基于听觉感知的顽健语音流检测新方法

A New Method of Robust Detection for Speech Stream Based on Auditory Perception

申丽然 徐 东 李雪燧 张汝波

(哈尔滨工程大学计算机科学与技术学院 哈尔滨150001)

Abstract Based on the human auditory perceptual properties, a new method for speech detection is proposed. The new method consists of the classical RASTA-PLP filter in logarithmic spectrum domain and a new differential filter in time domain. In this manner, all kinds of noise especially impulse noise can be removed effectively. Proved by experiments, the new method in speech stream detection in adverse environments outperforms the conventional detection method.

Keywords Speech stream detection, Speech perception, Differential filter

引言

语音流检测在很多领域具有重要的意义,在大多数场合语音流检测是靠人工来完成的。长时间的侦听,尤其是在强噪声环境下会使人的听觉感到非常疲劳,这样一方面会造成对人耳的损伤,使听力下降;另一方面造成工作效率的下降,在很疲劳的状况下,很容易将某些语音信息漏掉。因此,用机器代替人进行语音流检测就显得非常必要了,它可以把语音段从噪声中分离出来,便于人的倾听和处理。

语音流检测是一个古老的问题,最早可以追溯到60年代[8],但至今没有得到完全的解决。最早的、最简单的方法是基于能量和过零率的方法。在背景噪声较小的情况下,用传统的方法我们可以比较容易地检测出语音流并进行端点检测。但是在现实中,语音常常被较复杂的噪声污染,如机器的运转声、汽车的引擎声等等。在这种情况下,传统的语音流检测方法性能开始恶化,端点检测变得十分困难,不准确的端点检测导致后续处理过程中的各种识别器性能的急剧下降。

人耳是一个良好的频谱分析仪,它能很好地分辨各种声音并对自己感兴趣的声音进行有效的提取。本文正是基于人耳的听觉特性[1]:临界带宽;等响曲线;听觉幂律;对各种噪声进行抑制。由于语音信号是由声道运动编码的,而非语音的干扰成份的变化速率通常位于声道形状变化的典型值之外。基于此,Hermanskey 提出语音信号经典的 RASTA 滤波技术[2]。这种方法对于卷性噪声有很好的抑制作用。

通常,经典的 RASTA 滤波器常常用于语音识别[3],本文首次把它用到了语音流检测上。实验表明,J-RASTA 滤波在强噪声下性能表现良好,唯一的不足是对脉冲噪声比较敏感。为了消除这种敏感,本文对经典的 J-RASTA 滤波进行了改进,在时域内对预处理信号进行差分平滑滤波,压制了脉冲噪声的影响。取得了良好的效果。

1 新的特征提取方法(D-RASTA-PLP 滤波)

1.1 平滑差分滤波器

这是一种把数据的平滑和差分相结合的算法。其目的是:在低频部分更好地接近最佳的差分滤波器($H_d(e^{jw})=jw$);而在高频部分具有较好的衰减,以期获得好的低通特性。在低频部分语音的信息得到了良好的保存,而在高频部分给予了平滑,从而脉冲噪声得到了抑制。这种性质正是语音信号处理所需要的。算法如图1所示。

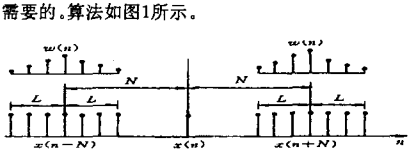

图1 平滑差分滤波器示意图

图中 $w(n)$是加权窗,它满足 $w(n)=w(-n)$;$\sum_{n=-L}^{L} w(n)=1$。差分滤波器的形式很多,本文用形式:$y(n)=(\Delta_1+\Delta_2)/6$移动平均 N=1,L=1,$\Delta_2=x(n+2)-x(n-2)$,$\Delta_1=x(n+1)-x(n-1)$。

本文采用差分滤波的目的是为了消除脉冲噪声的影响,消除脉冲噪声影响的方法很多。本文之所以采用这种形式,主要考虑在性能不是明显下降的情况下最小的消耗机器的时空资源。

1.2 经典 J-RASTA 滤波器

相对于语音信号,如果通信信道的频率特性是不变或缓慢变化的,则可用带通的 RASTA 滤波器滤除缓变化的信道因素,如图2所示。

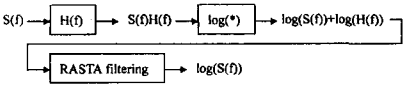

图2 经典 RASTA 滤波示意图

信道滤波器 H(f)在对数功率谱域同语音功率谱是可分离的(同态滤波)。因此,Hermansky 的经典 RASTA 滤波的计算包括以下步骤:

1)利用滤波器组计算听觉谱。如本文采用的19通道 PLP 滤波器组如图3:

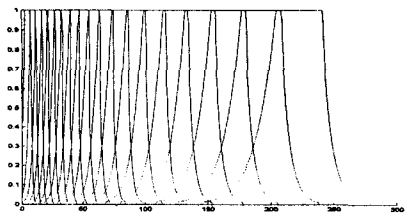

图3 19通道 plp 滤波器响应

2)对听觉谱幅度非线性压缩(如对数,$y=\log(x)$)

3)对数听觉谱通过 RASTA 滤波器,RASTA 滤波器的传递函数为

$$H(z)=0.1z^4*\frac{2+z^{-1}-z^{-3}-2z^{-4}}{1-0.98z^{-1}}$$

滤波后的对数听觉谱作余玄变换。

已经知道[2],当语音信号同时受到卷积噪声和加性噪声的污染,经典的 RASTA 滤波不能很好地工作,因为噪声与语音信号在对数功率谱域是不可分离的。为此,Hermansky 等提出 J_RASTA 算法,其核心是以下式替换步骤2)中的对数运算。

$$y=\log(1+J\cdot x)$$

其中 J 为与信号相关的常数,当 $J.x\gg1$时,上式近似于对数压缩,当 $J.x\ll1$时,上式近似于线性变换。Hermansky 等发现 J 的值与语音信号的 SNR 相关,当语音 SNR 较高时,应使 $J.x\gg1$才能获得较高的识别率,当语音 SNR 较低时,应使 $J.x\ll1$得到较好的识别性能[3]。也就是说,对低信噪比语音 RASTA 滤波滤除了加性噪声,对高信噪比语音则滤除了卷积噪声。

1.3 本文所提出的 D-J-RASTA-PLP 滤波

基于对经典 RASTA 的深入研究,本文对 RASTA 做了改进。

(1)在时域内进行了平滑滤波。

(2)对经典的 RASTA-PLP 次序做了调整,使其更接近人的听觉。

这一改进,在不失真的情况,不但充分保留了经典 RASTA 的优良特性而且使脉冲噪声得到了很好的抑制。具体流程如下:

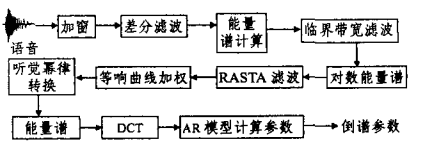

图4 D-J-RASTA-PLP 滤波流程

由上图可以看出 D-J-RASTA-PLP 滤波不但很好地反映了人耳的听觉特性,而且很好地应用了语音的声学的物理特性,为以后的各种操作打下了良好的基础。

2 D-J-RASTA-PLP 特征提取和语音流检测试验

2.1 试验准备

试验基于现场实录真实的无线话带通信语料,其中包括中、英等多个语种。采样频率为11025Hz,窗函数采用 Hamming 窗。帧长度为23.2ms,帧移为16.6ms。每帧产生5阶的倒谱参数。

2.2 实验结果

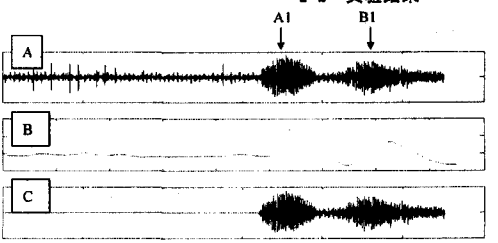

图5 D-J-RASTA-PLP 对脉冲噪声的抑制

根据倒谱参数，判定语音流的有无采用最差相似度准则。它的基本思想是人耳对语音的感知依靠当前和过去声音的比较，根据这个性质，可以求出相邻帧的差异，据此判断语音流的出现。D-J-RASTA-PLP 倒谱有效地提取了语音的参数，可以很方便地根据这些参数计算出相邻帧的差异度。即：$p(C_i)=\frac{1}{5}\sum_{p=1}^{5}(C_p^i-C_p^{i-1})^2$，其中 C_p^i 为第 i 帧的第 p 阶倒谱。根据研究工作的需要，我们需要对多语种、多信噪比的语音进行检测。图5是一些试验结果。

图5表明 A 是1s 的原始语料，从中可以看出有很多的脉冲噪声。A1出为英文 'O'，B1为英文 'K'。图 B 是 D-J-RASTA-PLP 倒谱参数，在相应的脉冲噪声处倒谱参数起伏很小。而真正的语音处，到谱参数却变化的相当有规律，很好地体现了语音的谱包络。图 C 是对语音流的提取。

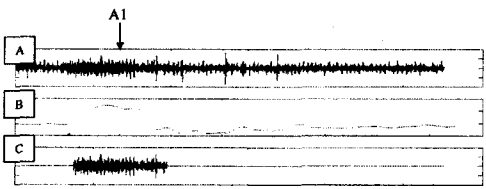

图6 D-J-RASTA-PLP 信噪比较差的情况下语音流的提取

图6表明 A 是1s 的原始语料，其中包含复杂的背景噪声，仅在 A1处有微弱的语音流出现。传统的方法很难检测出。但在图 B 中我们可以很清楚地看到有语音流出现是倒谱参数有了明显的变化。为语音流的提取提供了良好的特征。图 C 是根据倒谱参数提取的语音流。

通过以上两个例子，我们可以看到，本文提出的方法可以很好地抑制噪声，与人的听觉特性相一致，当环境中出现语音时，倒谱参数就有相应的变化，正如人耳听到有人说话会不由自主地倾听一样。

讨论 目前在语音信号处理中主要关注的问题之一是如何使在干净环境下表现良好的系统，在恶劣环境下也有突出的表现。我们相信造成系统性能恶化的原因之一是环境中不同频率噪声的影响。为了消除这种影响出现了很多优秀的去噪方法，像小波去噪、谱减法、Kalam 滤波等等。这些方法在一定程度上表现非常优秀。但它们往往忽略了听觉的特性，无论如何听觉在语音的交流中起到了不可替代的作用[6]。本文对听觉特性给予了足够的重视，并取得了良好的检测效果。但正如所有的事物一样，优秀只是相对的，对每一个系统来说都有它的不足所在，本系统最大的不足就是信噪比驱动下 J 值的确定。只有在 J 值比较合适的情况下，系统才能得到比较满意的效果。在实验中，用自适应的方法对信噪比进行了实时估计，取得了不错的效果，但这也付出时空上的代价。期待有一种良好的解决环境信噪比的方法，以使系统性能更优。总之，无论从抗噪性来看还是从应用的角度来看，D-J-RASTA-PLP 有很好的应用前景。

参 考 文 献

1 Hermansky H. Perceptual Linear Predictive (PLP) analysis of speech. Journal of Acoustical Society of America, 1990, 87(4): 1738～1752

2 Hermansky H, Morgan N. RASTA processing of speech . IEEE Trans. Speech and Audio processing, 1994, 2(4): 578～589

3 Vuurens, Hermansky H. Data-drivedesign of RASTA-like filter. Proc. Euro-speech, Rhodes, Greece, 1997

4 Kanedera N, Arai T, Hermansky H, Pavel M. On the importance of various modulation frequencies for speech recognition. In: Proc. of EUROSPEECH'97, Rodos, Greece, 1997

5 Hirsh H. Estimation of noise spectrum and its application to SNR estimation and speech enhancement: [Technique report TR-93-012]. issi uc Berkeley, 1993

6 Hermansky H. Should recognizer have ears?. Speech communication, 1998, 25(1-3): 3～27

7 Greenberg S. The ears have it: the auditory basis of speech perception. Proc. IcphS-95, 1995, 3: 34～44

8 Rabiner L R, Sambur y M R. An Algorithm for Determining the Endpoints of isolated utterance. The Bell System Technical Joural, 1975, 54(2)

(上接第57页)

泛应用于桌面视觉、物体建模、视觉导航等领域。

参 考 文 献

1 赵文伯，刘俊刚. CMOS 图像传感器发展现状. 半导体光电，20，(1)：11～18

2 Don Anderson 著，姜汉龙等译. FireWire 系统体系. 中国电力出版社，2001

3 杨云飞. 多目立体视觉图像获取和三维恢复技术：[北京理工大学硕士学位论文]. 2001

4 1394-based Digital Camera Specification, Version 1. 20. July 23, 1998, 1394 Trade Association