基于 Fisher 线性判别分析的语音信号端点检测方法

王明合 张二华* 唐振民 许 昊
(南京理工大学计算机科学与工程学院 南京 210094)

摘 要：传统的语音端点检测方法对辅音，特别是受到噪声污染的清音部分与背景噪声之间分离能力不足。针对上述问题，该文提出一种基于 Fisher 线性判别分析的梅尔频率倒谱系数(F-MFCC)端点检测方法。将清音信号和背景噪声视为两类分类问题，采用 Fisher 准则求解具有判别信息的最佳投影方向，使得投影后的特征参数具有最小类内散度和最大类间散度，从而增大清音与背景噪声的可分离性。在不同语音库上的实验结果表明，F-MFCC 能够在不同信噪比和背景噪声条件下提高语音端点检测的准确率。

关键词：语音处理；语音端点检测；梅尔频率倒谱系数；Fisher 线性判别分析
中图分类号： TN912.34 **文献标识码**： A **文章编号**：1009-5896(2015)06-1343-07
DOI:10.11999/JEIT141122

Voice Activity Detection Based on Fisher Linear Discriminant Analysis

Wang Ming-he Zhang Er-hua Tang Zhen-min Xu Hao
(*School of Computer Science and Engineering*, *Nanjing University of Science and Technology*, *Nanjing* 210094, *China*)

Abstract: Traditional Voice Activity Detection (VAD) approaches can not effectively detect consonant as well as noisy unvoiced consonant. To address this problem, this paper proposes a VAD approach Mel Frequency Cepstrum Coefficient (F-MFCC) based on Fisher linear discriminant analysis, in consideration of two-class issue regarding to consonant and background noise. Fisher criterion rule is used to solve the optimal projection vector, building upon which we can minimize the within-class scatter can be minimized and the between-class scatter can be maximized, as a result to enhance separability between consonant and background noise. Extensive experiments are conducted to evaluate the F-MFCC performance. The results demonstrate that, under different SNR and noise conditions, the proposed approach achieves higher VAD accuracy.

Key words: Speech processing; Voice Activity Detection (VAD); Mel Frequency Cepstrum Coefficient (MFCC); Fisher linear discriminant analysis

1 引言

语音端点检测(Voice Activity Detection, VAD)是指用来检测语音信号中语音起始点和结束点的技术，目的是把有声段和无声段分开。该技术广泛应用于语音识别、说话人识别、语音编码、信道传输及语音信号减噪等相关领域。研究表明，即使在安静环境下，语音识别系统大部分的错误是由端点检测精度不足造成的[1]。VAD 是语音信号处理中最基本的，但又极为关键的环节，仍然是当前研究的热点之一。早期阶段，其主要采用语音的短时能量和过零率相结合的双门限法进行检测，在纯净语音状况下具有良好的性能。然而，在真实环境下，采集的语音信号大多伴有各种各样的噪声，使得检测性能大幅下降，进而会降低语音自动识别系统的准确性以及语音通信系统重构语音信号的质量。

针对噪声干扰，研究人员提出了大量的 VAD 方法，从不同的角度可以分为多种类型。从所提取的特征参数来看，有基于短时能量及过零率、子带信噪比[2]、自相关函数、声道共振峰[3,4]、谱熵[5]、小波分解系数[6]、线性预测倒谱系数残差及高阶统计量[7]、梅尔频率倒谱系数(Mel Frequency Cepstrum Coefficient, MFCC)[8]、ERB 特征[9]、希尔伯特-黄变换特征[10]、稀疏表示[11]和多种特征相结合[12]等方法；从判决距离来分有基于欧式距离、明式距离、余弦夹角距离、相关系数距离[8]等方法；从机器学习的角度可分为有监督学习、无监督聚类和半监督学习方法。近年来还有研究者提出了基于多麦克风[13]和深度神经网络(deep neural networks)[14]等方法。上述方法通过噪声特性估计，虽然在一定程度上提高了 VAD 的鲁棒性，但对于一些受到噪声污染的辅音信号，特别是和噪声特征较为接近的清音部分，分离能力明显不足。

Fisher 线性判别分析 (Fisher linear

2014-08-29 收到，2014-12-19 改回
*通信作者：张二华 speechstudio@163.com

discri-minant analysis)[15]作为模式识别领域最具影响的算法之一，广范应用于人脸识别、医学图像分类、语音识别等系统。本文将 VAD 看作两类分类问题，提出基于 Fisher 线性判别分析的 F-MFCC 端点检测方法。在语音库上事先选取部分清音信号作为清音样本集，把待检测语音的前几帧作为背景噪声样本集，通过 Fisher 准则求解 MFCC 具有判别信息的最佳投影方向。特征参数经投影后，增强了清音和背景噪声之间的区分能力，使得清音分离能力大幅提高，从而 F-MFCC 端点检测方法的整体准确度得到提升。在增强清音分离能力的同时，浊音分离能力依然保持良好，只有极少部分受到了一些影响，可以通过和短时能量参数相结合来弥补。求得投影向量后，对每帧 MFCC 特征参数直接投影降维至 1 维标量，根据阈值判决该帧是否为有声段。

本文结构安排为：第 2 节介绍基于 MFCC 相似度方法；第 3 节提出 F-MFCC 算法并进行理论分析；第 4 节在不同信噪比和背景噪声条件下进行实验仿真和性能评价；第 5 节总结全文。

2 基于 MFCC 相似度方法

MFCC 是最常用的声学特征之一。由耳蜗的生理构造决定，人耳对不同频率的声音信号具有不同的感知能力，在频域上呈现非线性关系。MFCC 就是根据这种现象提出的特征参数。首先对语音信号预加重、分帧、加窗处理，然后对每帧进行离散傅里叶变换，得到在频率域上的能量分布。根据人耳特性设置一组三角滤波器组，计算每个滤波器输出的能量的对数，再经过离散余弦变换，得到一组系数 $c(i)$，即 MFCC。在实际应用中，通常保留前 12 维，$1 \le i \le 12$。将 MFCC 的向量形式记作 $\boldsymbol{c}_m$，其中 m 为帧序列号。选取相关系数作为相似度测度，根据式(1)计算 $\boldsymbol{c}_m$和 $\boldsymbol{b}$ 的 MFCC 相似度距离 $d(\boldsymbol{c}_m, \boldsymbol{b})$，并参照短时能量法，选取合适的阈值来判决该帧是有声段，还是背景噪声段。

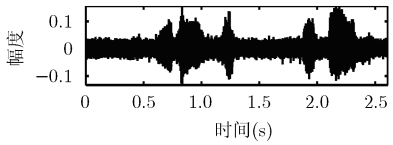
$$d(\boldsymbol{c}_m,\boldsymbol{b}) = 1 - \frac{(\boldsymbol{c}_m - \bar{c}_m\boldsymbol{I})(\boldsymbol{b}-\bar{b}\boldsymbol{I})^{\mathrm{T}}}{\left[(\boldsymbol{c}_m - \bar{c}_m\boldsymbol{I})(\boldsymbol{c}_m - \bar{c}_m\boldsymbol{I})^{\mathrm{T}}\right]^{1/2}\left[(\boldsymbol{b}-\bar{b}\boldsymbol{I})(\boldsymbol{b}-\bar{b}\boldsymbol{I})^{\mathrm{T}}\right]^{1/2}} \tag{1}$$

其中，$\bar{c}_m = \dfrac{1}{L}\sum_{i=1}^{L} c_m(i)$, $\bar{b} = \dfrac{1}{L}\sum_{i=1}^{L} b(i)$，$L$ 为 12，$\boldsymbol{I}$ 为单位向量，$\boldsymbol{c}_m$为当前帧的 MFCC 向量，$\boldsymbol{b}$ 为背景噪声的 MFCC 均值。将前 10 帧视为背景噪声，将其 $\boldsymbol{c}_n$ 的平均值作为 $\boldsymbol{b}$ 的初始值，$1 \le n \le 10$。考虑到背景噪声可能随时间变化，对背景噪声进行自适应更新。

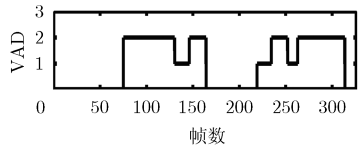
3 基于 Fisher 线性判别分析的 VAD

虽然传统 VAD 能够降低噪声对端点检测的影响，但是对受到噪声污染的辅音以及和噪声特征较为接近的清音部分分离能力明显不足。VAD 所采用的语音特征主要有能量、过零率、信噪比、MFCC 等，下面分别从这几个方面来分析清音信号分离能力弱的原因。图 1(a)为加入白色噪声的含噪语音，信噪比为 0 dB；图 1(b)为人工端点标注，用 0 表示背景噪声段，1 表示清音段，2 表示浊音段；图 1(c)为含噪语音的频谱图。相对于浊音，清音的能量本来就较低，且多数噪声和清音的过零率同样较高，显然，在强噪声背景下，很难从能量和过零率上把清音和噪声区分开来。图 1(a)中含噪语音的平均信噪比为 0 dB，在部分元音段，信噪比峰值高达 27 dB，而在部分清音段，信噪比则低至 −9 dB。因此对背景噪声的估计和自适应更新中产生的误差偏移很容易导致基于信噪比阈值的端点检测产生错误；从图 1(c)中可以看出，受噪声污染的清音信号和背景噪声的频谱极为相似，这导致基于相似度距离的 VAD 很难实现清音和背景噪声的有效分离。

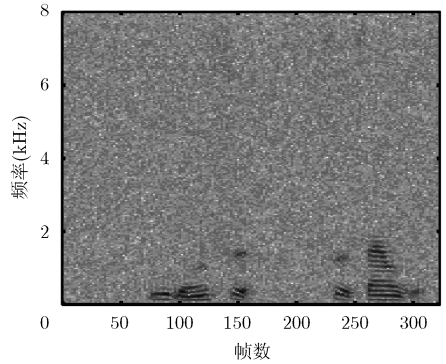
从发声原理角度分析，清音可以被认为是通过声门的气流噪声经过声道的滤波产生的，和自然生



(a)加噪语音信号波形



(b)人工端点标注



(c)加噪语音信号频谱图

图 1 加噪语音信号

成的各类背景噪声不尽相同。因此，可以把背景噪声和清音看作两类分类问题，通过将 Fisher 线性判别分析[15]引入 VAD，增大清音与背景噪声的类间散度和减小类内散度，以此来提高两者的判别能力。其基本思想是将高维的特征参数投影降维到最佳判别矢量空间，投影后保证模式样本在新的子空间类内紧凑和类间分离(即最小的类内散度和最大的类间散度)，模式在该空间中有最佳的可分离性。VAD 属于两类分类，可以投影降维到 1 维空间，在此基础上可选取合适的阈值区分有声段和背景噪声段。不同人之间，甚至男女之间清音的 MFCC 差别很小，因此，我们在已有纯净语音库中随机选取清音段组合成一个约 3 s 的清音样本集，预加重、分帧、加窗后，提取出 N_1 帧 MFCC 参数，记作 $\boldsymbol{Q}_k, 1 \le k \le N_1$。取待检测语音信号的前 N_2 帧，作为背景噪声样本集，同样处理得到 N_2 帧 MFCC 参数，记作 $\boldsymbol{G}_k, 1 \le k \le N_2$，通常 N_2 取值 10。清音样本集、背景噪声样本集以及二者合并后样本集的均值向量分别记作 $\boldsymbol{u}_1$,$\boldsymbol{u}_2$,$\boldsymbol{u}_0$，根据式(2)计算。

$$\left.\begin{aligned}\boldsymbol{u}_1 &= \frac{1}{N_1}\sum_{k=1}^{N_1}\boldsymbol{Q}_k \\ \boldsymbol{u}_2 &= \frac{1}{N_2}\sum_{k=1}^{N_2}\boldsymbol{G}_k \\ \boldsymbol{u}_0 &= \frac{1}{N_1+N_2}(N_1\boldsymbol{u}_1 + N_2\boldsymbol{u}_2)\end{aligned}\right\} \tag{2}$$

给定投影向量 $\boldsymbol{w}$，取维数 12，则投影后的类间散度为

$$\begin{aligned}\mathrm{SS}_B &= N_1\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{u}_1 - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{u}_0\right)^2 + N_2\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{u}_2 - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{u}_0\right)^2 \\ &= \boldsymbol{w}^{\mathrm{T}}\Big[N_1(\boldsymbol{u}_1-\boldsymbol{u}_0)(\boldsymbol{u}_1-\boldsymbol{u}_0)^{\mathrm{T}} \\ &\quad + N_2(\boldsymbol{u}_2-\boldsymbol{u}_0)(\boldsymbol{u}_2-\boldsymbol{u}_0)^{\mathrm{T}}\Big]\boldsymbol{w}\end{aligned} \tag{3}$$

类内散度为

$$\begin{aligned}\mathrm{SS}_W &= \sum_{k=1}^{N_1}\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{Q}_k - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{u}_1\right)^2 + \sum_{k=1}^{N_2}\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{G}_k - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{u}_2\right)^2 \\ &= \boldsymbol{w}^{\mathrm{T}}\left[\sum_{k=1}^{N_1}(\boldsymbol{Q}_k-\boldsymbol{u}_1)(\boldsymbol{Q}_k-\boldsymbol{u}_1)^{\mathrm{T}}\right. \\ &\quad \left. + \sum_{k=1}^{N_2}(\boldsymbol{G}_k-\boldsymbol{u}_2)(\boldsymbol{G}_k-\boldsymbol{u}_2)^{\mathrm{T}}\right]\boldsymbol{w}\end{aligned} \tag{4}$$

令

$$\begin{aligned}\boldsymbol{S}_B &= N_1(\boldsymbol{u}_1-\boldsymbol{u}_0)(\boldsymbol{u}_1-\boldsymbol{u}_0)^{\mathrm{T}} \\ &\quad + N_2(\boldsymbol{u}_2-\boldsymbol{u}_0)(\boldsymbol{u}_2-\boldsymbol{u}_0)^{\mathrm{T}} \\ \boldsymbol{S}_W &= \sum_{k=1}^{N_1}(\boldsymbol{Q}_k-\boldsymbol{u}_1)(\boldsymbol{Q}_k-\boldsymbol{u}_1)^{\mathrm{T}} \\ &\quad + \sum_{k=1}^{N_2}(\boldsymbol{G}_k-\boldsymbol{u}_2)(\boldsymbol{G}_k-\boldsymbol{u}_2)^{\mathrm{T}}\end{aligned}$$

Fisher 鉴别准则表达式为

$$\max J_{\mathrm{fisher}}(\boldsymbol{w}) = \frac{\mathrm{SS}_B}{\mathrm{SS}_W} = \frac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_B\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_W\boldsymbol{w}} \tag{5}$$

其中，$\boldsymbol{S}_B$,$\boldsymbol{S}_W$ 均为对称半正定矩阵，$(\boldsymbol{S}_W)^{1/2} = (\boldsymbol{S}_W^{\mathrm{T}})^{1/2}$，且 $\boldsymbol{S}_W = (\boldsymbol{S}_W)^{1/2}(\boldsymbol{S}_W)^{1/2}$。

令 $\boldsymbol{v} = (\boldsymbol{S}_W)^{1/2}\boldsymbol{w}$，则 $\boldsymbol{w} = (\boldsymbol{S}_W)^{-1/2}\boldsymbol{v}$，代入式(5)得

$$\max J_{\mathrm{fisher}}(\boldsymbol{w}) = \frac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_B\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_W\boldsymbol{w}} = \frac{\boldsymbol{v}^{\mathrm{T}}\left(\boldsymbol{S}_W^{\mathrm{T}}\right)^{-1/2}\boldsymbol{S}_B\left(\boldsymbol{S}_W\right)^{-1/2}\boldsymbol{v}}{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}} \tag{6}$$

等价于求最大特征值 $\lambda_{\max}[(\boldsymbol{S}_W^{\mathrm{T}})^{-1/2}\boldsymbol{S}_B(\boldsymbol{S}_W)^{-1/2}] = \lambda_{\max}[(\boldsymbol{S}_W)^{-1}\boldsymbol{S}_B]$ 对应的特征向量，即

$$\begin{aligned}\lambda_{\max}\boldsymbol{w} &= (\boldsymbol{S}_W)^{-1}\boldsymbol{S}_B\boldsymbol{w} \\ &= (\boldsymbol{S}_W)^{-1}(\boldsymbol{u}_1-\boldsymbol{u}_2)(\boldsymbol{u}_1-\boldsymbol{u}_2)^{\mathrm{T}}\boldsymbol{w} \\ &= (\boldsymbol{S}_W)^{-1}(\boldsymbol{u}_1-\boldsymbol{u}_2)\left(\boldsymbol{u}_1^{\mathrm{T}}\boldsymbol{w} - \boldsymbol{u}_2^{\mathrm{T}}\boldsymbol{w}\right)\end{aligned} \tag{7}$$

其中，$\lambda_{\max}$ 及 $(\boldsymbol{u}_1^{\mathrm{T}}\boldsymbol{w} - \boldsymbol{u}_2^{\mathrm{T}}\boldsymbol{w})$ 为标量，$\boldsymbol{w}$ 与 $(\boldsymbol{S}_W)^{-1}\cdot(\boldsymbol{u}_1-\boldsymbol{u}_2)$ 同方向，若忽略系数，则最佳投影方向 $\boldsymbol{w}$ 为

$$\boldsymbol{w} = (\boldsymbol{S}_W)^{-1}(\boldsymbol{u}_1-\boldsymbol{u}_2) \tag{8}$$

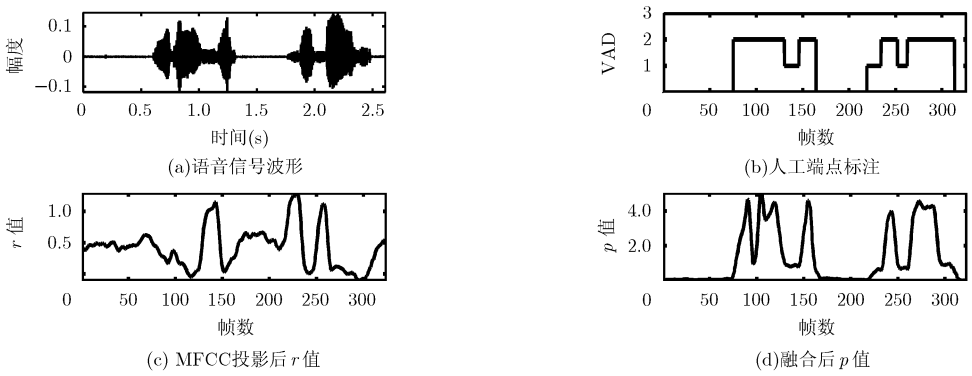
将待检测的语音信号提取出每帧的梅尔倒谱系数 $\boldsymbol{c}_m$，其中 m 为帧序列号。根据式(9)，将投影降维到 1 维后的参数记作 r_m。

$$r_m = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{c}_m \tag{9}$$

图 2(a)为在安静环境下录制的一段语音对应的波形。考虑到在无声段也有录音设备的本底噪声存在，绝对纯净的语音信号现实世界中是不存在的。不失一般性，本文将所有非语音信号视为噪声，亦即将所有无声段均视为背景噪声段。图 2(b)为人工标注，其中，0 表示背景噪声段，1 表示清音段，2 表示浊音段。通过观察图 2(c)中 MFCC 的投影值曲线可知，清音段和背景噪声段的可分离性显著提高，浊音段和背景噪声段的可分离性保持良好。将语音信号的短时能量值记作 e_m，根据式(10)和 r_m 进行融合后，记作 p_m。

$$p_m = |r_m - R| + \alpha e_m \tag{10}$$

其中，R 为背景噪声 r 值的估计。将待检测语音信号前 N_2 帧 r_m 的平均作为 R 的初始值，α 为权重系数。设 E 为 e_m 前 N_2 帧的平均值。若 E 小于 τ，则令 $E = \tau$。设 $\alpha = a/E$，其中 a 和 τ 为常数，分别取值 0.1 和 0.05。当背景噪声较小时，e_m 在 p_m 中权重较大，可有效避免将说话过程中和清音特征相近的换气、呼吸等噪声误检测为有声段；当背景噪声较大时，e_m 在 p_m 中权重变小，$|r_m - R|$ 的权重变大，可减弱背景噪声的能量对端点检测的干扰。在检测过程中，如果第 j 帧对应的信号被判决为背景噪声，



(a)语音信号波形 (b)人工端点标注

(c) MFCC投影后 r 值 (d)融合后 p 值

图 2 语音信号 MFCC 投影后的参数

则根据式(11)对 R 进行更新，以自适应地跟踪背景噪声的变化。

$$R=(1-\beta)R+\beta r_j \tag{11}$$

式中，$\beta\in[0,1]$ 决定 R 自适应背景噪声的速度，本文取固定值 0.01。

图 2(d)给出了参数 p_m 对应的曲线。对于背景噪声段，$|r_m-R|$ 和 e_m 均较小，故 p_m 值较小；对清音段，$|r_m-R|$ 较大，但 e_m 较小，故 p_m 值中等；对浊音段，$|r_m-R|$ 和 e_m 均较大，故 p_m 值较大。因此，将参数 p_m 作为最终 VAD 的判决依据，在保证浊音有效检出的情况下，清音的分离能力明显增强。

4 性能分析与评价

4.1 实验环境

仿真实验所采用的语音信号选自 NUST603_2014 及 TIMIT 语音库，混叠的噪声选自 NOISEX-92 噪声库。NUST603_2014 语音库由南京理工大学“高维信息智能感知与系统”教育部重点实验室录制完成，包含男 210 人，女 213 人，是在日常办公室环境下，分别通过麦克风、固定电话、手机 3 种传输信道录制的，并混有真实自然的背景噪声。TIMIT 语音库由 Texas Instruments 和 Massa-chusetts Institute of Technology 联合录制完成，包含男 438 人，女 192 人，是在安静环境及高质量麦克风条件下录制的连续语音。

实验在联想 PC 机(CPU:E7500, 2.93 GHz)上进行，操作系统采用 Windows XP，在 MATLAB R2011a 环境下执行 F-MFCC。在不同语种和噪声条件下，以人工标注为标准，重点考察如下 3 方面的性能指标：

(1)清音分离能力：指将清音信号和背景噪声进行区分的能力。

(2)整体检测准确率：指端点检测正确的帧数在被测试语音信号总帧数中所占据的比例。

(3)实时性能：统计 F-MFCC 的执行时间，以此衡量实时性能。

在后续结果分析中，我们将(1)对比 F-MFCC 和 MFCC 相似度方法的清音分离能力；(2)对比 AMR-1[2]，基于 MFCC 相似度方法和 F-MFCC 的整体准确率；(3)分析 F-MFCC 的实时检测性能。

4.2 结果分析

在 NUST603_2014 语音库麦克风目录下，随机选取清音片段，组成约 3 s 的清音样本集。将待检测语音信号的前 N_2 帧作为背景噪声的样本集。利用 Fisher 线性判别分析找到最佳投影向量，将语音信号提取出 MFCC，逐帧投影降维，并和能量参数融合后，作为 VAD 的判决参数。

图 3~图 6 对选自 NUST603_2014 语音库中的麦克风语音、固定电话语音、TIMIT 语音库纯净语音、NUST603_2014 语音库的麦克风语音混入白色噪声后(SNR=0 dB)的部分语音信号清音分离能力进行了对比，分别用 0 和 1 表示背景噪声段和有声语音段(包括清音段和浊音段)。在人工标注过程中，分别用 0,1 和 2 表示背景噪声段、清音段和浊音段。由图可知，在不同的语种和信噪比条件下，F-MFCC 在清音分离能力方面都明显超过了传统方法中具有代表性的基于 MFCC 相似度距离检测方法。图 3 和图 5 中语音信号的背景噪声虽然较小，但噪声类型主要是录音设备的电路噪声和说话人的呼吸、换气噪声。通过适当扩大参数 α，F-MFCC 可以有效降低此类噪声对 VAD 的影响，所以清音分离性能明显优于 MFCC 相似度方法。图 4 中语音信号采集自固定电话，并伴有随时间波动的周期性环境噪声，该环境下 F-MFCC 的清音分离性能略优于 MFCC 相似度方法。图 6 中语音信号为 NUST603_2014 语音库的麦克风语音混入白色噪声，背景噪声几乎将清音信号完全淹没。通过适当减小参数 α，调节 p_m 中 $|r_m-R|$ 的权重，提高清音分离能力。

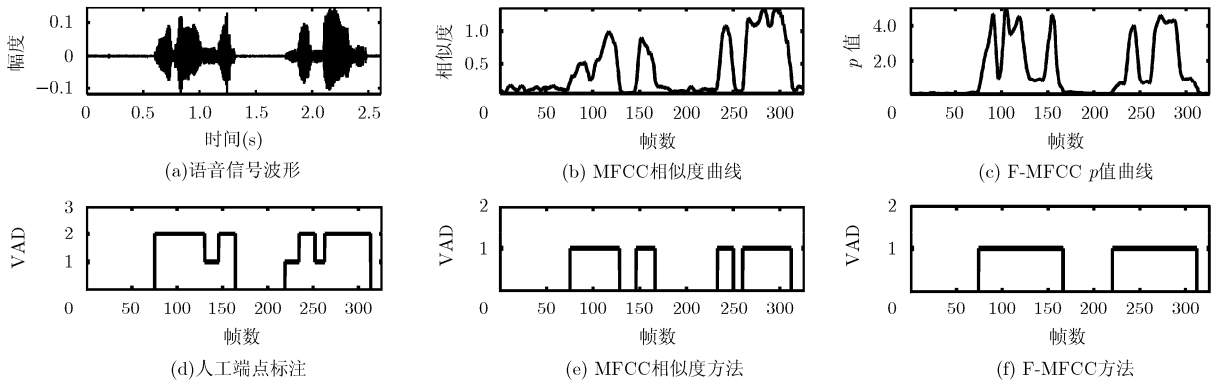


图 3 清音分离能力对比(麦克风语音)

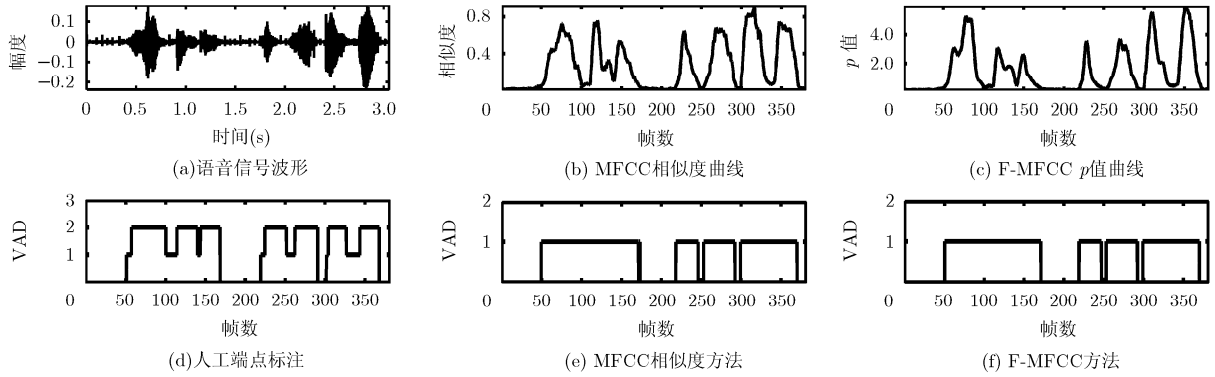


图 4 清音分离能力对比(带噪电话语音)

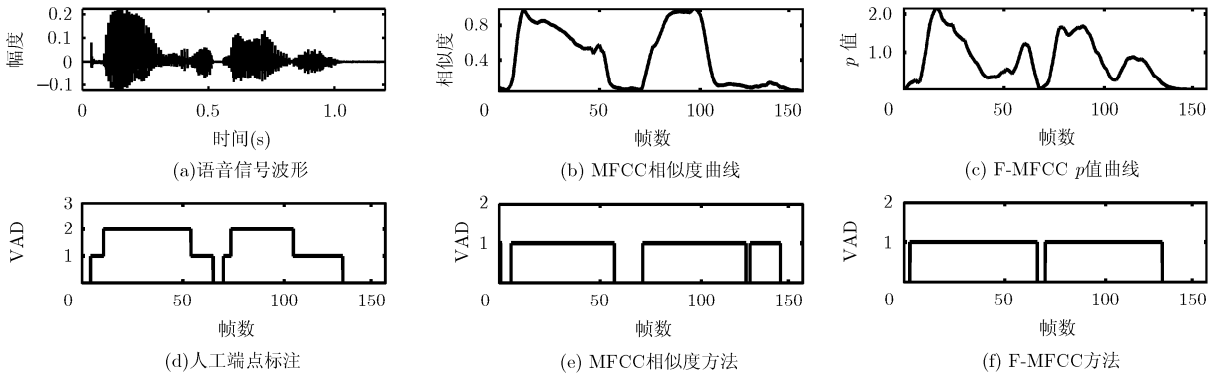


图 5 清音分离能力对比(TIMIT 语音库语音)

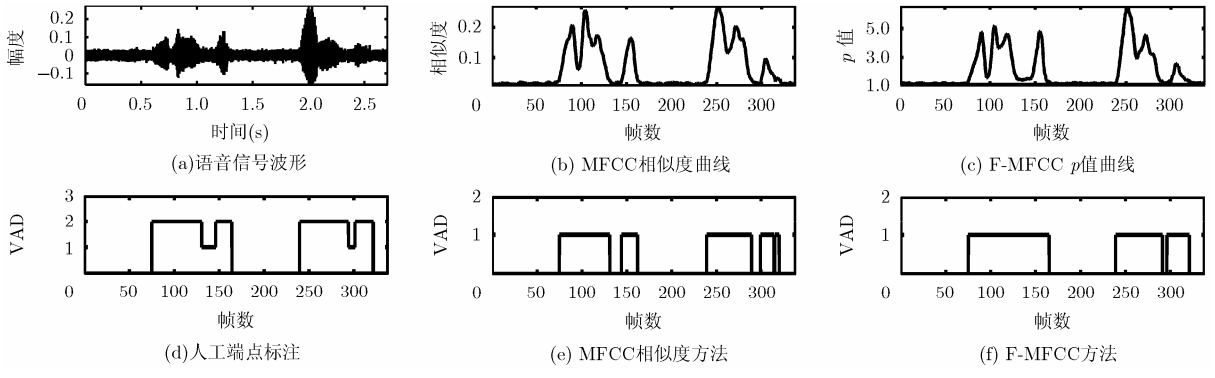


图 6 清音分离能力对比(麦克风语音混入白色噪声)

下面对 AMR-1，基于 MFCC 相似度方法和 F-MFCC 的准确率进行结果分析。受噪声污染清音信号的误判是导致传统 VAD 错误的重要原因之一。清音的分离能力改善了，VAD 的整体准确率自然会得到提高。在不同语音库和信噪比条件下，F-MFCC, AMR-1 和基于 MFCC 相似度方法的整体准确率对比如表 1 所示。其中 NUST603_2014 语音库的麦克风语音、带噪电话语音、TIMIT 语音库纯净语音的信噪比由 NIST STNR Tools(V2.7)评估得出。

由表 1 可知， F-MFCC 端点检测方法在不同条件下的整体准确率均明显高于其它两种方法。目前所广泛使用的 AMR 并不精确[16]，原因是该方法在检测到有声段时，为保证经编码和传输后语音的可懂度，将有声段分别向前、向后延展几帧，降低了有声段的漏检率，但明显增加了将无声段误检测成有声段的虚警率。高精度的 VAD 可进一步提高多速率语音编码的压缩率，并降低对传输信道的带宽要求。在对 TIMIT 语音库纯净语音的试验中，基于 MFCC 相似度的方法准确率只有 75.2%。这是因为相比汉语普通话，清音在英语中所占的比例明显更多，所以清音分离能力对 VAD 整体准确率的影响更大。为了提高实验结果的参考价值，这里考虑了两种检测方案，分别命名为 F-MFCC(I)和 F-MFCC(II)。前者的清音样本取自 NUST603_2014 语音库麦克风目录下的汉语普通话语音信号；后者的清音样本取自 TIMIT 语音库英语语音信号。根据表中的数据可知，在 5 种情况下的 VAD 准确度，F-MFCC(II)均优于 F-MFCC(I)，这是源于英语和汉语普通话的语言结构、发音方式等存在差异。取自 TIMIT 语音库英语语音信号的清音样本集音素更丰富，代表性更强。实验结果表明：相对于基线方法 AMR-1 和基于 MFCC 相似度的 VAD，在所有 5 种测试条件下，F-MFCC 端点检测方法获得了相对更高的整体准确率。

我们统计了 NUST603_2014 语音库中语音端点检测所需的时间,以此评价 F-MFCC 的实时检测性能。为了提高实验结果的参考价值，我们从语音库随机选取 400 段(约 6h)语音进行以上的实验。根据实验统计数据，每 60 s 语音信号的端点检测平均执行时间为 1.211 s，表明 F-MFCC 可以满足实时性要求。

5 结论

本文在 Fisher 线性判别分析的基础上，提出了 F-MFCC 端点检测方法。首先，用 Fisher 准则求解具有判别信息的最佳投影方向，目的是增大噪声和清音间的可分离性。然后，把 MFCC 作为语音信号的特征参数，并将其投影值和短时能量相结合，增强了对易受噪声污染的清音信号的分离能力，提高了端点检测的整体准确率。实验结果表明，该方法在不同语种、环境噪声和信噪比条件下，端点检测的清音分离能力、整体准确率始终优于目前具有代表性的 AMR-1 和 MFCC 相似度方法。

表 1 检测准确率对比(%)

方法	NUST603_2014 麦克风语音 SNR=42 dB	NUST603_2014 带噪电话语音 SNR=40 dB	TIMIT 语音库纯净语音 SNR=53 dB	NUST603_2014 麦克风语音混入白色噪声 SNR=0 dB	TIMIT 语音库语音混入工厂噪声 SNR=0 dB
AMR-1	88.5	70.9	89.9	78.2	77.6
MFCC 相似度	83.0	88.7	75.2	83.8	62.7
F-MFCC(I)	98.1	92.4	96.7	92.9	**85.3**
F-MFCC(II)	**98.3**	**93.3**	**97.0**	**96.8**	**85.3**

参 考 文 献

[1] Junqua J C. Robustness and cooperative multi-model man-machine communication applications[C]. The Structure of Multimodal Dialogue, Maratea, Italy, 1991: 101–112.

[2] ETSI. Universal Mobile Telecommunication Systems (UMTS); Mandatory Speech Codec speech processing functions, AMR speech codec; Voice Activity Detector VAD[S]. ETSI TS 126 094 v11.0.0(2012-10): 1–26.

[3] Wan Yu-long, Wang Xian-liang, Zhou Ruo-hua, *et al*. Enhanced voice activity detection based on automatic segmentation and event classification[J]. *Journal of Computational Information Systems*, 2014, 10(10): 4169–4177.

[4] 宫朝辉, 刁麓弘. 改进共振峰提取的语音端点检测[J]. 计算机辅助设计与图形学学报, 2013, 25(8): 1230–1236.
Gong Zhao-hui and Diao Lu-hong. Improved speech endpoint detection based on formant[J]. *Journal of Computer Aided Design & Computer Graphics*, 2013, 25(8): 1230–1236.

[5] 李晔, 张仁志, 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): 1397–1440.

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): 1397–1440.

[6] Chen Shi-huang and Wang Jhing-fa. A wavelet-based voice activity detection algorithm in noisy environments[C]. Proceedings of the 9th IEEE International Conference on Electmnics, Circuits and Systems, Dubrovnik, Croatia, 2002: 995–998.

[7] Ghosh P K, Tsiartas A, and 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] 王宏志, 徐玉超, 李美静. 基于 Mel 频率倒谱参数相似度的语音端点检测算法[J]. 吉林大学学报(工学版), 2012, 42(5): 1331–1335.

Wang Hong-zhi, Xu Yu-chao, and Li Mei-jing. Voice activity detection algorithm based on Mel-frequency cepstrum coefficient (MFCC) similarity[J]. *Journal of Jilin University* (*Engineering and Technology Edition*), 2012, 42(5): 1331–1335.

[9] Oh Sang-yeob and Chung Kyung-yong. Improvement of speech detection using ERB feature extraction[J]. *Wireless Personal Communications*, 2014, 79(4): 2439–2451.

[10] 卢志茂, 金辉, 张春祥, 等. 基于 HHT 和 OSF 的复杂环境语音端点检测[J]. 电子与信息学报, 2012, 34(1): 213–217.

Lu Zhi-mao, Jin Hui, Zhang Chun-xiang, *et al*.. Voice activity detection in complex environment based on Hilbert-Huang transform and order statistics filter[J]. *Journal of Electronics & Information Technology*, 2012, 34(1): 213–217.

[11] Deng Shi-wen and Han Ji-qing. Statistical voice activity detection based on sparse representation over learned dictionary[J]. *Digital Signal Processing*, 2013, 23(4): 1228–1232.

[12] Zhang Yan, Tang Zhen-min, Li Yan-ping, *et al*.. A hierarchical framework approach for voice activity detection and speech enhancement[J]. *The Scientific World Journal*, 2014, Vol. 2014: Article ID 723643, 8 pages.

[13] Choi Jae-hun and Chang Joon-hyuk. Dual-microphone voice activity detection technique based on two-step power level difference ratio[J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2014, 22(6): 1069–1081.

[14] Ryant N, Liberman M, and Yuan Jia-hong. Speech activity detection on YouTube using deep neural networks[C]. Interspeech: 14th Annual Conference of the International Speech Communication Association, Lyon, France, 2013: 728–731.

[15] Fisher R A. The use of multiple measures in taxonomic problems[J]. *Annals of Eugenics*, 1936, 7(2): 179–188.

[16] Mak M W and Yu H B. A study of voice activity detection techniques for NIST speaker recognition evaluations[J]. *Computer Speech & Language*, 2014, 28(1): 295–313.

王明合：　男，1970 年生，博士生，研究方向为信号处理、语音识别、说话人识别.

张二华：　男，1967 年生，副教授，主要研究方向为信号处理、语音识别、3 维数据可视化方面.

唐振民：　男，1961 年生，博士生导师，教授，主要研究方向为语音识别、图像处理、智能机器人.