强噪声背景下汉语语音端点检测和音节分割

杨崇林
(哈尔滨工程大学水声工程系 ,哈尔滨 150001)
李雪耀　孙　羽
(哈尔滨工程大学计算机与信息科学系 ,哈尔滨 150001)

摘　要　根据汉语语音的特点 ,提出在强噪声背晾下对汉语语音进行端点检测和音节分割的新算法. 在 85dB 的噪声环境中 ,实验考察了端点检测的正确性和音节分割的稳定性. 结果表明 ,算法在这两方面达到了很高的性能 ,且与话者无关. ①

关键词　噪声 ;端点检测 ;音节分割 ;语音识别
分类号　TN912. 34

Endpoint Detection and Syllable Division for Chinese Speech in High Noise Condition

Yang Chonglin
(Underwater Acoustic Dept. Harbin Engineering University ,Harbin150001)
Li Xueyao　Sun yu
(Computer and Information Dept. Haribn Engineering Univeristy ,Haribn 150001)

Abstract　In this paper ,according to the character of chinese speech ,a new algorithm of endpoint detection and syllable division for chinese speech in high noise condition is proposed. The experiments in the presence of 85dB noise showed that the algorithm achieved pretty high performance when tested with the correction of endpoint detection and the reliability of syllable division ,and it is less dependent of different speakers.

Key wrods　noise ;endpoint detection ;syllable division ,speech recognition

0　引　言

语音识别作为人机交互的一种手段 ,在自动化、通信、指挥控制等许多领域具有广泛的应用价值 ,所以它越来越受到人们的重视 ,在船舶工业中的应用也越来越广泛. 然而尽管各行业对语音识别的实用化、商品化抱极大期望 ,这一步却难以实现. 原因在于 ,虽然目前的很多语音识别系统在实验室条件下都能达到很高的识别率 ,但到了存在着一定背景噪声的

① 收稿日期 :1996 - 12 - 30
责任编辑 :王义成

应用场合,则性能无一例外地将会急剧下降.

研究表明[1],即使在安静环境下,语音识别系统一半以上的识别错误来自端点检测器.作为语音识别系统的第一步,端点检测的关键性不容忽视.传统的端点检测算法都是针对实验室安静环境的,近年来人们才开始研究噪声环境下语音的端点检测.

汉语语音与其它语言相比,有自己的一些特点[2].汉语中每个字都是单音节,每一个音节又都是由声母(包括零声母)和由若干音素组成的韵母拼音而成.汉语的辅音大部分是清辅音,受到噪声干扰时,极易被噪声淹没.从这点来说,噪声中汉语语音端点检测,关键是寻找语音的准确的起始点.

现有的对含噪汉语语音的端点检测方法[3],检测起始点时多从浊音段固定往前取.统计表明[2],汉语声母发音长度差别很大,短的只有 10ms 左右,长的可达 200ms 以上.还有的方法[4]存在着一定的局限性.此外,根据汉语语音的音节特点对语音进行音节分割,对大词表的汉语语音识别来讲可以作为一个很好的粗分类手段,然而目前尚未有含噪语音进行音节分割的方法.

目前船舶工业已提出了对语音识别实用化的具体要求,一些研究项目已经立项,本文的工作正是在这一背景下开展的.文中提出了一种在强噪声背景下对汉语语音进行端点检测和音节分割的有效算法.算法用到了时频(Time - Frequency, TF)参数.文中还给出了从端点检测的正确性和音节分割的稳定性两方面对算法的测试结果.

1　TF 参数

一般的端点检测算法常用的参数有短时能量、短时平均过零率、零能比、零能积等简单的时域参数以及 LPC 残差、基音信息等.当背景噪声的干扰较严重时,除零能比还能保持对元音和噪声较好的区分外,其它参数已显力不从心.J - C Junqual[5]在 1994 年提出了一种新的参数,即 TF(Time - Frequency, TF)参数,它的计算原理如图 1.

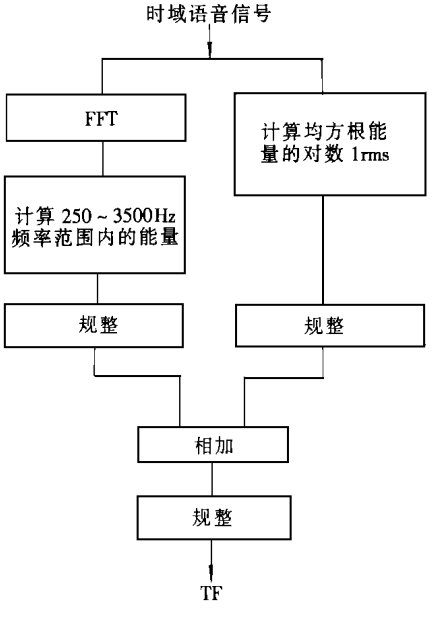


图 1　TF 参数的计算原理

虽然 TF 参数仍然是一个能量参数,但是它不仅统计了语音信号在时域上的有效能量,还统计了语音信号在频域上 250Hz～3500Hz 频率范围内的能量,而且两者相加.汉语语音的能量主要集中在元音上,而元音 3 个共振峰的频率范围主要集中在 250Hz～3500Hz 之间[2].根据 TF 参数的计算原理,无论背景噪声的强度如何,对含噪语音所计算的 TF 的谱必然在汉语语音的元音部分形成一个强峰,而在语音的其它部分和噪声段则很平缓,如图 2 对发音"准备"所计算的 TF 值所示.由此它必然能在检测汉语语音的能量集中区时,起到十分显著的作用.这一点是本文的算法所依据的重要原理,将在后面的算

法中得到体现.

2 端点检测算法

端检测算法共使用了4个参数:短时能量、TF参数、短时能量的均方根对数1rms和短时平均过零率.其中,短时能量用来粗略检测出包含完整语音部分的信号段,TF参数用来检测第一音节的能量集中区,1rms和过零率用来精确检测包含辅音的语音起始段.所以,整个算法是由3个步骤组成:

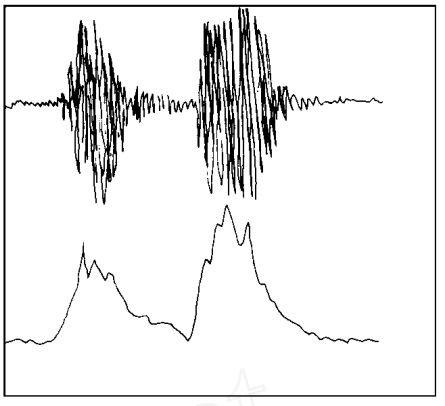


图2 发音"准备"的时域波形及TF曲线

步骤1 用简单的能量参数粗略检测出含语音的信号段.在此之前取噪声参考帧用来求噪声的能量参考值.算法参考了文献[6].此过程中为后面的检测计算3个参数:每帧能量的均方根对数值1rms,每帧的过零率和频域的部分能量.限于篇幅,具体实现过程这里不再赘述.

步骤2 用TF参数检测出语音第一音节的能量集中区(元音部分).

定义阈值 $th1$

$$th1 = (Emax - aver) \cdot aver \cdot A \tag{1}$$

式中,$Emax$表示语音第一音节的1rms的最大值,$aver$是噪声参考帧的1rms的平均值,A是一个常数.

将此阈值应用于各帧的TF参数,即可检测到第一音节的能量集中区.

步骤3 从在步骤2中检测到的能量集中区的起点开始前推,按如下过程搜索语音信号的起始点:

(1)以能量集中区的起点为起点;

(2)从原起点前推20ms作为新的起点;

(3)若新起点处的1rms的均值大于 $th2$ 转2,否则继续;

(4)若新起点前的20ms的1rms的均值大于 $th3$ 转2,否则继续;

(5)若新起点前的20ms的过零率均值大于 $izct$ 转2,否则继续;

(6)定现新起点为语音信号的起始点,结束.

算法中所用的3个阈值定义为

$$th2 = \frac{Emax - aver}{B} + aver \tag{2}$$

$$th3 = C \cdot aver \tag{3}$$

$$izct = D \cdot E \quad \{噪声参考帧的过零率\} \tag{4}$$

式中,$Emax$和 $aver$ 意义同式(1),B,C,D都是常数,$E\{\cdot\}$表示取平均.

终点的检测不必遵循这个过程,可直接在步骤1的检测结果的尾部去掉几帧后作为终点.

图3给出了按以上的3步端点检测算法对发音"准备"检测出的语音段、能量集中区和语音的起始点和终点,图中,A和C标记的是语音段的起点和终点,B标记的是能量集中区

的起点.

3　音节分割算法

根据 TF 参数的原理,它既然能很好地应用于检测汉语语音的元音部分的能量集中区,汉语每个音节只有一个能量集中区,那么 TF 参数轨迹必然能很好的地体现了汉语语音的音节特征. 根据图 4 给出的对发音"计算机"所得到的 TF 值的轨迹可证实. 据此提出一个简捷的音节分割算法,通过检测语音段的 TF 值的轨迹峰值的个数来判定所发语音的音节数. 算法描述如下:

步骤 1　置音节数为 0,置 n 为 0;

步骤 2　$n \leftarrow n+1$,若 $n = N+1$ 转步骤 4;否则取第 n 个 TF 值,若其值小于 $H1$,继续步骤 2,否则转步骤 3;

步骤 3　$n \leftarrow n+1$,若 $n = N+1$ 转步骤 4;否则取第 n 个 TF 值,若其值大于 $H2$,继续步骤 3,否则转步骤 2;

步骤 4　结束.

算法中 $H1$, $H2$ 为如图 4 中所示的下上两条横线, N 为 TF 值的个数.

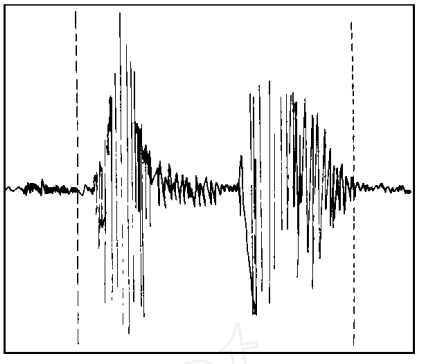


图 3　对发音"准备"的端点检测结果

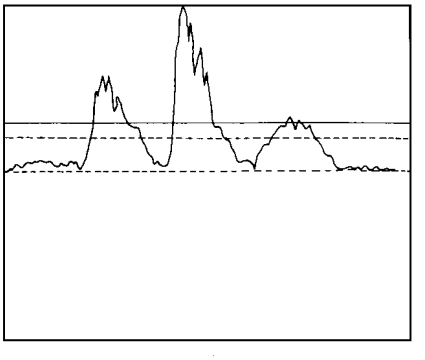


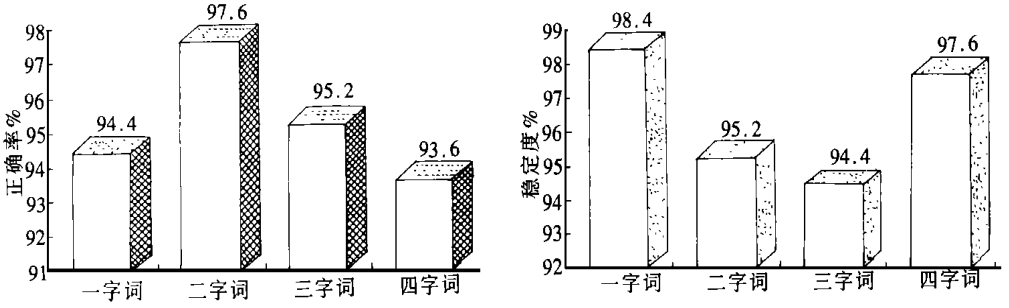
图 4　发音"计算机"的 TF 轨迹

4　实验与测试结果

上述的端点检测和音节分割算法,都已在一个由高速数字信号处理器 TMS320C30 和 PC486 组成的系统上实现. 其中,采样是通过 14 位 A/D 芯片 TLC32044C 完成的,采样频率 10kHz. 为了验证这两个算法,作者在高噪声背景中对它们进行了实验. 背景噪声是在某型舰艇指挥舱里的噪声的录音,其中包括螺旋桨等机械噪声、嘈杂的人声等. 噪声级为 85dB(线性计权). 实验中语音帧长取 6.4ms,发音者为 25 岁男性,发音包括汉语 1,2,3,4 字词汇各 25 个,共 100 个,每个词发音 5 次. 所用的话筒无指向性.

对端点检测算法考察的是它的正确率,这是通过与人工检测的结果相比较得到的. 对音节分割算法,因为音节分割在汉语大词汇量语音识别中通常是作为粗分类的一个手段,而对粗分类最重要的指标是分类的稳定性,因此作者考察的也是音节分割的稳定性.

实验结果如图 5 所示. 从图中可以看到,本文所提出的在强噪声背景下进行端点检测和音节分割的算法都达到了很高的性能,完全可以应用于有强噪声背景的汉语语音识别系统. 在端点检测中出现的错误和音节分割中的不稳定,是由于某些音节的元音能量不太强,如图 4 中的第 3 音节,这些可以在算法实现时采用一些特殊的手段加以解决.



(a) 端点检测的正确率统计结果 (b) 音节分割的稳定度统计结果

图5 噪声背景下端点检测和音节分割的实验结果

5 结 论

本文基于 TF 参数,给出了在强噪声背景下,对汉语语音进行端点检测和音节分割的算法. 该算法的显著特点是:

(1)针对汉语语音设计,适用于汉语语音识别应用系统;

(2)因为 TF 参数体现说话语音的频率范围,算法不对特定人,与话者无关;

(3)算法基于对噪声的学习,抗噪声能力强.

此外,算法实时、准确、稳定,它对于在高噪声背景如飞机、舰船指挥舱室等环境下进行语音识别具有很好的应用前景,本文的工作为这方面的应用研究提供了理论基础.

论文写作过程中承花栅教授热情指导,在此谨致谢意.

参 考 文 献

1 J C junqua. Robustness and cooperative multimodel man - machine communication applications. Proc Second Venaco Workshop and ESCA ETRW, Sept. 1991

2 陈永彬,王仁华. 语音信号处理. 合肥:中国科学技术大学出版社,1990

3 杨家沅. 语音识别与合成. 成都:四川大学出版社,1994

4 杨子云,徐近霈. 高噪环境下命令语音识别的特殊方法. 计算机智能接口与智能应用论文集,1993

5 J C Junqua. A robust algorithm for word boundary detection in the presence of noise. IEEE Trans on Speech and Audio Processing, 1994, 2(3): 406～412

6 陈尚勤,罗承烈. 近代语音识别. 成都:电子科技大学出版社,1991