小波变分辨率频谱特征静音检测和短时自适应混音算法

薛 卫[1] 都思丹[2] 叶迎宪[2]
(南京农业大学计算机系 南京 210002)[1] (南京大学电子科学与工程系 南京 210093)[2]

摘 要 静音检测算法使用两种语音感觉特征与变分辨率频谱的Mel频率倒谱系数组合成音频特征,采用多门限过零率对静音进行初判,并通过二分类支持向量机对组合语音特征进行分类;实时混音算法使用每一路音频的短时能量作为混音权重。测试表明,静音检测算法在不同信噪比下语音识别正确率高于G.729b静音检测算法;实时混音算法听觉测试优于传统的算法,并且混音计算延时低,满足网络实时传输的要求;两种算法同时应用于视频会议系统,视频会议服务器的运算量低于使用了G.729b静音检测算法的视频系统。

关键词 静音检测,小波,支持向量机,短时自适应权重
中图法分类号 TP391.42 **文献标识码** A

Voice Activity Detection Using Wavelets Multiresolution Spectrum and Short-time Adaptive Audio Mixing Algorithm

XUE Wei[1] DU Si-dan[2] YE Ying-xian[2]
(Department of Science and Technology, Nanjing Agricultural University, Nanjing 210002, China)[1]
(Department of Electronics Science and Engineering, Nanjing University, Nanjing 210093, China)[2]

Abstract The proposed VAD uses MFCC of multiresolution spectrum and two classical audio parameters as audio feature, and prejudges silence by detection of multi-gate zero cross ratio, and classifies noise and voice by Support Vector Machines. New speech mixing algorithm used in Multipoint Control Unit(MCU) of conferences imposed short-time power of each audio stream as mixing weight vector, and was designed for parallel processing in program. Various experiments show, proposed VAD algorithm achieves overall better performance in all SNRs than VAD of G.729b and other VAD, output audio of new speech mixing algorithm has excellent hearing perceptibility, and its computational time delay is small enough to satisfy the needs of real-time transmission, and MCU computation is lower than that based on G.729b VAD.

Keywords Voice activity detection, Wavelet, SVM, Short-time adaptive weighted

1 引言

静音检测(Voice Activity Detection,简称VAD)与实时混音技术是网络语音交流平台的重要技术。静音检测在语音处理中有重要的作用:在语音识别中有助于提高识别率;在用户接口中消除回声提高合成语音的音质;在通讯系统中应用静音检测可以降低系统平均传输率。一般VAD算法通过提取音频信号特征值与预先设定好的门限值的比较来确定静音。早期VAD使用的参数包括短时过零率、短时能量、自相关系数和LPC[1-3]等,但语音信号和某些背景噪声信号具有非平稳性,故以这些参数作为特征的检测系统识别率都不高。

Guo Guodong等[4-6]提出用小波分解的方法将信号按频段进行分解,然后在各个频段提取不同参数。然而由于未考虑信号的整体频谱,其识别率不高。目前常用的一些静音检测方法(如GSM[7],G.729B[8]等)会将部分气流噪声识别为正常语音。为了克服这些问题,本文提出基于小波变分辨率频谱(wavelets-based multiresolution spectrum,简称WBMS)的MFCC特征[9],并且利用支持向量机(Support Vector Machines,简称SVM)实现静音检测。

在具体应用中,计算机能表示的声强范围非常有限,因此一般的实时混音算法不可避免地会出现叠加结果溢出[10],从而产生语音的失真。文献[11]中给出了几种改进的混音算法:平均调整权重法、强对齐权重法、弱对齐权重法和自对齐权重法,前3种算法存在混音后音量降低和求和溢出的缺点,自对齐权重法虽然有效避免了这些问题,但由于单路声音所乘的权重是一个随时间变量变化的值,会使得混音后语音功率谱分布变得较为分散,引入噪声,特别是混音路数达到4时。为了解决这些不足,本文提出了一种短时自适应权重(short-time adaptive weighted,简称SAW)混音算法。

2 结合静音检测的通信系统架构

本文研究的通信系统架构将静音检测、实时混音技术同

到稿日期:2008-08-29 返修日期:2009-01-08 本文受国家自然科学基金(60472026)资助。
薛 卫(1979—),男,博士,讲师,主要研究方向为音视频处理技术、信息安全、嵌入式系统,E-mail:xwskyzq@sina.com;都思丹 女,教授,博士生导师,研究方向为信号处理等。

时应用在基于集中混音策略[12]的多媒体通讯架构，音频编解码器基于G.729AB实现。G.729是ITU于1995年制定的使用CS-ACELP(共轭结构代数码激励线性预测)的8kbit/s语音编码标准。G.729A是G.729的附件A，是G.729语音编码标准减少算法复杂度的版本。G.729B是G.729的附件B，主要描述了VAD(活动话音检测)的静音压缩算法。本文系统客户端语音编码器首先检测语音，如果语音静止，就停止发送数据，以减少网路传输的开销。然而如果出现同一时刻所有通信端完全不发声的情况，就会使听者觉得通信中断，因此需要人为地加入一些噪声，使听者觉得通信没有中断。本文采用G.729B自适应噪声生成(CNG)技术，在编码器检测到输入数据有语音时，输出是正常的80比特的帧，而当编码器在话音段后面第一次检测到静音时，它会产生一个15比特的静音描述帧，告诉解码端当前噪声的情况。如果背景噪声的特征没有发生明显的变化，编码器就持续产生空帧，直到背景噪声的特征发生了明显的改变，则重新发送静音描述帧，或者检测到有话音了，则切换到发送正常语音帧。

系统的混音器在多点控制单元(Multipoint Control Unit，MCU)对语音信号进行混音，音频混合单元从各个通讯终端取得信号，进入混音器前判断其是否为正常语音，并且只对同一时刻正常语音包解码混音；如果出现同一时刻流入的数据包均为静音，系统随机选择其中一路直接发送到接收者。通过以上策略，系统降低了网络传输的负担；将混音计算集中到MCU，与分布式混音相比减少了客户端的运算量，同时，由于采用了静音检测技术，也从一定程度上降低了MCU的混音运算负担。

3 静音检测

G.729b等协议的静音检测存在一定的误判，当发言者不发言，且周围无其它噪声源时，仍有30%左右的音频帧被发送，表明较弱的背景声音被判断为正常语音，如与话筒接近的人鼻孔或嘴里呼出的气流产生的信号。本文提出用多门限过零率静音检测与二分类SVM[13]来区分出气流等外界引起的声音(伪发音)，这里将语音分为三类：静音、伪发音和正常语音。

3.1 音频特征参数提取

静音检测的参数包括两部分：感觉特征、基于小波变分辨率频谱的MFCC。

1)感觉特征

Zci：Zci为第i帧的过零率。

Ei：Ei为第i帧的短时能量值。

2)基于变分辨率频谱的MFCC

MFCC从人耳对频率高低的非线性心理感觉角度反映了语音短时幅度谱的特征，其识别效果要优于传统的线性预测倒谱参数，但在信噪比较低时，识别效果并不理想。而小波变换在静音检测中的难点是如何将小波变换系数转换成高效的特征参数，Stegmann，Lin[5,6]尝试从不同的频段分别提取特征，但这样做会忽略语音的整体频谱特征，因此我们将分解后的各个分辨率下的频谱系数再拼接成完整的频谱，从小波变换的压缩特性[14]我们知道，拼接后的音频信号小波系数将是稀疏的；同时，小波变换也会去除音频中一部分噪声谱，提高了信号的信噪比。因此，我们提出基于变分辨率频谱的MFCC，以分解合并后的小波频谱作为MFCC的输入。

变分辨率频谱提取流程如图1所示，对时域语音信号采用Daubechies4小波包[15]变换把加窗信号分解成N个子带的系数，在各个子带进行重构至第一次小波分解后系数尺寸，并对各子带系数进行归一化处理，随后对系数作FFT变换，将各子带系数求和组成变分辨率频谱，最后将变分辨率频谱送交MFCC提取模块。

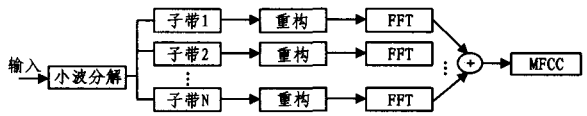

图1 WBMS计算流程

Mel尺度倒谱系数(CMFCC)计算公式如下：

$$c_{MFCC}(i)=\sqrt{\frac{2}{L}}\sum_{l=1}^{L}\log m(l)\cos\left\{(l-\frac{1}{2})\frac{i\pi}{L}\right\} \tag{1}$$

其中：

$$m(l)=\sum_{k=o(l)}^{h(l)}W_l(k)\,|X_n(k)|\quad ,l=1,\ 2,\cdots,L \tag{2}$$

$$W_l(k)=\begin{cases}\dfrac{k-o(l)}{c(l)-o(l)} & o(l)\leqslant k\leqslant c(l)\\ \dfrac{h(l)-k}{h(l)-c(l)} & c(l)\leqslant k\leqslant h(l)\end{cases} \tag{3}$$

$o(l)$，$c(l)$和$h(l)$分别是l个三角形滤波器的下限、中心和上限频率。

3.2 多门限过零率[17]纯静音检测

本文采用多门限过零率对静音进行预判，目的是通过简单的运算去除明显静音。多门限过零率检测法设3个高低不同的门限，$T_1<T_2<T_3$，对每一帧用式(4)分别求相应于T_1，T_2，T_3的3种门限过零率Z_1，Z_2和Z_3。

$$Z_n=\sum\{|\mathrm{sgn}[x(n)-T_n]-\mathrm{sgn}[x(n-1)-T_n]|+|\mathrm{sgn}[x(n)+T_n]-\mathrm{sgn}[x(n-1)+T_n]|\}*w(n-w) \tag{4}$$

总过零率：

$$Z=W_1Z_1+W_2Z_2+W_3Z_3$$

其中：W_1，W_2，W_3为过零率权值；Z为过零率和。Z_0定义为总过零率分界值。当$Z>Z_0$时，判为语音帧；否则，判为静音帧。

3.3 静音检测算法

静音检测流程如图2所示，提取一帧音频数据的复合参数，使用多门限过零率检测纯静音，如果小于Z_0，就判断其为静音，否则由二分类SVM分类器判断出噪声、正常语音，本文将伪发音、预判中遗漏的静音归为噪声。

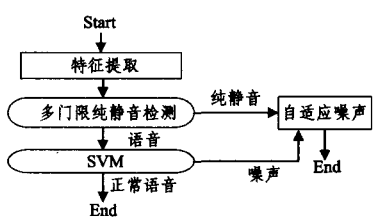

图2 静音检测流程

4 短时自适应权重混音算法

定义权重$w(j)$，每K个音频帧变化一次。对于语音8kHz的采样频率，以80点作为一帧。首先计算每路声音在

K 个数据帧中的平均幅度值：

$$Avg(j)=\frac{1}{Kl}\sum_{i=0}^{Kl-1}|data(j,i) \tag{5}$$

其中，$data(j,i)$表示第 j 路语音的第 i 个样本值，字母 l 代表一个数据帧中声音的样本数。然后根据 $Avg(j)$计算出第 j 路语音应占有的权重 $w(j)$：

$$w(j)=\frac{Avg(j)}{\sum_{p=0}^{n-1}Avg(j)} \tag{6}$$

然后根据 $w(j)$对声音进行混合：

$$MixData(i)=\sum_{j=0}^{n-1}data(j,i)*w(j) \tag{7}$$

其中：

$$\sum_{j=0}^{n-1}w(j)=1 \tag{8}$$

注意到算法中各路的 $Avg(j)$的计算是相互独立的，可以设计出高度并行化的计算结构，并且用 MMX，SSE，SSE2 指令集对程序进行优化。

5 实验结果

实验由 4 部分组成，首先是用复合音频特征对 SVM 分类器进行训练，并用新的静音检测算法对数据进行检测，主要对其误检率与 G. 729b 静音检测进行比较，同时也对新特征与传统 MFCC 特征识别结果进行比较；对 SAW 混音算法进行测试，测试的重点是在 MCU 多路音频解压、混音、再压缩共消耗时间；另外也对混音算法和 MCU 性能进行了测试。

本文采用 kHz 的音频采样频率，以 80 点作为一帧进行检测，每一帧 10ms。静音检测用的所有训练和测试数据来自“Aurora 2”数据库，这些数据分别被附加了实际噪声，另外采集 20 个人每人 60 秒的语音用于测试混音。

5.1 音频特征检测

多门限过零率检测纯静音环节需要确定最佳权重向量和门限值，具体做法是，对于训练用数据，以多门限过零率检测产生的静音误判率为目标函数，遍历每一个权重向量和门限值取值范围，找出产生误判率最低的权重向量和门限值，这就是最佳权重向量和门限值。

音频特征由两部分组成：感觉特征、变分辨率频谱的 MFCC，其中感觉特征两个，MFCC 特征为 $L=12$ 个，SVM 的内积函数选用径向基函数（$\sigma^2=0.3$），训练采用 SMO 算法[17]。表 1 中数据为本文静音检测算法、MFCC＋SVM、G. 729B VAD在不同噪声信号下的测试结果，总体来看本文的静音检测算法在语音检出率、气流音检出率、误检率都是最优的，当背景噪声增大后，其气流音检出率以比较大的速度降低，但仍可检出 50%，相对于其它算法有一定优势。P_d 为正常语音正确检出率，P_s 为气流噪声正确检出率，P_c 为纯静音（不包括气流噪声）误检率（将纯静音判为正常语音的比率），P_f 为 G. 729B VAD 算法静音误检率（将静音、气流噪声误判为正常语音的比率）。

表 1 本文 VAD 与 MFCC+SVM VAD，G. 729B VAD 比较

Noise type	SNR (db)	Proposed VAD			MFCC+SVM			G. 729B VAD	
		Pd(%)	Ps(%)	Pc(%)	Pd(%)	Ps(%)	Pc(%)	Pd(%)	Pf(%)
白噪声	25	99.81	95.32	1.48	99.72	88.92	2.36	99.72	16.3
	15	98.47	89.18	2.18	97.43	85.17	3.45	96.61	23.86
	5	94.63	50.28	2.52	91.28	38.48	4.16	85.05	36.75
会议	25	99.71	93.45	1.72	99.61	87.44	2.52	99.53	19.42
	15	98.26	90.36	2.25	97.27	84.53	3.65	95.76	27.57
	5	93.52	52.63	3.12	90.56	38.23	4.28	84.56	41.83

5.2 混音器性能测试

测试不同混音路数时混音器的实时运行特性，实验时，$k=10$，每路音频持续时间长度为 60s。分别对 3－20 路音频进行混音，音频解、压均采用 G. 729a 算法，测试结果为平均每次混音时间，如图 3 所示。商业上衡量视频会议是否明显延迟的时间一般是在 0. 5 毫秒以内，所以本文提出的混音算法是高效实时的，而且从目前文献、专利以及商业系统发布的数据来看，支持 16 路以上混音的系统还未使用。

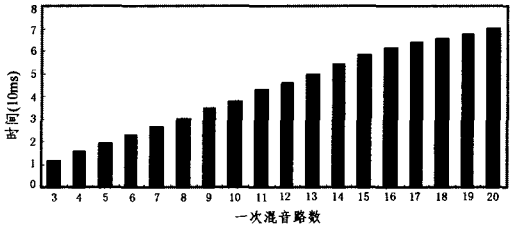

图 3 基于自适应权重混音的时间消耗图

5.3 混音算法测试

为了对 SAW 混音测试有直观的判断，本文给出了两路音频处理结果，并且在混音前不过滤掉其中的静音。根据测试人员的主观评价，混音效果明显好于平均调整权重法、ASW[11]等的输出，混音输出的波形保留了两路音频的波形特征。合成后的音频流连续，没有跳音和断续的感觉，也没有爆破噪声，ASW 合成后的音频流有一定的嘈杂感，并且会随着混音路数的增加而加剧，如图 4 所示。

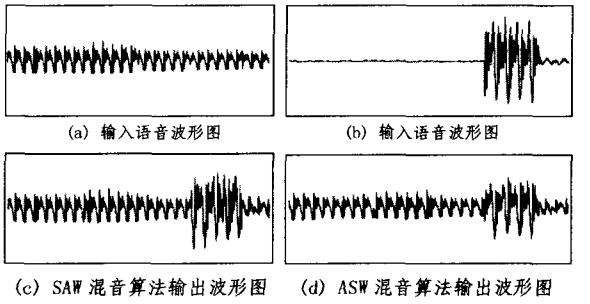

图 4 两路音频处理结果图

5.4 MCU 性能测试

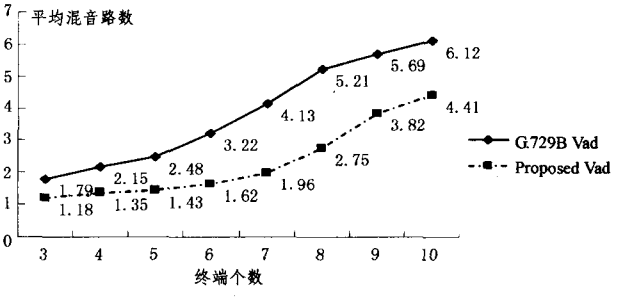

图 5 混音器每次混音时的混音路数均值

基于文中提出的静音检测、混音算法开发出一套视频会议系统，在系统编解码协议、混音算法均相同的情况下，测试 G. 729B 静音检测和本文的静音检测算法对 MCU 混音器性能的影响。结果表明，在不同会议发言端参与的情况下，本文

提出的静音检测算法相对于 G.729B 静音检测使得 MCU 的混音路数明显减少，从而减少了 MCU 的运算量，并允许更多的客户端参加语音讨论，如图 5 所示。

结束语 本文提出了结合新的静音检测技术的集中式混音系统。在新的静音检测算法中，采用变分辨率频谱 MFCC 参数及两个感觉参数作为语音特征，为了将纯静音、气流产生的伪发音与正常语音区别开，使用多门限过零率检测首先对纯静音进行预判，然后用支持向量机对语音特征进行分类。相比于 G.729B 中的静音检测技术和基于 MFCC+SVM 静音检测技术，在噪声比较大的情况下仍能达到比较高的语音识别率。

对于混音器的重要组成部分，本文采用 SAW 混音算法对各路解压音频进行混音，听觉测试比较优秀，并且通过对算法结构进行优化，获得了比较低的混音延时，即使是对 20 路音频混音，其运算延时仍很低，可以满足实时传输的要求。将新的静音检测技术与 SAW 混音技术使用在视频会议系统后，MCU 的混音计算量大大小于采用 G.729B 静音检测技术的视频会议系统 MCU，而且设备条件也允许更多的客户端连接服务器参加语音讨论。

参 考 文 献

[1] Nemer E, Goubran R, Mahmoud S. Robust voice activity detection using higher-order statistics in the LPC residual domain [J]. IEEE Transactions on Speech and Audio Processing, 2001, 9: 217-231

[2] Junqua J C, Reaves B, Mak B. A study of endpoint detection algorithms in adverse conditions: Incidence on a DTW and HMM recognize[C]// Proc. Eurospeech'91. 1991: 371-1374

[3] Sangwan A, Chiranth M C, Jamadagni H S, et al. VAD techniques for real-time speech transmission on the Internet[C]// IEEE International Conference on High-Speed Networks and Multimedia Communications. 2002: 46-50

[4] Guo Guodong, Li S Z. Content-Based Audio Classificationand Retrieval by Support Vector Machines[J]. IEEE Trans. on Neural Networks, 2003, 14(1): 209-215

[5] Stegmann J, Schroeder G. Robust Voice Activity Detection Based on the Wavelet Transform[C]// Proc. IEEE Workshop on Speech Coding. September 1997: 99-110

[6] Lin Chien-chang, Chen Shi-huang, Truong T K, et al. Audio Classification and Categorization Based on Wavelets and Support Vector Machine[J]. IEEE Transactions on Speech and Audio Processing, 2005, 13(5): 644-651

[7] ETSI: Draft Recommendation prETS 300 724: GSM Enhanced Full Rate(EFR) speech codec, 1996

[8] ITU-T: Draft Recommendation G.729, Annex B: Voice Activity Detection, 1996

[9] Rabiner L, Juang B H. Fundamentals of Speech Recognition. Englewood Cliffs[M]. NJ: Prentice-Hall, 1993

[10] Agustín JG, Hussein AW. Audio mixing for interactive multimedia communications[C]// JCIS'98. Research Triangle, NC, 1998: 217-220

[11] Fan Xing, Gu Wei-kang. Research on fast real-time adaptive audio mixing in multimedia conference[J]. Journal of Zhejiang University Science, 2005, 6a(6): 507-512

[12] Venkat RP, Harrick MV, Srinivas R. Communication architectures and algorithms for media mixing in multimedia conferences [J]. IEEE/ACM Trans. on Networking, 1993, 1(1): 20-30

[13] Cortes C, Vapnik C. Support Vector Networks [J]. Machine Learning, 1995, 20: 273-297

[14] Cvetkovic Z, Vetterli M. Discrete-time Wavelet Extrema Representation Design and Consistent Reconstruction [J]. IEEE Trans. SP, 1995, 143: 681-693

[15] Daubechies I. Ten Lectures on Wavelets. SIAM, Philadelphia, 1992

[16] Thomas Parsons W. Voice and Speech Processing[M]. McGraw-Hill Book Company, 1986

[17] Platt J C. A Fast Algorithm for Training Support Vector Machines[R]. MSR-TR-98-14. April 1998

(上接第 152 页)

通过合成数据集的对比实验，验证了 IR²-Tree SJKS 算法的性能优越性。

参 考 文 献

[1] Yiu Man Lung, Dai Xiangyuan, Mamoulis N, et al. Top-k Spatial Preference Queries[C]//Proceedings of ICDE. 2007: 1076-1085

[2] Shekhar S, Chawla S. Spatial Databases: A Tour [M]. Prentice Hall, 2003

[3] Grossman D A, Frieder O. Information Retrieval: Algorithms and Heuristics [M]. Springer, 2006

[4] Manning C D, Raghavan P, Schutze H. Introduction to Information Retrieval[M]. Cambridge University Press, 2007

[5] Brinkhoff T, Kriegel H-P, Seeger B. Efficient Processing of Spatial Joins Using R-trees[C]// Proceedings of SIGMOD. 1993: 237-246

[6] Lo M-L, Ravishankar C V. Spatial Joins Using Seeded Trees[C]//Proceedings of SIGMOD. 1994: 209-220

[7] Huang Yun-Wu, Jing Ning. Spatial Joins Using R-trees: Breadth-First Traversal with Global Optimizations [C]// Proceedings of VLDB. 1997: 396-405

[8] Lo M-L, Ravishankar C V. Spatial Hash-Joins[C]//Proceedings of SIGMOD. 1996: 247-258

[9] Guttman A. R-Trees: a dynamic index structure for spatial Searching[C]//Proceedings of SIGMOD. 1984: 47-57

[10] De Felipe I, Hristidis V, Rishe N. Keyword Search on Spatial Databases[C]//Proceeding of ICDE. 2008: 656-665

[11] Faloutsos C, Christodoulakis S. Signature Files: An Access Method for Documents and Its Analytical Performance Evaluation [J]. ACM Trans. Inf. Syst, 1984, 2(4): 267-288

[12] Lee Dik Lun, Kim Young Man, Patel G. Efficient Signature File Methods for Text Retrieval [J]. IEEE Transactions on Knowledge and Data Engineering, 1995, 7(3): 423-435

[13] Larson P A. A method for speeding up text retrieval[C]//Proceedings of SIGMOD. 1983: 117-123

[14] Christodoulakis S, Faloutsos C. Design Considerations for a Message File Server [J]. IEEE Trans. on Software Engineering, 1984, 10(2): 281-210