一种改进的基于时域参数的语音切分算法*)

林 帆 徐明星
(中山大学计算机科学系 广州 510275)
(清华大学计算机科学与技术系智能技术与系统国家重点实验室 北京 100084)

摘 要 本文探讨了基于时域的语音切分算法,在前人研究的基础上,提出一种改进算法——自适应、前后搜索和检测短时脉冲噪音算法。该算法主要利用语音信号的短时参数,采用统计的方法定出切分所需要的阈值;根据背景音和静音过零率的不同,进一步搜索符合要求的静音帧;同时滤去短时脉冲噪音。实验证明,该算法准确率很高,有很好的鲁棒性,允许误差在 60 ms 的范围内,对于原始语音切分错误率为 5.04%;在信噪比(SNR)大于等于 2 dB 的情况下,对带噪语音的切分错误率为 10%～20%。

关键词 语音切分,短时参数,自适应,前后搜索,检测短时脉冲噪音

An Improved Speech Detection Algorithm Based on Time-domain Parameter

LIN Fan XU Ming-Xing
(Department of Computer Science, Sun Yat-Sen University, Guangzhou 510275)
(State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science and Technology, Tsinghua University, Beijing 100084)

Abstract This paper researches on speech detection algorithm based on time domain, and describes an adaptive, both forwards and backwards search, detecting short-term pulse noise algorithm. This algorithm uses a variety of features including the frame amplitude and zero crossing rate to calculate threshold using statistical method. And it searches much further for the unvoiced frame according to the ZCR(zero crossing rate), which is differ from unvoiced frame to background frame. This algorithm also detects impulse noise that last little. Experimental results show that this improvement has good performance, even in noisy condition. Testing the original speech, the error rate is 5.04%, and in noisy environment with a SNR of beyond 2 dB, the rate is around 80%～90%.

Keywords Speech detection, Short-term variety, Adaptive, Search forwards and backwards, Detecting short-term pulse noise

1 引言

在很多应用领域,把语音切分成语音段跟无声段(包括噪音)是非常必要的,因为往往只需要研究语音信号,过多的无声段以及噪音会造成影响,比如语音识别中,准确标出语音信号的起点和终点,对整个识别系统性能的提高是很有帮助的;另一方面,对语音段和无声段分别进行处理,比如可以忽略无声段,这也有利于减少处理的时间以及系统消耗的功率。

语音切分方法主要有两类:基于时域和基于频域两种。基于时域的方法主要通过计算语音信号的短时参数,比如短时能量、短时幅度、短时过零率,利用清音、浊音与无声段的这些短时参数的不同概率分布,分辨声音类型,达到切分的目的。基于频域的方法则是利用线性预测编码(LPC)参数、倒谱系数、共振峰频率及带宽、反射系数等等,根据频谱,可以发现语音与噪音在这类参数上的区别,根据这些特征进行切分。基于频谱熵的语音切分方法也很成熟,ABSE(adaptive band-partition spectral entropy)方法[8]利用 BSE(band-partitioning spectral entropy)参数,将频谱分成不同的子带,采用有用子带的信息切分语音,能够有效地应用于噪音环境中,甚至包括背景噪音逐渐增强的情况。此外,还有基于模式识别方法[5,6]以及基于贝叶斯方法[7]的语音切分算法在鲁棒性方面也取得很好的效果。而利用 HOS(higher order statistics)属性的方法[9]性能也较好,但是需要统计大量语音的信息,计算量大。相对而言,基于时域的语音切分算法不需要做很多的变换,算法简单、运算量小,速度快,能很好地应用在允许少量延时的实时系统中。

对基于时域的语音切分方法的研究时间很长,很多人做了有益的尝试。早在 1975 年,L. R. Rabiner 和 M. R. Sambur 就提出过语音切分的基本算法,在此基础上,其他人提出一些新的算法[4,10,11]。这些方法在高信噪比的情况下切分准确度较高,但是在信噪比很低的情况下,效果不佳。此外,缺乏一种通用的切分算法评价方案,不同的方法统计出来的结果可能会出现较大偏差。本文在前人研究的基础上,提出了一种改进算法。该算法能够很好地适应不同信噪比的噪音环境下,鲁棒性高。实验结果表明,改进算法与原方法相比,鲁棒性得到较大提高,错误率大幅降低。

本文的结构如下:第 2 节介绍文[1]提出的语音切分基本算法,具体分析其不足之处;第 3 节详细说明改进算法,分别介绍自适应的实现、前后搜索以及检测短时脉冲噪音;第 4 节是实验部分,对上述两种算法在不同的语音环境下的实验结果进行对比;最后对改进算法进行总结,以及对以后研究工作进行展望。

*)本文受国家自然重点基金资助,基金号:60433030。徐明星 副教授。

2 基于时域的语音切分基本算法

本文提出的改进算法是基于文[1]提出的语音切分基本算法，下面简单描述该算法。

2.1 算法基本原理

该基本语音切分算法基于语音信号的短时性。语音信号随着时间的变化而变化，只有在一段相对短的时间内才有较一致的特性，这短段时间一般可取为 5～50 ms[2]。根据统计学原理，对计算出来的短时特征和短时参数进行分析，以区别语音信号和噪音信号。

实际应用中，选取 20 ms 作为一帧的长度，而以 10 ms 作为帧偏移。M 表示一帧语音短时平均幅度，Z 表示语音信号的短时过零率。

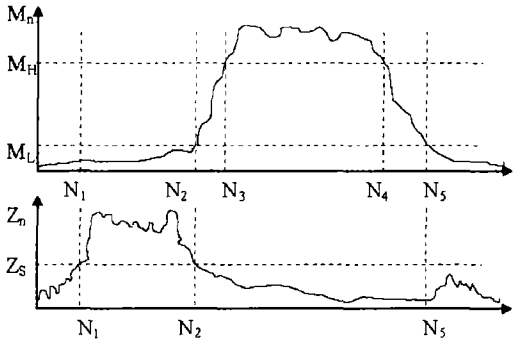


图 1 基本的语音切分方法示意图

基本算法的实现就是利用语音信号与静音信号的幅度与过零率存在的差别。如图 1 所示，首先设定 3 个阈值 M_H、M_L、Z_S，当信号的幅度大于 M_S 时，该点所在帧可以确定为语音帧，该点记为 N_3，但 N_3 还不能准确标记语音信号的起点，要向前寻找第一个等于阈值 M_L 值的点，记为 N_2，该点比较准确地确定了语音的起点。

但某些清音幅度也不高，跟静音相差不大，上述方法并不能很好加以区分，所以要根据静音和清音短时过零率的不同，进一步判断。换句话说，从 N_2 出发，向前搜索过零率刚好低于阈值 Z_S 三倍(经验值)的点，为了缩小偏差，向前搜索的最大时间定为 25 ms。找到的点就定为语音的起点，同理，可以确定语音的终点。

该算法的主要难点在于调整阈值 M_H、M_L、Z_S，不同的阈值对切分的准确度的影响是很大的。

2.2 基本算法的不足

采用基本算法能够较有效地分辨语音段和无声段，切分效果较好，但基本算法也有一些不足。主要表现在以下几个方面：

(1)对不同噪音环境的适应能力不高，缺乏自我调节能力。该算法的实现依赖于阈值的设置，如果每次遇到新的噪音背景就要设置新的阈值。在复杂的情况下，时间效率往往低得不可接受，每改变一次信噪比的值或者噪音类型都要调整阈值，这是不现实的。

(2)对某些幅度较低的清音，切分效果很差。主要是因为，如果阈值 M_H 定得太低，则会将幅度稍高的噪音当作语音处理，从而插入许多错误的切分点，降低结果的可靠性；阈值 M_H 定得太高，则会忽略幅度稍低的语音，使得若干静音与语音连成一片，删除掉一些切分点。这两个问题往往不能同时解决。具体实现中，总是选择使总的错误率最低的阈值，这就造成了幅度低的清音被当成静音，归入静音段，造成较大偏差。

(3)对于背景音出现的脉冲噪音，分辨能力不足。因为脉冲噪音的短时幅度大，能够达到阈值 M_H，所以基本的切分方法无法分辨其是否为语音，这就在原来没有切分点的地方插入错误的切分点。

3 改进的语音切分算法

针对基本算法的不足，本文提出了一种改进算法，新算法对不同的噪音类型和不同的信噪比有自适应能力，无需人来设定阈值；通过前后搜索调整切分点，改进了切分的性能；通过加入对短时脉冲噪音的检测与去除，降低了脉冲噪音的影响，下面对此改进算法进行详细说明。

3.1 算法阈值的自适应确定

基本算法的性能依赖于阈值的设置，在实验数据很多的情况下，每次都要靠人工修改阈值是很不现实的，工作量相当大。因此很有必要进行改进，以适应各种噪音的环境中。

在实际应用中，从开始录音到说话人说话可能有一定的延迟，前 0.5 s 很可能仅仅是背景音，该段语音或者是静音或者是噪音，并且在整个语音段中，这个背景音通常是相对稳定的。在录音开始后的 2 s 内，说话人一般已经开始说话了。基于这一事实，统计前 0.5s 背景音的短时参数和前 2s 语音段的短时参数，结合这些参数定出合适的阈值——M_H、M_L、Z_S，从而实现自适应。

用 $\hat{M}_n$ 表示背景音的最大帧幅度(下标 n 表示 noise)，$\hat{Z}_n$ 表示背景音的最大过零率，$\bar{M}_n$ 表示背景音的平均短时幅度，$\bar{Z}_n$ 表示背景音的平均短时过零率，$\bar{M}$ 表示前 2s 的平均短时幅度，$\bar{Z}$ 表示前 2s 的平均短时过零率，阈值 M_L、M_H、Z_S 具体计算如下：

$$M_L=\alpha\cdot\bar{M}_n+(1-\alpha)\cdot\hat{M}_n,\alpha=\frac{2}{3} \tag{3}$$

$$M_H=\beta\cdot\bar{M}+(1-\beta)\cdot\hat{M}_n,\beta=\frac{2}{3} \tag{4}$$

$$Z_S=\frac{1}{6}\cdot\bar{Z}+\frac{1}{12}\cdot\hat{Z}_n+\frac{1}{6}\cdot\bar{Z}_n \tag{5}$$

式中的常数项是权重，都是经验值，在实际应用中，能够取得较好效果。

该方法考虑到背景音的短时参数和语音的短时参数，根据各种情况，自动调节阈值，实验表明，对不同噪音的敏感度高，自我调节能力较强，从而提高总体的鲁棒性。

3.2 前后搜索调整切分点

定出阈值之后，根据基本切分算法初步定出的切分点，并不精确，因为在静音段中间可能出现清音，比如 f 字母的尾音[f]和 c 字母的始音[s]，幅度很小，很容易被误认为噪音，所以希望能够在原来定出的静音段进一步搜索，将清音帧找出。

清音和噪音明显不同在于清音的过零率往往很大，例如在采样频率 44100 Hz 的情况下，每帧(20 ms)清音的短时过零率达到 100～200。该算法就是利用这一特征，在初步定出的静音段起点向后、终点向前搜索，对于过零率很高的帧再进行分析。为了减少短时脉冲噪音的影响，搜索的帧长取 40 ms，帧移不变。这样的设置使得清音和静音过零率的差距更加明显，更有利于判别。具体实现如下：每次记录静音段的起点、终点，记为 N_S、N_E。从 N_S 开始，在不超过 N_E 的范围内，计算长度为 40 ms 帧的过零率，如果大于 Z_T，则认为是静音帧。其中阈值 Z_T 的计算公式如下：

$$Z_T=\begin{cases}260, \bar{Z}>\bar{Z}_n \\ 8\cdot Z_S, \bar{Z}\leqslant\bar{Z}_n\end{cases} \tag{6}$$

其中，$\bar{Z}$ 是语音段的平均过零率，$\bar{Z}_n$ 是背景音的平均过零率。260 是一个经验值，在背景音的平均过零率比语音的平均过零率低时，采用该经验值能够得到很好的效果；如果背景音的过零率很高，那么阈值 Z_T 取 $8Z_S$（常数 8 也是经验值），随着 Z_S 增大而增大，此时清音帧达到这样的值也是很困难的，所以该能力在背景音幅度高、过零率高的情况下，也不能很好区分清音帧和噪音帧。但另一方面，该阈值的设立有利于防止错误地向前后搜索。如果背景音过零率高，那么用于判断的值也高，背景音本身难以达到这样的值，所以能有效防止把过零率高的背景音也当作清音切分进语音段。此外，为了保证可靠性，规定前向搜索的长度不超过 0.2 s。从静音段的终点向前搜索清音帧也采用类似的方法。

3.3 短时脉冲噪音检测与去除

如前所述，基本算法对短时脉冲噪音的区分能力较弱。经过分析发现，短时脉冲噪音的持续时间短，跟语音段的持续时间存在明显差距。可以利用这一点进行滤除：在上述算法得出的静音切分段中，计算相邻静音段的距离，即计算前一段的结束标记与当前段开始标记的差，如果小于 0.16s，说明这两段中间很大可能不是语音信号，而是短时脉冲噪音。对这两段进行合并，前一段的开始标记作为当前段的开始标记，当前段的结束标记不变，并在输出文件中删除前一段的标注。

4 实验结果及评价

4.1 实验数据

实验的数据是女声发音的英文字母串，每个字母单独发音，每几个字母之间有稍长的停顿，比如“nb titi fg titu ny tary eigh”，共 200 句。格式为 WAVE 格式，双声道，采样频率为 44100 Hz，采样精度为 16 位，帧长 20 ms（882 个采样点），帧偏移 10 ms。使用 WaveSuffer 软件，手工将数据库中语音和静音之间的切分点标注出来，作为比照。

4.2 算法性能的评价

实验时，为了更好地进行判别，切分的对象是每几个字母之间的静音段，而不是切分至单个字母。因此，引入时间阈值 T_S。如果静音的长度小于 T_S，可以认为是两个音节中间的短暂停顿，归入语音段。

基本算法及改进算法均设置时间阈值 T_S 为 0.34s（实验数据中，单个字母之间的时间间隔均小于该值）。采用基本算法进行切分时，要先设置阈值。将语音信号的最大值量化为 1、幅度阈值 M_H、M_L 分别取 0.168 和 0.068，过零率阈值取 30。这些阈值是经过多次尝试定出的，能够使基本算法对原始语音的切分取得更低的错误率。改进算法有自适应能力，不需预先设定阈值。

对算法进行评价的参数采用错误率。设人工标注切分点 L，自动切分点 L'，定义误差 $\varepsilon=|L-L'|$。如果切分点相对于人工切分点的误差大于 ε，称为替换错误（substitution error），错误个数用 S 表示；如果人工切分点相应的位置没有切分点，称为删除错误（delete error），用 D 表示出错个数；如果在人工切分点的中间增加了切分点，称为插入错误（insert error），出错个数用 I 表示。人工切分点的总数记做 N。定义错误率（$Err\%$）如下：

$$Err\%=(S+D+I)/N\cdot 100\% \tag{7}$$

4.3 实验结果

对上述语音库进行实验，当误差 ε 取如下几个典型值（20 ms，40 ms，60 ms）时，实验结果如表 1 所示。实验数据表明，改进算法总体上的错误率低于基本算法，并且精度更高，在更低允许误差的情况下效果更好。

表 1 原始语音切分结果

	切分点错误率（%）		
	$\varepsilon<20$ ms	$\varepsilon<40$ ms	$\varepsilon<60$ ms
基本算法	20.65	8.46	5.31
改进算法	17.98	7.99	5.04

对具体的实验数据分析也表明，基本算法出现的问题得到了有效的解决：

（1）如图 2(a)所示，白色部分表示语音段，灰色部分表示静音段，中间（即切分点处）用直线分隔。箭头所指为 f 的尾音[f]，由于幅度小，基本算法将其切分在静音的区域里。而采用改进算法，前向搜索时将 f 的尾音[f]辨别出来，切入语音段。如图 2(b)所示，箭头指向的清音切分入语音段。

（2）如图 2(c)所示，在每几个字母之间原来是静音，但是由于中间出现了脉冲噪音，所以多切分出几段语音段（图中窄带空白处）。而改进算法能辨别出短时脉冲，将其归入背景音（图 2(d)）。

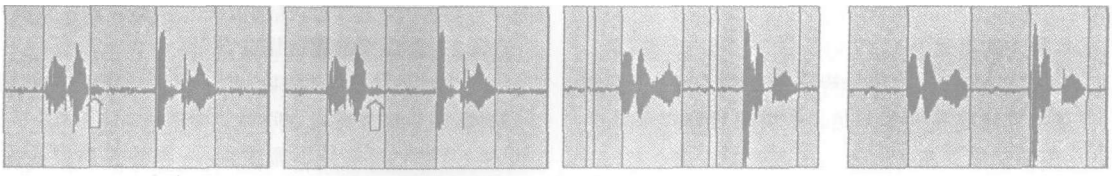


图 2 对于具体的实验数据，基本方法以及改进算法的切分效果对比：(a)基本算法对清音的切分的不足（箭头所示为清音）。(b)改进算法对清音的切分。(c)基本算法对脉冲噪音分辨能力不足。(d)改进算法对短时脉冲的处理。

4.4 新算法对噪音环境的鲁棒性

为了验证这两种算法在噪音环境下的表现，验证其鲁棒性，我们在原始语音中添加噪音，再进行切分。添加噪音的类型有：粉红噪音和白噪音，实验结果如图 3(b)(c)(d)所示，是采用基本算法和改进算法，分别切分带粉红噪音和白噪音的语音，在不同信噪比下错误率的曲线，(b)图是误差 $\varepsilon=20$ ms 的情况，(c)图是误差 $\varepsilon=40$ ms 的情况，(d)是误差 $\varepsilon=60$ ms 的情况。不同算法，不同噪音类型分别用不同的图标表示。

实验结果表明，改进算法在噪音环境下的表现远比基本算法好，信噪比越低，差距越大。对噪音环境下语音切分的过程中，基本算法的阈值统一采用上文所述，取在原始环境下切分效果较好的值，即 M_H 和 M_L 分别取幅度最大值的 16.8% 和 6%，Z_S 取 30。适当调节这些值会使基本算法的切分错误率降低，但是用于对比的实验数据太多，信噪比的不同值、噪

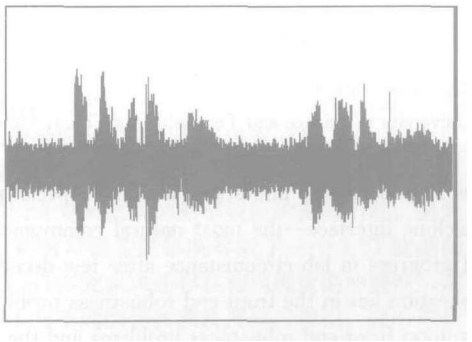
音的不同类型都要重新调整阈值，效率太低，所以这里选择统一的阈值，结果也仅供参考。

从图 3 中对比采用改进算法对带不同噪音类型的语音进行切分的错误率曲线。可以发现加入粉红噪音对切分的影响更大，信噪比越高，错误率越低，总体上呈线性下降趋势，而白噪音尽管过零率高，但是幅度分布均匀，切分算法能较好地处理这种情况，所以白噪音对切分性能的影响并不大，切分结果的错误率基本保持在一个较低的水平(10%～20%左右)。

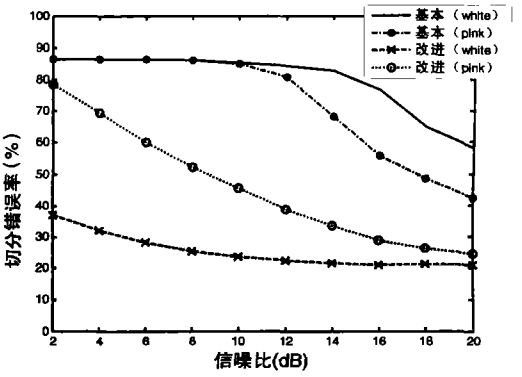
总结与展望 本文提出的改进算法，可以在一定程度上弥补基本切分算法的不足，能够有效辨别语音和背景音，包括幅度相差不大的清音和背景音，可以滤去短时脉冲噪音，有比较高的正确率。由于能够自适应调节阈值，因此对不同的噪音环境有较好的适应能力，尤其是在白噪音环境下，本文提出的新算法能够保持很好的性能，鲁棒性很高。

在实验中，我们也发现，改进算法还存在一些不足，如没有办法滤去持续时间较长(持续时间在 0.5s 以上)的脉冲噪音；当背景音过零率高于清音时，不能判别清音和背景音；而且没有考虑语音信号一开始就是声音，以及录音开始较长一段时间(2 s)仍然没有语音出现的情况。此外，调整时间阈值 T_S，并对该改进算法做相应调整，可以进一步提高切分精度，实现每个单词的切分。这些方面将是今后的努力方向。

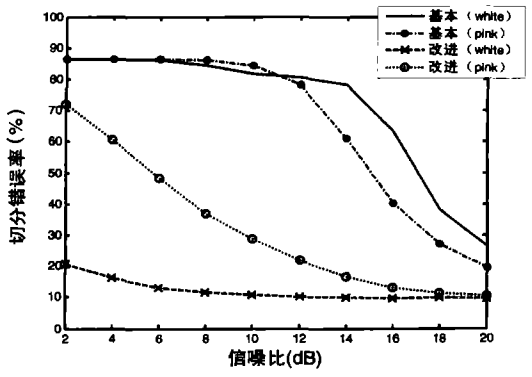
另一方面，可以分析、研究各种噪音对语音切分的影响因素，进一步提高算法的鲁棒性。文[10]提出一种利用基音周期和短时参数进行判断的方法，在结合基于时域的方法和基于频域的方法进行语音切分方面进行有益的尝试。这也是今后可以研究的课题。



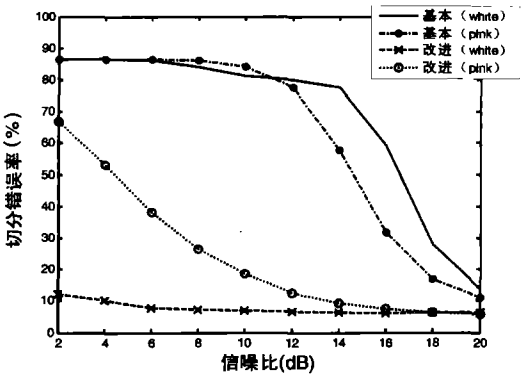
(a)



(b)



(c)



(d)

图 3 (a)是加入白噪音的语音在信噪比为 2 dB 情况下的波形图。(b)(c)(d)是采用基本算法和改进算法，分别切分带白噪音和粉红噪音的语音，在不同信噪比下错误率的曲线：(b)图是误差为 20 ms 的错误率；(c)图是误差为 40 ms 的错误率；(d)图是误差为 60 ms 的错误率。

参 考 文 献

1 Rabiner L R, Sambur M R. An Algorithm for Determining the Endpoint of Isolated utterances. Bell Syst. Tech. J., 1975, 54(2): 297～315

2 杨行峻，迟惠生. 语音信号数字处理. 北京：电子工业出版社，1995

3 张继勇，郑方，杜术，等. 连续汉语语音识别中基于归并的音节切分自动机. 软件学报，1999，10(11)

4 Burileanu D, Pascalin L, Burileanu C, et al. An Adaptive and Fast Speech Detection Algorithm. In: Sojka P, Kopeek I, Pala K, et al eds. Text, Speech and Dialogue: Third International Workshop, TSD 2000. Brno, Czech Republic: Springer-Verlag GmbH, 2000. 177～182

5 Beritelli F, Casale S, Cavallaro A. A robust voice activity detector for wireless communications using soft computing. IEEE J Sel Areas Commun, 1998, 16(9): 1818～1829

6 Beritelli F, Casale S, et al. Performance evaluation and comparison of G. 729/AMR/fuzzy voice activity detectors. IEEE Signal Process Lett, 2002, 9(3): 85～88

7 张文军，谢剑英，李聪. 基于贝叶斯方法的鲁棒语音切分. 数据采集与处理，2002，17(3)

8 Wu Bing-Fei, Wang Kun-Ching. Robust Endpoint Detection Algorithm Based on the Adaptive Band-Partitioning Spectral Entropy in Adverse Environments. IEEE Transactions on Speech and Audio Processing, 2005, 13(5)

9 Li K, Swamy M N S, Ahmad M O. An Improved Voice Activity Detection Using Higher Order Statistics. IEEE Transactions on Speech and Audio Processing, 2005, 13(5)

10 贾卓燕，申瑞民. 一种利用声音特性快速切分英文单词音节的算法. 计算机仿真，2005，22(2)

11 何致远，胡起秀，徐光佑. 说话人识别中的语音切分算法的研究. 计算机工程与应用，2003. 6