噪声更新,且当前帧为 $t$,通过对幅度谱 $X_{t-L},\cdots,X_t$ 进行统计,得到当前噪声的统计量 $\hat{E}_h$ 和 $\hat{\sigma}_h$,然后分别将 $\hat{E}_h$ 和 $\hat{\sigma}_h$ 与第 $h-1$ 次更新时噪声的估计值 $E_{h-1}$和 $\sigma_{h-1}$加权求和得到第 $h$ 次噪声估计值 $E_h$ 和 $\sigma_h$,其更新公式如下:

$$\hat{E}_h(k) = \frac{1}{L}\sum_{l=0}^{L-1}X_{t-l}(k) \tag{3}$$

$$E_h(k) = \alpha E_{h-1}(k) + (1-\alpha)\hat{E}_h(k) \tag{4}$$

$$\hat{\sigma}_h(k) = \sqrt{\frac{1}{L}\sum_{l=0}^{L}[X_{t-l}(k) - \hat{E}_h(k)]^2} \tag{5}$$

$$\sigma_h(k) = \alpha\sigma_{h-1}(k) + (1-\alpha)\hat{\sigma}_h(k) \tag{6}$$

其中,$\alpha$ 为惯性系数,它表示对于历史噪声的依赖程度。

### 2.4 门限判别

在 VAD 算法中,门限的设置对于算法的性能影

响很大。如果设定固定门限,在噪声变化不大的情况下,算法尚能正常运行。如果噪声变化较大,就需要门限能够随着噪声变化进行自我调整。

本文中,通过将最新的噪声估计值 $E_h$ 与 $\beta$ 倍噪声根方差估计值 $\sigma_h$ 求和得到当前噪声上界的估计值,然后计算出该上界的 LTSD 特征值,将其作为当前的判决门限 $\gamma$:

$$\gamma = \frac{1}{NFFT}\sum_{k=0}^{NFFT}\frac{[E_h(k)+\beta\sigma_h(k)]^2}{E_h^2(k)} \tag{7}$$

在算法的应用中,$\beta$ 值会随着信噪比的不同变化。当 $SNR \geqslant 30$ dB时,$\beta = 5$;当 $SNR \leqslant 5$ dB时,$\beta = 2.8$;其余 $SNR$ 条件下,$\beta$ 值为一个加权平均值。

$$\beta = 5 - \frac{SNR - 5}{30 - 5} \times (5 - 2.8) \tag{8}$$

### 2.5　基音比例特征

在平稳噪声环境中,LTSD 特征结合上下文信息,体现了噪声与语音频谱的差异,能够很好地区分出语音与噪声。但是当出现敲击键盘、碰撞麦克风等突发噪声时,由于此类噪声的 LTSD 特征与语音的 LTSD 特征极为相似,故常常被误判为语音。此时需要利用语音中的另一重要特征——基音比例来辅助判别。

语音中浊音是由周期性脉冲激励声道而产生,其声带振动频率称为基音频率[8]。我们对一段信号进行基音检测,统计出其中含有基音的帧数 $M_{\text{pitch}}$,然后计算 $M_{\text{pitch}}$与这段语音的帧数 $M$ 的比值 $\theta_{\text{pitch}} = M_{\text{pitch}}/M$,通过 $\theta_{\text{pitch}}$是否达到语音的判别门限 $\theta_v$ 可以区分出该段信号是语音还是噪声。

$$voice = \begin{cases} 1, & \theta_{\text{pitch}} > \theta_v \\ 0, & \theta_{\text{pitch}} < \theta_v \end{cases}$$

从图 3 和图 4 的对比可以看出,引入基音比例特征可以有效减少将突发噪声误判为语音的错误。

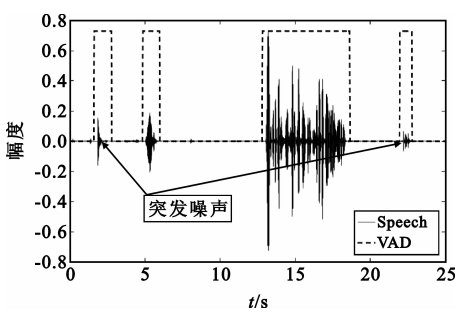


图 3　未引入基音比例判别的语音检测
Fig.3 VAD without introducing pitch ratio decision

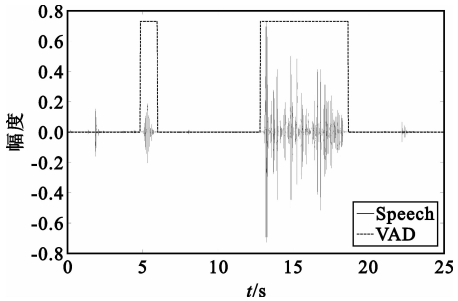


图 4　引入基音比例判别的语音检测
Fig.4 VAD with introducing pitch ratio decision

### 2.6　算法流程

结合 LTSD 特征以及基音比例特征的语音检测算法的基本思想如下:首先,利用 LTSD 特征对语音进行初步判断;然后,利用基音比例特征对已判断为语音的片段再次进行检测;最后,依据最短语音长度 $L_{\text{MinSent}}$以及最短非语音长度 $L_{\text{MinNoise}}$两个指标对判别结果进行平滑,消除过短的语音,以及句中停顿。

图 5 描述了整个算法的流程。

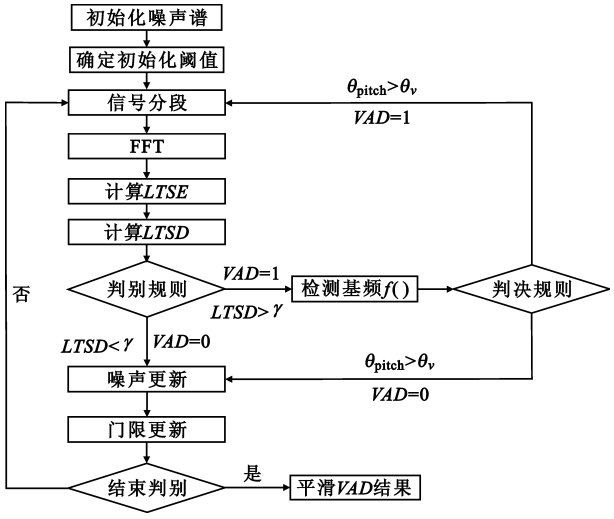


图 5　算法流程图
Fig.5 Flow chart of the proposed algorithm

## 3　实验及分析

### 3.1　实验数据

实验数据来自于美国国家标准化局 2008 年说话人识别评测(NIST Speaker Recognition Evaluation 2008)的训练和测试数据集。数据中主要为实际生活中的电话录音,以及面对面的采访录音。录音中以英语为主,同时含有其他国家语言。录音采用 8 kHz采样,8b $\mu$ 率编码。

由于不是针对语音端点检测的数据集,该数据

集并无标注文件。实验从中随机抽取了 5 个人的数据文件,将所抽取的文件合并形成一个大约 300 句话、时长为30 min的测试文件,然后对测试文件进行手工标注。

为了测试不同噪声及信噪比条件下的效果,实验中加入 4 种信噪比(5 dB、10 dB、15 dB和20 dB)条件下 NOISEX-92 库中的 4 种噪声(white、volvo、factory 和 babble)。

### 3.2 实验配置

测试文件按照20 ms分帧,帧叠为10 ms,FFT 阶数 $NFFT = 512$,LTSD 窗长 $N = 6$,初始化噪声长度为文件起始的200 ms即文件开始的前 20 帧数据,基音比例门限 $\theta_v = 0.37$(该值为经验值),为了避免噪声门限更新过于迅速,噪声更新惯性系数为 $\alpha = 0.6$,最短句长 $L_{\text{MinSent}} = 20$ 帧,最短噪声长度 $L_{\text{MinNoise}} = 20$ 帧,这两个参数是根据人日常说话时的一般情况来设置。

### 3.3 评价标准及测试结果

在评价标准中采用非语音帧正确率(HR0)和语音帧正确率(HR1)以及总帧正确率(HR)。定义 $N_{00}$表示非语音帧判为语音帧的数量,$N_0$表示非语音帧的总数,$N_{11}$表示语音帧判为语音帧的数量,$N_1$表示语音帧的总数量,则评价标准可以通过以下公式计算:

$$HR0 = \frac{N_{00}}{N_0} \tag{9}$$

$$HR1 = \frac{N_{11}}{N_1} \tag{10}$$

$$HR = \frac{N_{00} + N_{11}}{N_0 + N_1} \tag{11}$$

实验结果在表 1 中列出。

**表 1 实验结果**

Table 1 Experiment result

| 噪声 | 指标 | 信噪比/dB | | | |
|---|---|---|---|---|---|
| | | 20 | 15 | 10 | 5 |
| white | *HR*0 | 97.15 | 97.62 | 98.21 | 98.99 |
| | *HR*1 | 98.66 | 97.20 | 96.33 | 93.12 |
| | *HR* | 97.57 | 97.50 | 97.68 | 97.34 |
| volvo | *HR*0 | 96.30 | 96.25 | 96.58 | 97.08 |
| | *HR*1 | 99.72 | 99.76 | 99.26 | 98.91 |
| | *HR* | 97.26 | 97.23 | 97.33 | 97.60 |
| babble | *HR*0 | 95.75 | 90.65 | 88.96 | 82.74 |
| | *HR*1 | 96.58 | 91.76 | 81.95 | 74.43 |
| | *HR* | 95.98 | 90.97 | 86.99 | 80.41 |
| factory | *HR*0 | 91.52 | 93.56 | 94.18 | 92.95 |
| | *HR*1 | 97.32 | 94.78 | 88.58 | 72.77 |
| | *HR* | 93.14 | 93.90 | 92.61 | 87.29 |

### 3.4 实验结果分析

从实验结果中可以看出,在 volvo 噪声和 white 噪声条件下,算法的效果非常好,尤其是 volvo 噪声条件下,在各种信噪比条件下都能获得较高的帧判别正确率,究其原因应该是 volvo 噪声频谱能量相对比较集中且变化平稳,LTSD 特征能很好地将其与语音区分开来。在 white 噪声条件下,相比 volvo 噪声条件下,语音帧命中率略有降低,但仍然能获得较高的总帧正确率。

在 factory 噪声与 babble 噪声中,语音帧的检测正确率有明显降低,而非语音帧检测正确率较高。经观察这两种噪声的频谱发现,其频谱与语音的频谱比较接近,干扰较大,使得仅使用 LTSD 特征检测出的语音混入大量噪声,同时由于引入了基音比例的检测,使得部分含有较高噪声的语音片段被舍弃,故语音帧检测率降低。

上述的缺憾在说话人识别的应用中存在有利一面。更为严格的语音检测避免了混有大量噪声的脏数据参与到训练与识别中,这一特点有利于提高说话人识别系统的性能。

为验证算法对于说话人识别系统的影响,后续进行了说话人识别的 GMM-UBM 系统[8]实验。首先,在系统的前端使用了原始的基于短时能量的语音检测算法,说话人识别系统的等错误率为28.69%。在不改变后期处理步骤,仅仅更换了语音检测算法的条件下,使系统的等错误率降至了24.27%,新的语音检测算法使说话人识别系统性能提高了 15.41%。

## 4 结束语

语音检测作为说话人识别系统的前端具有重要的作用。本文提出的语音检测算法在发挥了 LTSD 特征的优点同时,利用了语音中基音比例的特点,有效减少了突发噪声的影响。实验证明,该方法在多种信噪比条件下能取得很好的检测效果。在说话人识别系统中,利用该算法可以有效提高说话人识别系统的性能。但是,算法目前还存在一些不足:在嘈杂的背景噪声中,由于基音频率特征的引入会导致语音检测率下降,该问题是目前算法需要完善的一个方面。

**参考文献:**

[1] Jongseo S, Kim N, Sung W. A statistical model-based voice activity detection[J]. IEEE Signal Processing Letters, 1999,

6(1):1 – 3.

[2] 陈振标,徐波.基于子带能量特征的最优化语音端点检测算法研究[J].声学学报,2005,30(2):171 – 176.
CHEN Zhen-biao, XU Bo. Research on Speech Endpoint Detection Algorithm Based on Sub-band Energy[J]. Acta Acustica, 2005, 30(2): 171 – 176. (in Chinese)

[3] 国雁萌,付强,颜永红.复杂噪声环境中的语音端点检测[J].声学学报,2006,31(6):549 – 554.
GUO Yan-meng, FU Qiang, YAN Yong-hong. Speech endpoint detection in real noise environments[J]. Acta Acustica, 2006, 31(6): 549 – 554. (in Chinese)

[4] Ramirez J, Segura J, Benitez C, et al. Efficient voice activity detection algorithms using long-term speech information [J]. Speech Communication, 2004, 42(3): 271 – 287.

[5] Kumar P, 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.

[6] 李晔, 张仁智, 崔慧娟, 等. 低信噪比下基于谱熵的语音端点检测算法[J]. 清华大学学报(自然科学版), 2005, 45(10): 1397 – 1400.
LI Ye, ZHANG Ren-zhi, CUI Hui-juan, et al. Voice activity detection algorithm with low signal-to-noise ratios based on the spectrum entropy[J]. Journal of Tsinghua University (Science and Technology), 2005, 45(10): 1397 – 1400. (in Chinese)

[7] 杨行竣, 迟惠生. 语音信号数字处理[M]. 北京: 电子工业出版牡, 1995: 318.
YANG Xing-jun, CHI Hui-sheng. A Real time Digital Speech Signal Processing System[M]. Beijing: Publishing House of Electronics Industry, 1995: 318. (in Chinese)

[8] Douglas R A, Rose R C. Robust text-independent speaker identification using Gaussian mixture speaker models [J]. IEEE Transactions on Audio, Speech and Language Processing, 1995(3): 72 – 83.

## 作者简介:

**孟一鸣**(1983—),男,河南开封人,硕士研究生,主要研究方向为语音信号处理、说话人识别;

MENG Yi-ming was born in Kaifeng, Henan Province, in 1983. He is now a graduate student. His research concerns speech processing and speaker identification.

Email: camelmengnudt@163.com

**欧智坚**(1975—),男,湖南衡阳人,副教授、博士生导师,主要研究方向为语音识别与机器智能。

OU Zhi-jian was born in Hengyang, Hunan Province, in 1975. He is now an associate professor and also the Ph.D. supervisor. His research concerns speech processing and machine intelligence.