基于声频特征的维吾尔语语音端点检测方法

杨雅婷[1,2]，马博[1,2]，王磊[1,2]，吐尔洪·吾司曼[1]，李晓[1]

1. 中国科学院新疆理化技术研究所，乌鲁木齐 830011；

2. 中国科学院研究生院，北京 100190

E-mail: yangyt_xj@sina.com

摘 要： 针对传统基于短时能量和短时过零率的端点检测方法中存在的对清音检测性能以及抗噪声性能较差的缺点，结合维吾尔语的声频发音特征，提出了一种计算较为简单，受噪声影响较小的语音端点检测新方法—基于声频特征的维吾尔语语音端点检测方法。实验表明，该方法能有效提高噪声环境下维吾尔语语音端点检测的性能。

关键词： 维吾尔语；声频参数；语音端点检测；语音识别

Speech Endpoint Detection Algorithm for Uyghur Based on Acoustic Frequency Feature

YANG Ya-ting[1,2]，MA Bo[1,2]，WANG Lei[1]，TURGHUN Osman[1]，LI Xiao[1]

1. Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011;

2. Graduate University of Chinese Academy of Science, Beijing 100190

E-mail: yangyt_xj@sina.com，Phn: +86-15026043990

Abstract: A major factor influencing the capability of speech recognition systems is the accuracy of endpoint detection. In order to solve the problem of the less effective detection of surd and the poor anti-noise performance in classic method, this paper combines Uyghur acoustic frequency feature which has a better detection of surd and has a good anti-noise performance to realize the endpoint detection. The experiment result proves that the method has better deletion result.

Key words: Uyghur; Acoustic Frequency; Speech Endpoint Detection; Speech Recognition

1 引言

语音信号的端点检测在语音识别中有着及其重要的作用，对语音识别的精确度影响很大。端点检测就是正确地标注出语音信号中的各种段落的始点和终点的位置，其目标是要在一段输入信号中将语音信号同其它噪声信号分离开来。一个好的端点检测方法应该具备可靠性，鲁棒性，准确简单，能够进行实时处理和不需要噪声的先验知识等特点[1]。

目前，语音端点检测可以分为两大类：基于阈值的方法和模式识别的方法。其中基于阈值的方法以其简单、快速的优点被广泛研究和使用。已有的各种基于阈值的方法均有其局限性，如短时能量法虽然较简单，但是该方法难以区分弱摩擦音与结尾时的鼻音；由于语谱的固有特征使得基于谱熵的端点检测方法能够有效地区分语音信号和噪声信号，但它对清音部分效果不太理想；基于短时过零率的方法虽然对清音的检测效果好，但是其抗噪声性能较差；同时现有方法在低信噪比情况下的检测效果较差[2]。

随着计算机应用技术的发展，民族语音文字信息处理研究工作也在深入展开。近年来，维语语音学的研究在声学特征分析、识别基元的选取、系统结构设计等方面做了大量的工作。但是，将维吾尔语的声频特征及发音规律的特殊性与语音识别关键技术结合方面还有待进一步加强。

本研究针对现有各种方法的不足，以及维语语音音节所特有的声频特征，提出基于声频特征的维吾尔语语音端点检测方法。

基金项目： 中国科学院“西部行动计划高新技术项目”（The western high technique program of Chinese Academy of Science. No. KGCX2-YW-507）；中国科学院“西部之光”项目。

作者简介： 杨雅婷（1985-），女，博士生，主要研究方向：多语种信息处理技术；马博，博士生；王磊，副研究员；吐尔洪.吾司曼，硕士；李晓，研究员，博导。

2 维吾尔语声频特征分析

维吾尔语属于阿尔泰语系突厥语族西匈语支，在语法上属于黏着语类型。它的音素、音节等发音单元具有本质发音特点。维语语音有元音 8 个、辅音 24 个[3]。由辅音和元音构成维语语音音节，每个音节必须且只能有一个元音。维语音节的三大块是:(起音) +领音+(收音)。如果用字母“V”代表元音，“C”代表辅音，维吾尔语的音节可以归纳为以下形式:

(1) V：一个元音构成的音节

(2) VC：由一个元音一个辅音构成的音节

(3) CV：由一个辅音一个元音构成的音节

(4) VCC：由一个元音和两个辅音构成的音节

(5) CVC：由一个辅音、一个元音、一个辅音构成

(6) CVCC：由一个辅音、一个元音、两个辅音构成

部分音节在语流中产生语流音变现象，常见的有同化、弱化、脱落以及元音和谐等现象。维语的发音规律和语音现象有很鲜明的特点，其元音、辅音以及语音结构的最小单位是音节。

2.1 元音

气流通过口腔，不受发音器官的阻碍而发出的音称为元音。元音字母如下:

ئە ئۇ ئۆ ئا ئو ئۈ ئې ئى

词中的后一个音节的元音要向前一个音节的元音看齐，既要同是前元音或后元音，或者同是圆唇元音或非圆唇元音，这种元音相互协调的现象就是元音和谐。发音时，声带颤动，气流较弱处于缓和状态，易于检测。

2.2 辅音

气流通过口腔，遇到发音器官的阻碍而发出的音称为辅音。辅音可分为两类:

(1) 发音时，声带不颤动，气流较强的辅音叫清辅音（十个）如:

ت س ف ھ خ چ ك پ ق ش

(2) 发音时，声带颤动，气流较弱的辅音叫浊辅音（十四个）如:

ر ۋ ي ز غ گ ب د ج ل م ن ژ

发音时，气流在发音器官的某一部分受到明显的阻碍。辅音发音的机制如鼻音 (通过鼻腔)、塞音 (气流被完全阻塞)、或是近音 (近似元音)。由于这些特点，辅音较元音难检测。

2.3 发音频谱能量分析

分别求出上述 8 个元音和 24 个辅音的频谱，对元音和辅音的频谱求平均并且归一化，得到维语元音和辅音的频谱能量分布图，如图 1 所示。由图 1 可知，维语 32 个音节的能量分布在三个区域内，分别为 300Hz～1KHz，1KHz～2.5KHz，2.5KHz～3.5KHz，由文献[3]可知，在 300Hz～1KHz 以内为第一共振峰，1KHz～2.5KHz 内为第二共振峰，2.5KHz 以上为其它共振峰。

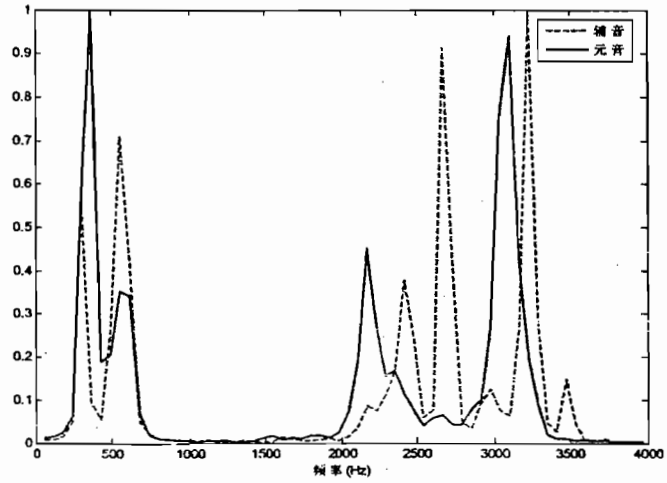

图 1 维语音节归一化频谱能量图

3 基于声频特征的维语语音端点检测方法

传统基于短时能量的方法对元音和浊辅音的识别效果较好，但是不能识别清辅音的真正起点，而基于短时过零率的方法在噪声干扰下效果并不明显[4]。维语ت的发音，基于短时能量的方法不能正确检测出真正的起点与终点，如图 2（b）所示；而基于短时过零率的方法难以区分噪声与语音，如图 2（c）所示。

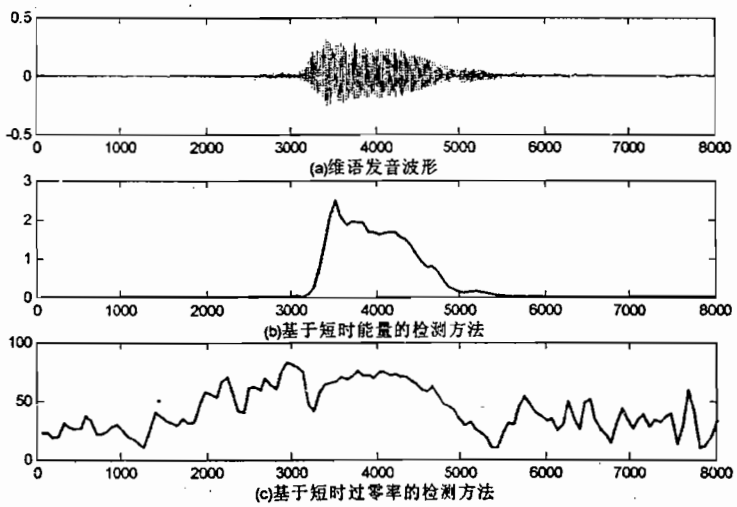

图 2 维语清辅音ت传统检测结果

通过对维语清辅音的频谱分析可知，清辅音除了在 300Hz～1KHz 频段内能量高，高频部分能量也相对较高，部分音节发音时需要先发摩擦音，例如：ت（清辅音）。发摩擦音时虽然时域整体能量低，但是高频能量较低频能量高，利用这一特性可以较好检测清辅音。

本研究通过分析三个维语音节频段的能量分布密度和前文所阐述的维吾尔语声频特征，提出以子带频谱密度为核心进行语音端点检测的方法，称之为基于声频特征的维吾尔语语音端点检测方法。

3.1 基本原理

根据上第二部分的实验与分析可知，基于声频特征的维吾尔语语音端点检测方法将语音频谱划分为三个子带，以这三个子带最大平均频谱密度作为语音帧的特征值进行端点检测。

3.1.1 频谱密度

定义 设 $\bar{x}_i = \{x(n) \mid n = 0,1...N-1\}$ 为第 i 帧语音时域信号，对其求傅里叶变换得到频谱 $\bar{X}_i = \{X(n) = \sum_{k=0}^{N-1} x(k)\exp(-j2\pi nk/N) \mid n = 0,1...N-1\}$，那么该帧频谱能量 $\bar{Y}_i$ 为：

$$\bar{Y}_i = \{Y(n) = |X(n)|^2 \mid n = 0,1...N-1\} \tag{1}$$

频谱密度 $\bar{P}_i$ 定义为：

$$\bar{P}_i = \{p(n) = \frac{Y(n)}{\sum_{k=0}^{N-1} Y(k)} \mid n = 0,1...N-1\} \tag{2}$$

3.1.2 子带平均频谱密度

定义 将整个频带分为 K 个子带，每个子带包含 N_k 个连续频率分量，那么第 k 个子带的平均频谱密度为：

$$P(k) = \frac{\sum_{i=k\times N_{k-1}}^{k\times N_{k-1}+N_k-1} p(i)}{N_k} \qquad k = 0,1...K-1; \quad N_k = 0 \quad if\ k < 0 \tag{3}$$

3.1.3 语音帧的特征值

根据 2.3 节分析，维语发音频谱可划分为 300Hz～1KHz，1KHz～2.5KHz，2.5KHz～3.5KHz 三个子频带。本文使用的语音采样率为 8KHz，语音帧长为 128 点，帧移 64 点。由此可得三个子带对应的频域点数为 $K_{low} = \{6,7...16\}, K_{mid} = \{17,18...40\}, K_{high} = \{41,42...56\}$，按照公式（3）计算这三个子带的平均频谱密度 P_{low}、P_{mid}、P_{high}。

$$\begin{cases} P_{low} = \frac{1}{11}\sum_{i=6}^{16} p(i) \\ P_{mid} = \frac{1}{24}\sum_{i=17}^{40} p(i) \\ P_{high} = \frac{1}{16}\sum_{i=41}^{56} p(i) \end{cases} \tag{4}$$

当前帧的特征值定义为

$$\lambda(i) = \max\{P_{low}, P_{mid}, P_{high}\} \tag{5}$$

3.2 端点检测方法

基于声频特征的维吾尔语语音端点检测方法，主要包含初始化、预处理、端点检测、噪声频谱更新这四个步骤[5]。这里说明我们选取元音“ئ”、浊辅音“ب”和清辅音“ت”作为实验样本验证检测结果（该实验样本具有同类代表性）。该方法对维语发音样本的检测效果如图 3 所示。

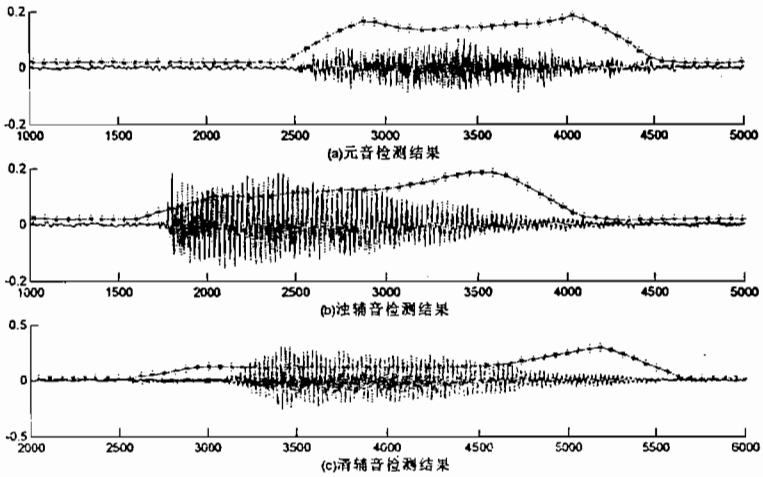

图3 本方法对维语元音“ش”、浊辅音“ب”和清辅音“ت”的检测结果

3.2.1 初始化

假设前M帧（$M=10$）为语音数据为无音片段，用来估计噪声频谱密度[6]。根据公式（1）计算每一帧的频谱能量，取其平均值作为噪声频谱能量$\bar{Y}_{noise}$，根据公式（2）计算噪声密度$\bar{P}_{noise}$，并计算平均噪声密度$P_{noise_avg}=\frac{\sum_{i=0}^{N-1}P_{noise}(i)}{N}$，则检测门限值为$TH=\alpha\times P_{noise_avg}$，本文中使用$\alpha=2$。

3.2.2 预处理

对第i帧语音数据，加hamming窗[7]。

$$W(n)=0.54-0.46\cos\frac{2\pi n}{N-1}\qquad n=0,1,2\cdots \tag{6}$$

根据公式（1）计算第i帧频谱能量$\bar{Y}$，根据公式（7）去除噪声频谱能量得到去噪后语音频谱能量$\bar{Y}_{clean}$：

$$\bar{Y}_{clean}=\bar{Y}-\bar{Y}_{noise}=\{Y(n)-Y_{noise}(n)\,|\,n=0,1\ldots N-1\}。 \tag{7}$$

利用去噪后的语音频谱能量$\bar{Y}_{cleon}$按照公式（2）计算语音信号的频谱密度$\bar{P}$，根据公式（5）计算当前帧的特征值$\lambda(i)$。为了消除突发噪声的影响，对求出的$\lambda(i)$做长度为K（本文$K=5$）的中值滤波及均值滤波。

3.2.3 端点检测

根据每帧语音信号的特征值$\lambda(i)$，与噪声门限值TH做比较。如果当前帧i所求特征值$\lambda(i)>TH$，则以该帧的帧号作为有音片段的起点N_1，触发端点检测过程。如果由过去帧已经得到了N_1，那么当小于TH时，就以该帧的帧号作为有音片段的终点N_2。相反，如果N_1还未得到，那么当小于TH时，表明当前帧仍处于无音片段。

3.2.4 噪声频谱更新

在确定某一帧为语音结束后，如果判断连续M帧都不是语音帧，则利用这M帧估计新的噪声频谱能量$\bar{Y}_{noise_cur}$及噪声频谱密度$\bar{P}_{noise_cur}$，按照$\bar{Y}_{noise}=\mu\bar{Y}_{noise}+(1-\mu)\bar{Y}_{noise_cur}$和

$\bar{P}_{noise} = \mu\bar{P}_{noise} + (1-\mu)\bar{P}_{noise\_cur}$ 更新噪声频谱能量及频谱密度（$\mu = 0.3$），利用更新后的 $\bar{P}_{noise}$ 计算 P_{noise_avg}，并更新噪声门限值 TH [8]。

4 实验与结果

本研究试验数据采自维吾尔语口语语音语料库，语音信号采样率为 8KHz，16 比特量化，帧长为 128 点，帧移 64 点。

图 4 是一段语音“تىلىفون قىل ئادەتتىكى خوش（再见，保持联系）”的基于短时能量及短时过零率方法检测结果与本研究方法的检测结果对比图。

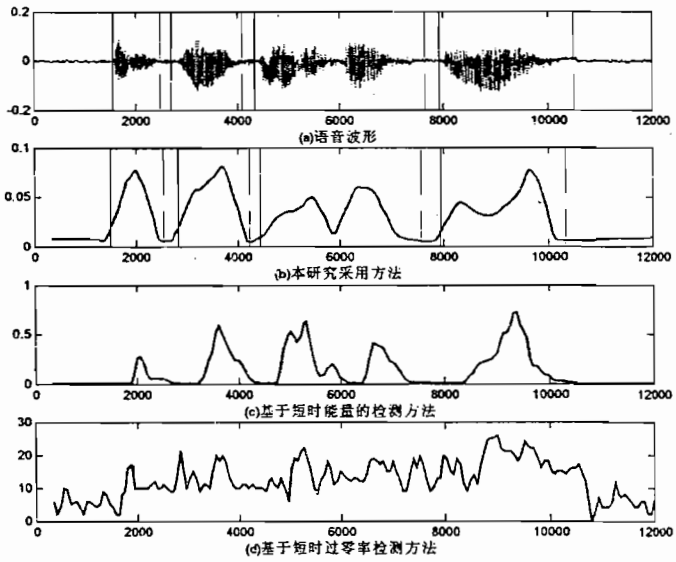

图 4 检测对比结果

实验结果表明，子带平均频谱密度的方法不仅能够很好的检测元音，而且还能较好的检测出清辅音的正确起始点，相对于传统方法在正确率上有所提高。而且，计算简单，可以用做实时端点检测。算法动态估计噪声密度，因此受噪声影响小。采用中值和均值滤波消除突发噪声的影响，而长度为 K 的中值滤波及均值滤波会对检测结果有 $\lfloor K/2 \rfloor$ 的误差，可以通过向前调整起点 $\lfloor K/2 \rfloor$ 帧和向后调整终点 $\lfloor K/2 \rfloor$ 帧修正此误差。

5 结束语

本文在研究传统基于短时能量和过零率的端点检测方法的基础上，基于维吾尔语语音的声频分布特征，提出了一种基于子带平均频率密度的维吾尔语语音端点检测方法，并通过实验证明了此方法的可靠性和有效性。

参考文献

[1] LI Jin，LIU Fu，WANG Ling，XU Hui-yan. Improved technology for speech endpoint detection[J]. Computer Engineering and Applications.2009，45(24)：133-135.

[2] Rabineer L R，Juang B H. Fundamentals of Speech Processing and Recognition[M].Prentice-Hall,1993.

[3]王昆仑，张贯虹，吐尔洪江·阿布都克里木. 维吾尔语元音的声频特征分析和识别[J]. 中文信息学报，2010，24(2)：

122-128.
[4]王炳锡，屈丹，彭煊.实用语音识别基础[M]. 北京：国防工业出版社，2005.
[5] ZHANG Jun-chang，JIANG Fei，LIU Hong. Study on endpoint detection based on multi-characteristic jointed in noisy environment[J]. Computer Engineering and Applications. 2009，45(32)：114-116.
[6] PIAO Chun-jun，MA Jing-xia，XU Peng. Study on noisy speech endpoint detection method[J]. Computer Applications.2006，26(11)：2685-2690.
[7] CHEN Zhen-biao XU Bo. Optimization of speech endpoint detection base on sub-band energy feature[J].2005,30(2):171-176.
[8] K. Yao, E .Shi, etc. Residual noise compensation for robust speech recognition non-stationary noise[C].Proc. ICASSP, (2):1125-1128,2000.