基于改进特征值的语音分割算法研究

任新社[1],缪 华[2],马青玉[3]

(1. 南京师范大学 教育技术系,江苏 南京 210097)
(2. 解放军国际关系学院 教育技术中心,江苏 南京 210039)
(3. 南京师范大学 物理科学与技术学院,江苏 南京 210046)

[摘要] 随着网络技术和媒体应用的迅速发展,传统的文本检索已不能满足需要,视频检索由于数据量大而得不到应用,语音检索就显示出重要的研究价值. 一个语音序列由多种不同类型的语音片段构成,而每一种类型的语音往往又包含不同的意义,因此通过语音特征进行语音分段来实现语音检索是现代媒体数据进行检索的重要手段. 通过对语音信号每一帧的基本特征值与整个语音序列的平均基本特征值进行比较,得到一个改进的特征值,并利用 K-Nearest Neighbor 算法进行语音分割,结果表明基于改进特征值的语音分割算法能够有效提高语音分割的准确性.

[关键词] 语音检索,语音分割,改进特征值
[中图分类号] TN912.3 [文献标志码] A [文章编号] 1672-1292(2011)03-0073-05

A Speech Segmentation Algorithm Based on Improved Eigenvalue

Ren Xinshe[1], Miao Hua[2], Ma Qingyu[3]

(1. Department of Educational Technology, Nanjing Normal University, Nanjing 210097, China)
(2. Center of Educational Technology, PLA Institute of International Relations, Nanjing 210039, China)
(3. School of Physicical Science and Technology, Nanjing Normal University, Nanjing 210046, China)

Abstract: With the rapid development of internet technology and media application, text-based retrieval cannot satisfy the requirements and auditory-visual processing can not be applied for the large data amount, so the emergence of speech retrieval is particularly important. An audio clip usually consists of many different types of audio segments with different meanings; therefore, it becomes a new method to perform speech retrieval with audio segmentation for modern media based on audio eigenvalue. In the article, the basic eigenvalue of each audio frame is compared with the average eigenvalue of the entire audio clip and then the improved eigenvalue can be obtained for audio segmentation by using the K-Nearest Neighbor algorithm. The experimental results show that the proposed algorithm based on the improved eigenvalue can efficiently improve the accuracy of audio segmentation.

Key words: speech retrieval, speech segmentation, improved eigenvalue

随着网络技术的不断发展,网络中的影音媒体数据迅速增长,约占网络信息总量的 20% 左右[1]. 据统计,国际知名网站 YouTube 每分钟上传视频的播放时长为 24 h,国内的优酷、土豆等网站也处于蓬勃发展阶段[2]. 对于如此巨大的影音数据,目前检索基本上还是采用基于标注的检索方式[3],如对音频数据标注为“某人的歌曲”、“某人的演讲”等. 由于基于人工标注方式的不完整性和主观性,人们很难快速找到满足具体要求的音频片段,同时人工标注不能解决听觉信息量迅速增长和对实时音频数据流进行检索等问题. 视频检索由于数据量过大而无法得到应用,而基于内容的语音检索技术通过对音频特征的分析,利用音频的幅度、频谱等物理特征,响度、音高、音色等听觉特征,词字、旋律等语义特征,实现基于内容的音频信息检索[4],这样既避开了人工标注的环节,同时也提高了视频检索的性能.

语音信号往往由多种不同类型的音频信号组成,例如说话音、环境音、音乐及噪声等. 语音检索一般针对某一特定类型的音频信号进行,首先需要将语音信号按照不同的类型进行分类,也就是语音分割,而后

收稿日期:2011-05-18.
基金项目:国家自然科学基金(10974098)、江苏省科技厅自然科学基金(BK2009407)和教育部博士点基金(20093207120003).
通讯联系人:马青玉,博士,教授,研究方向:声学技术和生物医学电子技术. E-mail:maqingyu@njnu.edu.cn

对分割的语音信号进行检索,这样可以大大提高检索的效率和准确度.

目前的语音分割主要基于简单的语音特征,如过零率、短时能量、线性预测系数(LPC)等,提取语音特征,并利用聚类算法对具有相同声学特征的语音进行聚类[5]. Saunders 提出了一种基于短时能量和过零率的从广播节目中分离音乐和说话音的方法[6],当窗长为 2.4 s 时,分割的准确率高达 98%. Scheirer 应用多种特征值和分类模型[7],如 GMM(Gaussian Mixture Model)、BP-ANN(Back Propagation Artificial Neural Network)和 KNN(K-Nearest Neighbor),当使用 2.4 s 窗长时,分割的错误率仅为 1.4%. 但对于更多种类的语音信号,这些简单的特征值不能满足语音分割的需求,因此找到区分度更大的新特征值对语音分割工作具有重要意义.

本文首先对不同类型的语音信号的基本特征值进行对比和分析,将语音信号每一帧的基本特征值与整个语音序列的平均基本特征值进行比较,得到一个改进的特征值,然后使用 KNN 算法进行分割实验,来验证改进之后的特征值能够提高分割的准确率.

1 改进特征值分析

1.1 短时过零率(Zero-Crossing Rate)的改进

语音信号的过零率是指单位时间内信号波形穿过横轴(零电平)的次数. 抽样后的语音信号是离散的时间序列,过零是指序列取样值改变符号,过零率则是指相对每个样本的改变符号的次数,因此定义如下:

$$\mathrm{ZCR} = \frac{1}{2(N-1)} \sum_{m=1}^{N-1} \left| \mathrm{sgn}[x(m+1)] - \mathrm{sgn}[x(m)] \right|. \tag{1}$$

式中,sgn[·]为符号函数. 语音信号的每一帧的过零率都是独立计算的,将各帧的过零率与语音信号的总体平均过零率的 1.5 倍进行比较,计算出语音信号的高过零率比率(High Zero-Crossing Rate Ratio,HZCRR):

$$\mathrm{HZCRR} = \frac{1}{2N} \sum_{n=0}^{N-1} [\mathrm{sgn}(\mathrm{ZCR}(n) - 1.5\,\mathrm{avgZCR}) + 1]. \tag{2}$$

式中,N 表示帧的总数量,n 表示帧索引,$\mathrm{ZCR}(n)$ 表示第 n 帧的过零率,avgZCR 为平均过零率. 经过分析可知,说话信号具有较高的高过零率比率,而音乐信号的却较低.

1.2 短时能量(*Short Time Energy*)的改进

短时能量是区分清音和浊音的重要特征参数,其定义为:

$$\mathrm{STE} = \sum_{m=1}^{N-1} x^2(m). \tag{3}$$

式中,$x(m)$ 为信号数据. 实验结果表明浊音的短时能量明显高于清音的短时能量. 可以通过设置一个短时能量门限值来区分浊音和清音. 在信噪比较高的情况下,短时能量还可以作为区分有声和无声的依据[8]. 类似于过零率,将语音信号的总体平均短时能量的一半与各帧的短时能量进行比较,得到改进之后的特征值,称之为低短时能量比率(Low Short-Time Energy Ratio,LSTER),计算如下:

$$\mathrm{LSTER} = \frac{1}{2N} \sum_{n=1}^{N-1} [\mathrm{sgn}(0.5\,\mathrm{avgSTE} - \mathrm{STE}(n)) + 1]. \tag{4}$$

式中,N 表示帧的总数量,n 表示帧索引,avgSTE 表示短时能量的平均值,$\mathrm{STE}(n)$ 为第 n 帧的短时能量.

1.3 频谱流量(*Spectrum Flux*)分析

频谱流量是一个语音序列内所有相邻帧间频谱变化的均值,它反映信号能量谱的变化快慢[9],定义为:

$$\mathrm{SF} = \frac{1}{(N-1)(K-1)} \sum_{n=1}^{N-1} \sum_{k=1}^{K-1} [\log(A(n,k) + \delta) - \log(A(n-1,k) + \delta)]^2. \tag{5}$$

式中,δ 是为了防止计算溢出而设定的一个很小的数值,$A(n,k)$ 为第 n 帧信号的离散傅里叶变换,其计算表达式为:

$$A(n,k) = \left| \sum_{m=-\infty}^{\infty} x(m)\, w(nL - m)\, \mathrm{e}^{\mathrm{j}\frac{2\pi}{L}km} \right| \tag{6}$$

式中,$x(m)$ 为输入的语音信号,$w(m)$ 为窗函数,L 为窗长,k 为离散傅里叶变换的点数.

1.4 线性频谱对(*Linear Spectral Pairs*)距离测量

El-Maleh 等人[10] 的研究结果表明,LSP 对不同的语音类型有较高的区分度. 假设 LSP 特征向量符合

高斯分布,即可用下式表示其概率分布函数($P_{\mathrm{LSP}}(\xi)$):

$$P_{\mathrm{LSP}}(\xi) = \frac{1}{2\pi\sqrt{|\hat{C}_{\mathrm{LSP}}|}}\exp\left[-\frac{\hat{C}_{\mathrm{LSP}}^{-1}}{2}(\xi-\hat{u}_{\mathrm{LSP}})^{\mathrm{T}}(\xi-\hat{u}_{\mathrm{LSP}})\right]. \tag{7}$$

式中,$\hat{C}_{\mathrm{LSP}}$ 是估计的 LSP 协方差矩阵,$\hat{u}_{\mathrm{LSP}}$ 是估计的均值向量,两个语音序列之间的 LSP 距离可以被定义为:

$$D = \int_{\xi}[P_{\mathrm{LSP}}(\xi) - P_{\mathrm{SP}}(\xi)]\ln\frac{P_{\mathrm{LSP}}(\xi)}{P_{\mathrm{SP}}(\xi)}\mathrm{d}\xi. \tag{8}$$

2 语音特征值分析

实验随机选取一些单声道的说话音、音乐和环境音信号,采样率为 44 100 Hz,分别计算出其改进前后的特征值,再进行比较,观察区分度. 由于语音信号有频繁的清音和浊音交替,而音乐信号一般没有这种特性,所以语音信号的过零率一般要比音乐信号的过零率高[11]. 过零率实验测量结果如图 1 所示,可以看出,说话信号基本都位于音乐信号之上,表明说话信号的过零率明显超过音乐信号,但二者具有一定的交叉性,其区分效果不是很明显.

图 2 所示为改进之后的高过零率比率的曲线图. 可以看出,说话信号的高过零率比率要明显高于音乐信号,并且具有较少的信号交叉,因此比基本的过零率有更高的区分度.

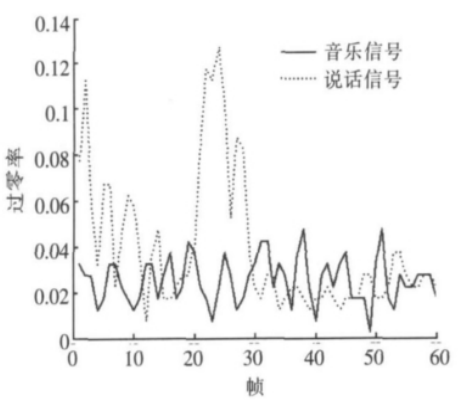

图 1 说话音和音乐的过零率分布

Fig.1 ZCR of speech and music

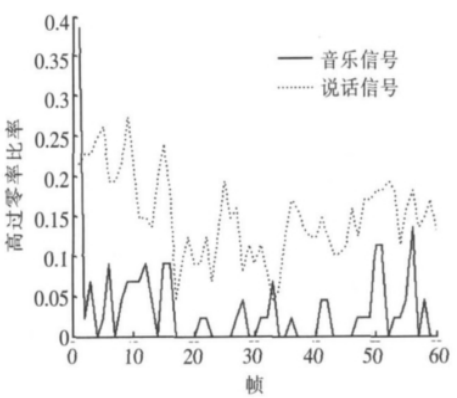

图 2 说话音和音乐的高过零率比率分布

Fig.2 HZCRR of speech and music

由于说话信号中经常会出现停顿,因此在语音信号中会出现更多的静音帧,这种特性使得说话信号的短时能量往往要大于音乐信号的短时能量,说话信号的短时能量的波动性也要强于音乐信号[12]. 图 3 给出了说话信号和音乐信号的短时能量分布,可以看出说话信号的跳变比较大,出现明显的静音帧,而音乐信号的短时能量分布较为平缓,但两者的交叉混叠较多,区分度较差.

通过低短时能量比率的计算得到的分布如图 4 所示. 可以看出,说话信号的低短时能量比率几乎都在音乐信号之上,存在很少的交叉,体现了低短时能量比率具有更大的区分度,能很好地解决静音帧的干扰问题.

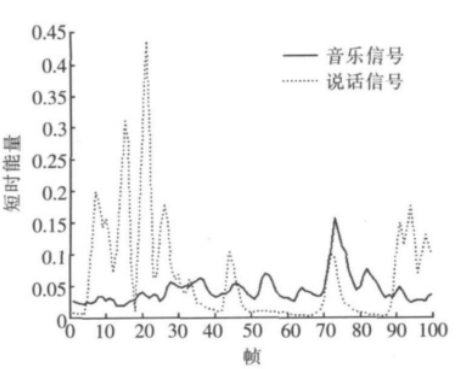

图 3 说话音和音乐的短时能量分布

Fig.3 STE of speech and music

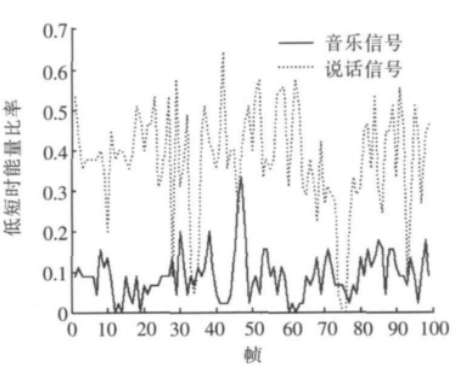

图 4 说话音与音乐的低短时能量比率

Fig.4 LSTER of speech and music

频谱流量通过对相邻帧之间的能量谱的比较来衡量信号频谱的变化程度. 由于说话信号中带有浊音与清音的交替,所以说话信号的频谱流量往往要大于音乐信号,环境音的频谱流量也往往大于音乐. 从图 5 可知,说话信号每一帧的频谱流量变化较大,反映出说话过程中的抑扬顿挫; 音乐信号由于声音的平缓连贯性,频谱流量相对较平; 环境噪声信号的变化比说话小,因此频谱流量有一定的变化,介于说话信号和音乐信号之间.

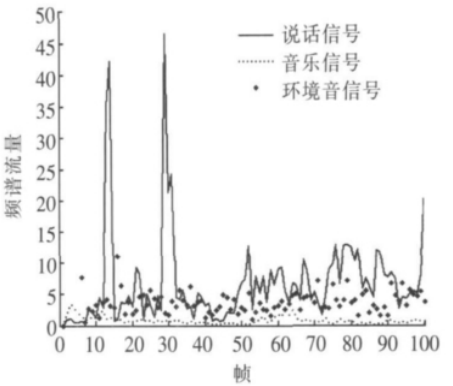

图 5　说话音、音乐、环境音的频谱流量

Fig.5　SF of speech, music and environment sound

3　语音分割技术

利用特征值对语音信号进行分割,需分两步实现: 第一步将音频信号分为说话音与非说话音,第二步将非说话音的语音信号进一步分为音乐、环境音等.

3.1　预分类技术

基于高过零率比率、低短时能量比率和频谱流 3 个特性,利用 KNN 分类器进行预分类,将音频信号分为说话音与非说话音两部分. 基本的流程如图 6 所示,此过程计算量少、速度快,可实现较精确的划分. 但是这些特征只是提高了过零率、短时能量和频谱的区分度,如果说话音中加入噪声,则说话音的这些特征值就十分接近音乐,因此需进行再次处理.

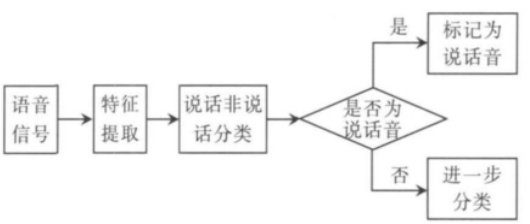

图 6　预分类过程

Fig.6　Flow chart of pre-classification

3.2　细化分类技术

在噪声环境下,用线性频谱对来区分说话音与非说话音具有较强的鲁棒性,因此可以利用线性频谱对将上一步的分割结果进行进一步的细分. 如图 7 所示,设定阈值 1 和阈值 2 来进一步区分说话音和非说话音,通过 LSP 的协方差矩阵与编码库内的数值进行比较,若两者之间的距离小于一个阈值,则该语音段为说话音,否则为非说话音.

一般情况下,阈值 1 大于阈值 2. 假设说话序列与编码库的 LSP 距离满足高斯分布 $N(\mu_p,\sigma_p)$,一般不会大于 $\mu_p+3\sigma_p$,其中 μ_p 和 σ_p 分别为数学期望和标准方差. 如果某个 LSP 距离大于该值即可被认定为非说话信号,因此阈值 1 被定义为 $\mu_p+3\sigma_p$. 同样,假设非说话信号与编码库的 LSP 距离满足高斯分布 $N(\mu_p,\sigma_p)$,阈值 2 可以被定义为 $\mu_p-3\sigma_p$. 因此可用阈值 1 和 2 来区分说话音和非说话音.

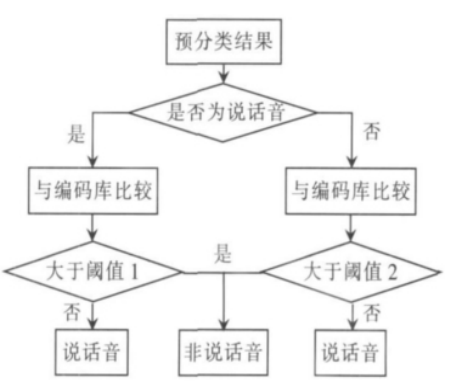

图 7　细化分类过程

Fig.7　Flow chart of refining classification

3.3　静音、音乐、环境音划分技术

由于静音片段的短时能量和过零率很低,故可通过对其过零率和短时能量的比较来找出静音片段. 如果平均过零率和短时能量小于某个阈值,则将其划为静音,否则划为非静音. 在多数情况下,环境音的频谱流量比音乐的大,可以此作为划分音乐和环境音的依据. 如果 SF 大于某个阈值,则该语音片段被划分为环境音,否则即为音乐.

4　实验结果分析

从电影、演讲、广播节目中截取了一些语音片段,然后对其进行切割与合并处理,其中说话音、音乐等各种语音类型均为 50 个. 语音预分类结果如表 1 所示.

表 1　语音预分类结果

Table 1　Result of pre-classification

语音类型	总数	分类结果		
		说话音	音乐	环境音
说话音	50	44	4	2
音乐	50	3	43	4
环境音	50	8	11	31

从表 1 的语音预分类结果可以看出,对于说话音和非说话音分类具有较高的准确性,但对于环境音的划分准确度还有待提高. 通过细化分类及音乐和环境音的划分,最终的语音分割结果如表 2 所示,环境音划分的准确度得到了提升.

为比较语音分割的效果,表 3 给出了使用基本的语音特征的语音分割结果. 通过实验结果的对比可见,使用改进特征值的语音分割具有比较满意的分割结果,其中说话音的准确率为 92%,音乐的准确率为 88%,环境音的准确率为 80%. 同时利用改进特征值的分割准确率比利用传统基本特征值的分割准确率也有所提高.

表 2 语音分割最终结果

Table 2 Final results of audio segmentation

语音类型	总数	分割结果		
		说话音	音乐	环境音
说话音	50	46	2	2
音乐	50	3	44	3
环境音	50	5	5	40

表 3 基于简单特征值的语音分割结果

Table 3 Results of eigenvalue based audio segmentation

语音类型	总数	分割结果		
		说话音	音乐	环境音
说话音	50	44	4	2
音乐	50	4	41	5
环境音	50	9	4	37

5 结论

本文通过对语音信号每一帧的基本特征值与整个语音序列的平均基本特征值进行比较,得到改进的特征值. 利用改进的特征值进行预分类和细化分类,得到语音分类和分割结果. 实验结果表明,改进特征值对语音分割的准确率有一定的提高. 但本研究的分割方法具有较强的噪声敏感性,同时对于 Hip-Hop 类型音乐的分割准确度也有待提高.

[参考文献] (References)

[1] 李恒峰,李国辉. 基于内容的音频检索与分类 [J]. 计算机工程与应用, 2000, 36(7): 54-56.
Li Hengfeng, Li Guohui. Content-based audio retrieval and classification [J]. Computer Engieering and Applications, 2000, 36(7): 54-56. (in Chinese)

[2] 朱爱红,李连. 基于内容音频检索综述 [J]. 微机发展, 2003, 13(12): 58-61.
Zhu Aihong, Li Lian. The summarization of content-based audio retrieval [J]. Microcomputer Development, 2003, 13(12): 58-61. (in Chinese)

[3] 张燕,唐振民. 基于 MFCC 和 HMM 的音乐分类方法研究 [J]. 南京师范大学学报: 工程技术版, 2008, 8(4): 112-114.
Zhang Yan, Tang Zhenmin. Research of music classification based on MFCC feature and HMM model [J]. Journal of Nanjing Normal University: Engineering and Technology Edition, 2008, 8(4): 112-114. (in Chinese)

[4] 张永皋,马青玉,孙青. 基于 MFCC 和 CHMM 技术的语音情感分析及其在教育中的应用研究 [J]. 南京师范大学学报: 工程技术版, 2009, 9(2): 89-92.
Zhang Yonggao, Ma Qingyu, Sun Qing. Investigation on speech emotion analyses and its application in education based on MFCC and CHMM techniques [J]. Journal of Nanjing Normal University: Engineering and Technology Edition, 2009, 9(2): 89-92. (in Chinese)

[5] Foote J. An overview of audio information retrieval [J]. Multimedia Systems, 1999, 7(1): 47-59.

[6] Saunders J. Real-time discrimination of broadcast speech/music [C] // Proc ICASSP96. Washington DC: IEEE Computer Society, 1996(2): 993-996.

[7] Scheirer E, Slaney M. Construction and evaluation of a robust multifeature music/speech discriminator [C] // Proc ICASSP97. Washington DC: IEEE Computer Society, 1997(2): 1-4.

[8] Zhang Y B, Zhou J. Audio segmentation based on multi-scale audio classification [J]. Multimedia Systems, 2004(4): 349-352.

[9] Lu L, Zhang H J, Jiang H. Content analysis for audio classification and segmentation [J]. IEEE Trans Speech Audio Process, 2002, 10(7): 504-516.

[10] Campbell J P, Jr. Speaker recognition: a tutorial [J]. Proceedings of the IEEE, 1997, 85(9): 1 437-1 462.

[11] Lu L, Jiang H, Zhang H J. A robust audio classification and segmentation method [C] // Proc 9th ACM Int Conf Multimedia. New York: ACM, 2001: 203-211.

[12] El-Maleh K, Klein M, Petrucci G, et al. Speech/music discrimination for multimedia application [C] // Proc ICASSP00. Istanbul: IEEE Press, 2000: 2 445-2 448.

[责任编辑: 严海琳]