ISSN 1000-9825, CODEN RUXUEW
*Journal of Software*, Vol.19, Supplement, December 2008, pp.96−103

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

# 用于自动字幕生成系统的语音端点检测算法*

李 祺+, 马华东, 冯 硕

(北京邮电大学 智能通信软件与多媒体北京市重点实验室,北京 100876)

## A Robust Endpoint Detection Algorithm for Video Caption Generation

LI Qi+, MA Hua-Dong, FENG Shuo

(Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecomm, Beijing 100876, China)

+ Corresponding author: E-mail: qi.liqi2001@gmail.com

**Li Q, Ma HD, Feng S. A robust endpoint detection algorithm for video caption generation. *Journal of Software*, 2008,19(Suppl.):96−103.** http://www.jos.org.cn/1000-9825/19/s96.htm

**Abstract**: With the development of multimedia technology, the use of video has increased in many fields, and captions are frequently inserted into video images to aid the understanding of audience. This paper proposes a robust endpoint detection algorithm for continuous speech in noisy environment, and it can be used in automatic video caption generation systems. In the proposed algorithm, we integrate the widely used energy, zero crossing and entropy to form a new feature, EZE-feature, which possesses advantages while compensating the drawbacks of each individual. Moreover, an adaptive endpoint detection method is proposed which makes the EZE-feature modify its environment parameters by adapting to the strength of background noise. The proposed algorithm has been used in an automatic video caption generation system, and the performance of the algorithm is very well.

**Key words**: endpoint detection; caption; video caption generation; audio analysis; speech recognition

**摘 要**: 字幕信息有助于观众对音视频内容进行理解,在音视频文件中起着不可或缺的作用.针对自动字幕生成系统的要求,提出了一种灵活、高效的语音端点检测算法,可以在复杂背景噪声的情况下,从连续的音频信号中提取语音端点.将短时能量、短时过零率、短时信息熵这 3 种基本音频参数进行结合,形成新的音频特征参数:短时能零熵(EZE-feature),在结合了音频信号时域特征和频域特征优点的同时,规避了它们各自的不足.在此基础上,还提出了一种环境自适应的语音端点判定算法,在端点检测过程中对背景噪声进行实时分析,并根据背景噪声的变化对短时能零熵参数进行调整.该语音端点检测算法已被成功应用于自动字幕生成系统中.

**关键词**: 语音端点检测;字幕;字幕生成系统;音频分析;语音识别

* Supported by the National Natural Science Foundation of China under Grant No.90612013 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2006AA01Z304 (国家高技术研究发展计划(863)); the Cosponsored Project of Beijing Committee of Education of China under Grant No.SYS100130422, the 111 Project of China under Grant No.B08004

Received 2008-05-01; Accepted 2008-11-25

## 1 研究背景

电视字幕是电视图像、声音的补充和延伸,辅助观众对电视节目所传播的信息进行理解,提高电视节目的传播效果,在电视节目制作中有着不可代替的地位和作用.随着多媒体技术的发展以及人们生活水平的提高,电视节目在人们生活中起到了越来越重要的作用,人们对字幕机的效率和功能也有了更高层次的要求.

当前的字幕制作方法是:首先准备好字幕文稿,它是指在制作电视节目之前,事先写好的一个文本文件,记录着节目的标题,以及节目中人物的对白.通常情况下,字幕文稿中的一行对应一条字幕信息.为了避免“声画不同步”现象(即字幕与声音不一致),在添加字幕时,字幕编辑人员需要一边听声音,一边逐句地确定每一条字幕的开始时刻与结束时刻,生成字幕文件.为了更进一步的节省制作字幕所花费的人力资源并提高字幕制作效率,自动字幕生成系统已成为电视技术以及数字媒体技术研究的热点.

自动字幕生成系统利用语音端点检测算法从音频流中分辨出语音和非语音,然后利用语音端点检测的结果确定每条字幕的起始时间和结束时间,在此基础上生成字幕文件.由于电视节目种类多样,持续时间通常较长,所以能够用于自动字幕生成系统的语音端点检测算法需要具有可靠性、准确性、稳健性、自适应性、简单性、实时性以及不需要背景噪声先验知识等特点.

经过几十年的发展,人们提出了许多语音端点检测方法,这些方法大体分为两大类[4,7].其中一类是基于阈值的方法[1,2],首先提取每一帧音频信号的声学特征,然后把这些特征的幅值与设定的阈值进行比较,从而对每一音频帧进行分类.另一类是基于机器学习的方法[8,10],首先需要选取大量的样本对语音信号和噪声信号分别进行训练,确定模型参数,然后利用已经建立好的模型对音频信号进行分类,其检测过程和识别过程类似.基于机器学习的方法具有准确性好、稳健性较强的特点,但是此类方法的复杂度高,运算量大,很难应用到实时语音信号处理系统中去.其次,基于机器学习的方法对训练样本的完备性要求很高,而电视节目类型多样,电视音频中的背景声音具有不确定性,无法对背景噪声进行准确的建模,于是,在自动字幕生成系统中,我们选用基于阈值的检测方法.

基于阈值的方法具有简单、直观的优点,从而被广泛研究和使用[2,5].在这类方法中,特征参数的选取尤为重要.作为音频信号的基本特征参数,短时能量[1,4]以音频信号在一段时间内的强度为判断依据,在信噪比较高的情况下,基于短时能量的方法可以有效的区分语音信号和非语音信号,而在信噪比较低的情况下,其检测效果会变得很差.短时过零率表示一帧语音中信号越过横轴(零电平)的次数,在一定程度上反应了频率的高低,便于检测噪声信号中的直流分量,而在背景噪声较为复杂的情况下检测效果较差.近几年来,针对语音端点检测问题,研究者们提出了各种能区分语音和噪声的特征参数或其衍生参数[4,7−9],用来提高算法的抗噪声性能.比如倒谱系数[2]、带方差[9]、自相关函数[3,6]、信息熵[7]等都被逐渐应用到端点检测技术中.另外,还有些工作通过将语音信号的几种特征结合起来进行端点检测,而对语音端点的判决也由原来的单一门限发展到多门限[5].

然而,现有的语音端点检测技术无法直接应用于自动字幕生成系统中:

(1) 现有的绝大多数语音端点检测算法都是依赖语音信号本身音节的特征来对语音和噪声进行区分[3].然而,在噪声环境下,某些以清音或摩擦音、爆破音开头的语音信号易被噪声淹没,从而导致误检.

(2) 在判决端点位置时,多数端点检测算法都是假设噪音信号在检测过程中是平稳的,并利用音频信号最初的几帧对背景噪声进行分析.然而,在电视节目中,背景噪声是在不断发生变化的(如电视剧中人物活动场景从街道转入办公大楼),如不动态的对背景噪声进行调整必将对检测结果造成影响.

(3) 目前大多数端点检测算法所选特征单一,具有一定的局限性,而且固定的门限阈值往往不能适应电视节目中各种不同的背景噪声环境,当信噪比降低时,性能明显下降[1,6].

(4) 目前大多数语音端点检测算法主要被应用于声纹分析、语音拨号等系统中[1,11].这一类应用所需处理的音频信号持续时间通常较短(通常为 2s~10s),而电视节目中的音频文件通常持续时间较长,且背景噪声具有不确定性.

(5) 目前大多数语音端点检测算法都是针对识别系统进行设计的,从而侧重于端点出现时刻的精确性,从而加大了特征提取的复杂度[2,11].然而在自动字幕生成系统中,我们对于精度的要求相对较弱.例如,一行字幕在

第 1 000ms时出现和它在 1 010ms时出现对观众是没有影响的,我们更侧重于端点的“存在性”以及算法的运行效率.

基于上述分析,我们认真学习了现有的语音端点检测方法并充分考虑到自动字幕机这一应用背景,将音频信号时域特征和频域特征相结合,提出一种新的音频特征参数,短时能零熵(EZE-feature),能够在复杂噪声条件下准确的区分语音和非语音.在此基础上,我们提出一种新的环境自适应的语音端点检测方法,对背景噪声进行实时分析,并在此基础上对短时能零熵进行动态的调整,从而能够在复杂的背景噪声环境下,对连续语音进行端点检测.本文提出的环境自适应的语音端点检测算法已被成功应用于项目合作单位新奥特公司研制的自动唱词机 NC8000 中.

## 2 特征参数的提取

音频信号是一种随时间变化比较缓慢的信号,可以认为在很短时间里(如 10ms~20ms 之内)信号近似不变,因此可以借用平稳过程的分析处理方法,把音频信号分成一些短时间段(音频帧)进行处理,这些短时间段具有相对的固定性.本论文中,我们取 10ms 为一帧,相邻帧之间重叠 5ms.

### 2.1 短时能量的提取

我们将音频帧的短时能量定义为该帧中所有采样值平方的和,即第 $i$ 帧的短时能量为:

$$E_i = \sum_{n=1}^{N} S_n^2 \tag{1}$$

其中,$N$表示第$i$帧中所包含的音频采样数量,$S_n$表示第$n$个采样的取样值.

### 2.2 短时过零率的提取

过零率是声音信号处理过程中一个常用的音频特征参数.当离散语音信号的时域波形通过时间横轴时,相邻时刻的采样值如果具有不同的符号,称为“过零”.单位时间的过零次数称为“过零率”,即单位时间内音频采样值符号变换的次数.第 $i$ 帧的短时过零率定义如下:

$$Z_n = \frac{1}{2}\sum_{n=1}^{N} |\operatorname{sgn}(S_n) - \operatorname{sgn}(S_{n-1})| \tag{2}$$

其中,$S_n$表示第$n$个音频采样的值;sgn( )为符号函数,定义为

$$\operatorname{sgn}(S_n) = \begin{cases} 1, & S_n > 0 \\ -1, & S_n < 0 \end{cases} \tag{3}$$

### 2.3 短时信息熵的提取

语音的感知过程与人类听觉系统具有频谱分析功能是紧密相关的.信息熵是频域的重要音频参数,它反应了语音信号所传达的信息量的大小.短时信息熵计算方法如下:

(1) 利用短时傅里叶变换(FFT)对每一帧的信号进行由时域向频域的转换:

$$X(\omega) = \sum_{n=-\infty}^{\infty} S_n \mathrm{e}^{-j\omega n} \tag{4}$$

(2) 计算每一频率的出现概率:

$$p_i = \frac{s(f_i)}{\sum_{k=1}^{M} s(f_k)} \tag{5}$$

其中,$s(f_i)$表示频率$f$的频谱能量,$p_i$表示相应频率的出现概率,$M$表示傅里叶变换计算得出的频率的总数.所规定的约束条件为

$$\begin{cases} 1.\ s(f_i)=0, \text{ if } f_i \le 50\text{Hz or } f_i \ge 3750 \\ 2.\ p_i=0, \text{ if } p_i \ge 0.9 \end{cases} \tag{6}$$

第 1 个约束公式用来保证语音信号的频率范围.因为人的发音频率基本集中在 50Hz~3 750Hz 之间,所以我

们把频率限定在这个范围之内.第 2 个约束公式用来滤除在某些频率上持续发生的噪声.

(3) 计算音频信号的信息熵:

$$H_i = -\sum_{j=1}^{M} p_j \log p_j \tag{7}$$

其中,$M$表示傅里叶变换计算得出的频率的总数,即窗口宽度,$p_i$表示相应频率的出现概率,$H_i$表示第$i$帧的短时信息熵.

### 2.4 短时能零熵参数的提取

在上述 3 个音频特征参数的基础上,本文提出了一个结合时域和频域的语音特征参数,短时能零熵,记作*EZE-feature*.第 $i$ 帧的短时能零熵定义如下:

$$EZE - feature_i = (E_i - E_b)\cdot(Z_i - Z_b)\cdot(H_i - H_b) \tag{8}$$

其中,$E_i$,$Z_i$和$H_i$分别表示第$i$帧的短时能量、短时过零率和短时信息熵;而$E_b$,$Z_b$和$H_b$则分别表示了当前背景噪声的短时能量、短时过零率和短时信息熵.短时能零熵综合考虑了音频信号时域和频域的特征,将时域和频域的特征参数结合在一起,能够发挥它们各自的长处,并在一定程度上规避它们各自的缺点,从而能够有效的应对各种不同类型的背景噪声.除此之外,与某些复杂的音频特征值相比,短时能零熵的提取方法相对简单,运算效率较高.

## 3 自适应的语音端点检测算法

### 3.1 自适应的语音端点检测过程

在传统的语音端点检测过程中,研究者通常选取音频文件的前几帧作为背景噪声进行分析[3,6],在语音端点检测的过程中,假设背景噪声不发生变化.然而,在电视节目中,背景噪声是不断变化且无法预测的.为此,本文提出了一个新的语音端点检测算法,在端点检测的过程中,可以及时地对背景噪声进行分析和调整.具体过程如图 1 所示.

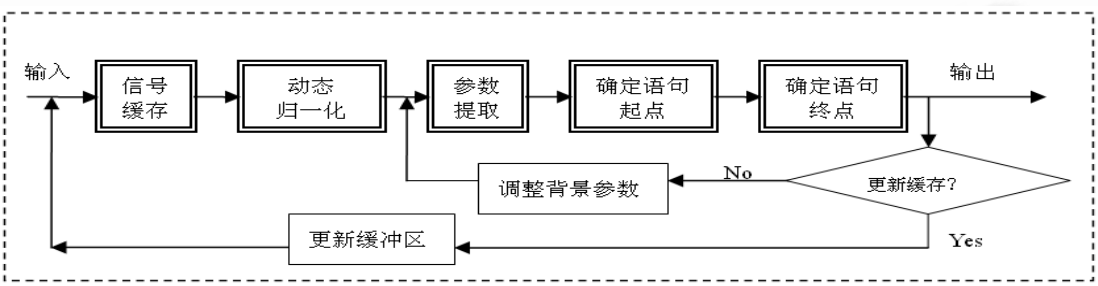



Fig.1 The process of the adaptive robust endpoint detection algorithm

图 1 自适应的语音端点检测过程

(1) 动态归一化.

假设对音频信号进行采样后得到采样序列,$g_n$表示第$n$个采样点的取值,进行归一化处理后,得到归一化的采样值$S_n$.

$$S_n = g_n / g_{\max} \tag{9}$$

其中,$g_{\max}$为缓冲区中所有音频采样绝对值的最大值.在语音端点检测过程中,我们首先根据当前缓冲区中存储的音频信号来确定$g_{\max}$的初始值.每检测到一个语句终点,便判断当前缓冲区是否需要更新.若需要,则从这一语句终点开始读入音频信号至缓冲区内,更新$g_{\max}$的取值,并对缓冲区内的音频信号进行归一化操作.

(2) 提取基本音频特征参数.将音频信号分帧后,提取每一帧的短时能量、短时过零率、短时信息熵,并对这些参数进行平滑处理.

(3) 提取背景噪声参数.提取背景噪声部分平均短时能量、平均短时过零率和平均短时信息熵.首先,我们选取音频信号最初的 5 帧进行背景噪声分析,并在语音端点检测的过程中,对背景噪声参数进行实时调整.

(4) 短时能零熵提取.利用背景噪声参数,对未进行分类的音频信号提取短时能零熵参数,并生成序列$X_1X_2X_3\ldots X_{n-1}X_n$.

(5) 确定语音的起始位置.

(6) 确定语音的终止位置.

每找到一个语音的终止点,需要判断缓冲区内未进行分类的音频信号长度是否大于 5s,若不足 5s 则从当前的语音终止点开始,读入音频信号至缓冲区内,转到(1).若不需要更新缓冲区,则转到(3),并取该语音终止点后的 5 帧音频信号进行背景噪声分析.

### 3.2 语音起始位置的判定

假设缓冲区内未进行分类的短时能零熵序列为$X_mX_{m+1}X_{m+2}X_{m+3}\ldots X_{n-1}X_n$.

(1) 对于某一时刻$t$,设其对应的短时能零熵参数为$X_t$,寻找短时能零熵的峰值点$X_{t+j}$使得:

$$X_t \le X_{t+1} \le X_{t+2} \le \ldots \le X_{t+j} \quad \text{且} \quad X_{t+j} \ge X_{t+j+1} \tag{10}$$

即 $t$ 时刻到 $t+j+1$ 时刻为音频信号短时能零熵的上升区间.

(2) 计算短时能零熵参数上升区间内的平均斜率

$$R_t = (X_{t+j} - X_t)/j \tag{11}$$

(3) 设定相对门限阈值$R_m$,如果有 $R_t \ge R_m$,则将$t$时刻标记为语音的终点,令$t=t+j+1$,去寻找与之相匹配的语音终点.如果 $R_t \le R_m$,则认为$t$时刻不是语音起点,令$t=t+j+1$,转到(1).

### 3.3 语音终止位置的判定

语音终止位置的判定与起始位置的判定相似.

(1) 对于某一时刻$t$,设其对应对短时能零熵参数为$X_t$,寻找短时能零熵的低点$X_{t+j}$使得:

$$X_t \ge X_{t+1} \ge X_{t+2} \ge \ldots \ge X_{t+j} \quad \text{且} \quad X_{t+j} \le X_{t+j+1} \tag{12}$$

即 $t$ 时刻到 $t+j+1$ 时刻为音频信号短时能零熵的下降区间.

(2) 计算短时能零熵参数下降区间内的平均斜率

$$R'_t = (X_t - X_{t+j})/j \tag{13}$$

(3) 相对门限阈值$R_m$同第 3.2 节,如果有 $R'_t \ge R_m$,则将$t$时刻标记为语音的终点,令$t=t+j+1$,确定是否需要更新缓冲区,并重新计算环境参数.如果 $R'_t \le R_m$,则认为$t$时刻不是语音起点,令$t=t+j+1$,转到(1).

## 4 实验与分析

### 4.1 音频参数有效性的验证

我们利用实验来验证本文所提出的音频特征参数是否能够有效的区分语音和非语音,找出语音的端点.实验中,我们通过语音的识别率来说明语音端点检测的准确性.

语音识别系统平台的构建:在 Matlab 环境下,用 HMM 模型作语音识别模型.实验中用到的语料是自己在实验室环境下通过 PC 录制的,音频格式 wav、平均数据速率 32.00kb/s、采样速率 16.00kb/s、音频采样大小 16 位、单声道.语音短时处理中,帧长 10ms、相邻帧之间重叠 5ms,特征参数是短时能量加 12 维的 MFCC 参数.

用于实验的语料包括 500 个英文数字.由 10 个人(男女各 5 人)分别从“0”读到“9”,每个英文数字读 5 遍.我们分别在无噪声环境、有音乐的环境和汽车内环境中,将本文提出的参数与短时能量、短时信息熵进行比较.比较结果如图 2 所示.

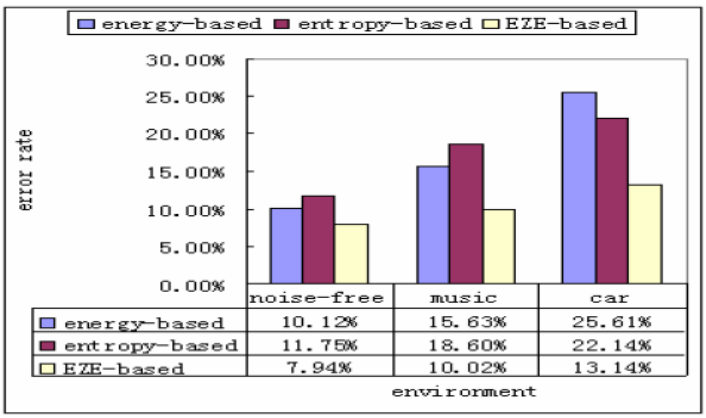



Fig.2 Error rate of three endpoint detection algorithms in different environments

图 2 3 种语音端点检测算法在不同背景噪声条件下的错误率

实验结果表明,在无噪声环境下,三种音频参数的检测效果都很好.然而,由于基于能量的音频特征参数只对音频信号的强度进行分析,难以区分语音和无法预知的背景噪声,所以在引入车内噪声的情况下检测效果较差.基于信息熵的音频特征参数难以区分语音和乐音,于是在引入音乐等背景噪音的情况下,表现出了明显的不足.本文提出的音频参数,短时能零熵(EZE-feature),将短时能量与短时信息熵进行结合,并加入了另一个重要的音频参数短时过零率对其进行辅助.在融合了音频信号的时域特征与频域特征的优点的同时,规避了二者各自的不足,从而在引入音乐背景以及车内噪声背景的情况下的都取得了较好的检测效果.此外,与一些复杂的音频特征参数相比,短时能零熵提取算法简单,提高了检测效率.

### 4.2 语音端点检测算法在自动唱词系统中的应用

自动字幕生成系统的框架结构如图 3 所示:首先自动字幕生成系统接受用户输入的音频文件(wav 格式)以及对应的字幕文稿文件(txt 格式).系统通过语音端点检测模块对音频文件进行语句端点检测,并根据这些端点找出音频中每一句话的起始点和结束点,生成一个语句端点文件.文件的每一行都有一对时间戳组成,对应一行字幕的起始时间和结束时间.最后,字幕生成模块将语句端点文件与字幕文稿进行整合,最终生成通用的 srt 格式的字幕文件.图 4 为使用自动字幕生成系统生成语音端点文件并与字幕文本相结合,最终生成 srt 格式的字幕文件并使用视频播放器播放字幕的过程.

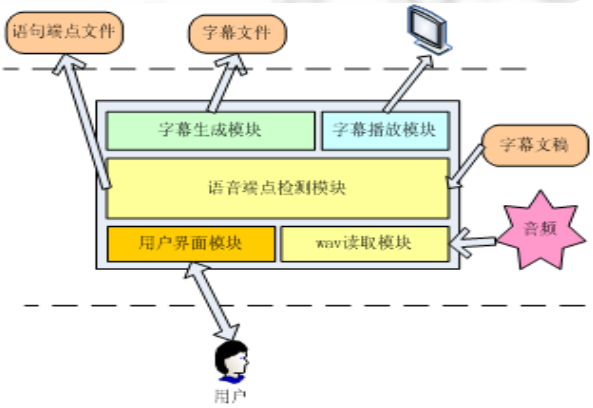
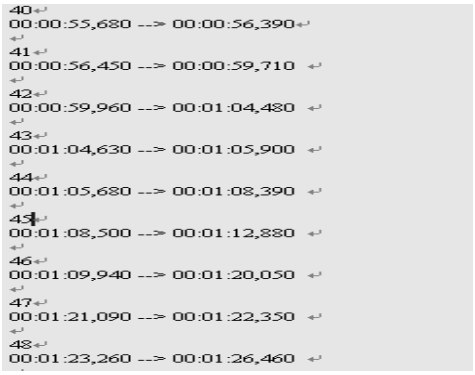


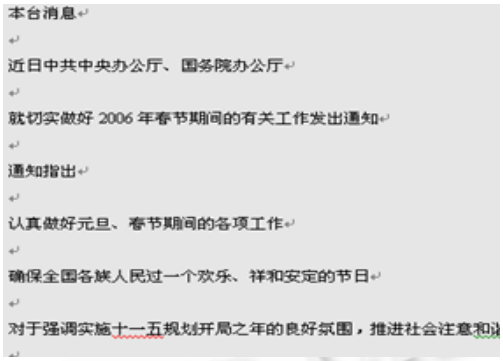

Fig.3 Framework of the video caption generation

图 3 自动字幕生成系统的框架结构




(a) File of speech endpoints

(a) 语音端点文件




(b) File of caption text

(b) 字幕文本文件


(c) Video caption file

(c) 视频字幕文件


(d) Playing result of the video caption generation system

(d) 自动字幕生成系统的运行效果

Fig.4 The generating process of the video caption

图 4 自动字幕生成系统的框架结构

我们将本文提出的语音端点检测算法用于字幕生成系统.实验数据全部截取于电视节目,共计 20 段音频.10 段截取于中央电视台新闻联播(包括开场音乐、新闻内容以及天气预报),10 段截取于中央电视台的访谈节目.每段音频长约 30 分钟.针对每一段音频文件,先由工作人员通过手动的方式对语句端点进行标注,我们将语音端点检测的结果与人工标注的结果进行比对,当两者相差超过 50ms 或发生漏检时,我们认为语音端点检测结果错误.实验中,我们将本文提出的环境自适应的语音端点检测算法与传统的基于门限的语音端点检测算法进行比较,实验结果如表 1 所示.从实验结果,可以看出,通过对背景噪声进行动态分析,本文提出的语音端点检测算法相对于传统的语音端点检测算法,其精确性提高了 10%.

**Table 1** The error rate of two sentence endpoint detection algorithms

**表 1** 两种语句端点检测算法的错误率

| 检测算法 | 实际语句端点数 | 检测正确的语句端点数 | 检测错误率(%) |
|---|---|---|---|
| 基于短时能零熵的自适应端点检测算法 | 3 900 | 3 494 | 10.41 |
| 基于短时能零熵的传统语音端点检测算法 | 3 900 | 3 075 | 21.15 |

## 5 总 结

本文提出了一种环境自适应的语音端点检测算法,可以在复杂背景噪声的情况下,从连续的音频信号中准确地提取语音端点.该算法检测效果好,计算效率较高,已被成功应用于自动字幕生成系统中.

在今后的研究中,我们将考虑如何对关键字、词进行准确定位,从而提高字幕生成系统中语句端点的检测精度.

**致谢** 在此,我们向对本文的工作给予支持和建议的同行,尤其是北京新奥特硅谷视频有限公司的姚景平老师、蔡常军经理等工作人员表示感谢.

**References**:

[1] Evangelopoulos G, Maragos P. Multiband modulation energy tracking for noisy speech detection. IEEE Trans. on Audio, Speech and Language Process, 2006,14(6):2024–2038.

[2] Junqua JC, Mak B, Reaves B. A robust algorithm for word boundary detection in the presence of noise. IEEE Trans. on Speech and Audio Process, 2004,2(3):406−412.

[3] Koichi Y, Firas J, Klaus R, Akinori K. Robust endpoint detection for speech recognition based on discriminative feature extraction. In: Proc. of the IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. 2006. 805−808.

[4] Li Q, Zheng J, Tsai A, Zhou Q. Robust endpoint detection and energy normalization for real-time speech and speaker recognition. IEEE Trans. on Speech Audio Process, 2002,10(3):146–157.

[5] Li Q, Zheng J, Zhou Q, Lee CH. A robust, real-time endpoint detector with energy normalization for ASR in adverse environments. In: Proc. of the IEEE Int'l Conf. Acoust. Speech, Signal Process. 2001. 233–236.

[6] Wu BF, Wang KC. Robust endpoint detection algorithm based on the adaptive band-partitioning spectral entropy in adverse environments. IEEE Trans. on Speech Audio Process, 2005,13(5):762–775.

[7] Yamamoto K, Jabloun F. Robust endpoint detection for speech recognition based on discriminative feature. In: Proc. of the IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. 2006. 114–119.

[8] Zhang X, Li G, Qiao F. A speech endpoint detection algorithm based on entropy and RBF neural network. In: Proc. of the IEEE Int'l Conf. on Granular Computing. 2007. 506–509.

[9] Liu HP, Li X, Zheng Y, Xu BL, Jiang N. Speech endpoint detection based on improved adaptive band-partitioning spectral entropy. Journal of System Simulation, 2008,20(5):1366−1371 (in Chinese with abstract English).

[10] Yan BF, Zhu XY, Zhang ZJ, Zhang F. Robust speech recognition based on neighborhood space. Journal of Software, 2007,18(4):878−883 (in Chinese with abstract English). http://www.jos.org.cn/1000-9825/18/878.htm

[11] Tang Y, Liu WJ, Xu B. Mandarin digit string recognition based on segment model using posterior probability decoding. Chinese Journal of Computers, 2006,29(4):635−641 (in Chinese with abstract English).

**附中文参考文献**:

[9] 刘华平,李昕,郑宇,徐柏龄,姜宁.一种改进的自适应子带谱熵语音端点检测方法.系统仿真学报,2008,20(5):1366−1371.

[10] 严斌峰,朱小燕.基于邻接空间的鲁棒语音识别方法.软件学报,2007,18(4):878−883. http://www.jos.org.cn/1000-9825/18/878.htm

[11] 唐赟,刘文举,徐波.基于后验概率解码段模型的汉语语音数字串识别.计算机学报,2006,29(4):635−641.

**李祺**(1981－),女,北京人,主要研究领域为音频信号处理,音频传感器网络.

**冯硕**(1983－),男,硕士,主要研究领域为音频信号处理.

**马华东**(1964－),男,博士,教授,博士生导师,主要研究领域为多媒体网络与系统,传感器网络,网格计算.