ISSN 1000-0054 清华大学学报（自然科学版）2013 年 第 53 卷 第 6 期 31/40
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2013, Vol. 53, No. 6 878-882

快速准确的自动音乐/语音分段方法

万玉龙， 周若华， 颜永红

（中国科学院 声学研究所，语言声学与内容理解重点实验室，北京 100190）

摘 要：自动音乐/语音分段是语音识别技术的重要部分。该文采用回声器时频分析计算平均能量谱及定长片段的优化短时低能量比，用 Bayes 分类器判定类型，并根据内容连续性对分段结果修正；最后采用振幅包络匹配滤波器求所有起始点，对分段结果进一步优化。实验基于多语种电视电台录音和国内电话录音数据展开，结果显示该方法的性能 FMeasure 可达 0.987，较已有分类系统性能有大幅提升，同时处理速度也有大幅度改进。

关键词：音乐/语音分段；回声器时频分析(RTFI)；优化短时低能量比；起始点侦测

中图分类号：TP 391.4；TP 391.3 **文献标志码**：A

文章编号：1000-0054(2013)06-0878-05

Fast and precise automatic music/speech segmentation

WAN Yulong, ZHOU Ruohua, YAN Yonghong

(Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: This article describes a fast and robust method for automatic music/speech classification and segmentation. A resonator time-frequency image (RTFI) is used to represent the average energy spectrum of the input data with the modified short-time low energy ratios then extracted for each constant length segment. Then, the system uses the Bayesian maximum-a-posteriori (MAP) classifier to decide the audio class of each segment and refine the classification results based on the fact that the audio types are continuous over a short time. An onset detection method is then used to rectify the beginnings and ends of each segment. The system is evaluated using recordings from multi-language radio and television shows and Chinese telephone calls. Tests show that the system outperforms the state-of-art methods with an FMeasure of up to 0.987 and much faster processing speed.

Key words: music/speech segmentation; resonator time-frequency image (RTFI); modified short-time low energy ratio; onset detection

音乐/语音的自动切分技术对于音频处理研究很重要。例如，在自动语音识别技术中，对于输入的音频数据，如果将非语音自动切分屏蔽，可以大大提高识别的准确率和稳定性；在哼唱检索及其他音乐信息检索应用中，将输入数据中非音乐部分自动切分屏蔽，同样可以有效地增强检索性能。在基于内容的说话人识别及多国语种识别等领域[1]，自动音乐/语音分类及切分技术也是非常关键的前端处理手段之一[2]。

目前国内外已有不少相关研究。Saunders[3]以 2.4 s 为固定窗长，提取 4 维基于过零率的统计特征和 1 维能量相关特征，利用多 Gauss 分类器对电台的音频进行实时语音/音乐分类，准确率可达 98%。Scheirer 和 Slaney[4]用 13 维特征来描述语音和音乐的分布特性，同时用 3 种分类器进行分类，准确率也高于 90%。Carey 等[5]对比评估了 4 种音频特征的分类性能，分别是幅度、倒谱、基频以及过零率，发现倒谱及其差分项具有最佳区分性。El-Maleh 等[6]融合了每帧的线性频谱和过零率特征，采用 Gauss 分类器和 K 近邻分类器，并采用了特定的决策规则，性能同文[3]相当。Karneback[7]利用 20 个频带的低频调制振幅及其标准差有效地区分了音乐和语音，该特征对通道质量的依赖性较低，维数上大大低于 Mel 频率倒谱系数（Mel frequency cepstral coefficient，

收稿日期：2013-04-27

基金项目：国家自然科学基金项目（10925419，90920302，61072124，11074275，11161140319，91120001，61271426）；
中国科学院战略性先导科技专项（XDA06030100，XDA06030500）；
国家“八六三”高技术项目（2012AA012503）；
中科院重点部署项目（KGZD-EW-103-2）

作者简介：万玉龙(1988—)，男(汉)，河南，博士研究生。

通信作者：周若华，研究员，E-mail：zhouruohua@hccl.ioa.ac.cn

MFCC)。Pinquier 等[8]提出一种模型匹配方法，鉴别语音的准确率为 99.5%，鉴别音乐的准确率为 93%。

同上述方法相比，本文提出的音乐/语音分段算法计算复杂度低且有效，不需要大量数据及多维特征用于模型训练，也不需要复杂的分类器融合及得分判定过程，仅利用 1 维基于回声器时频分析(resonator time-frequency image，RTFI)[9]的能量特征，通过 Bayes 判别进行分类，同时采用了基于内容连续性的后决策法及时域振幅包络匹配的起始点侦测法，进行切分性能优化。

1 时频分析

图 1 为本文方法的系统流程图。原始数据通过时频分析、计算优化低能量比率以及 Bayes 判定得到初始片段分布，同时对原始数据进行预处理、振幅包络匹配滤波及峰值提取与修正得到所有起始点，最终融合结果得到优化后的分段结果。

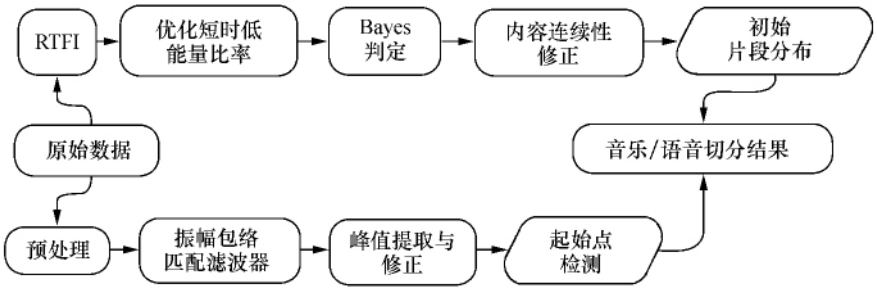

图 1 音乐/语音分类切分系统流程

1.1 回声器时频分析

最常用的音频时频分析方法为短时 Fourier 变换(short-time Fourier transform，STFT)法，但它只能提供均匀的时频分析，即频率分辨率在频率范围内保持一致。然而，对于许多音频分析任务，在不同的频率范围所需要的频率分辨率是不同的，低频段需要更好的频率分辨率，高频段需要更好的时间分辨率，这也符合人耳听觉模型的感知规律[9]。由此，本文定义了频变的时频分析方法(frequency-dependent time-frequency analysis，FDTF)，

$$\mathrm{FDTF}(t,\omega)=\int_{-\infty}^{+\infty}s(\tau)w(\tau-t,\omega)\mathrm{e}^{-\mathrm{j}\omega(\tau-t)}\,\mathrm{d}\tau. \tag{1}$$

其中：$s(\tau)$为采样点值；与 STFT 不同，FDTF 的窗函数 w 与分析频率 ω 相关，窗长(时频分辨率)随 ω 变化而变化。式(1)还可表示为基于带通滤波器组的实现方法，

$$\mathrm{FDTF}(t,\omega)=s(t)*I(t,\omega). \tag{2}$$

其中：$I(t,\omega)=w(-t,\omega)\mathrm{e}^{\mathrm{j}\omega t}$ 为带通滤波器组的脉冲响应函数，ω 为中心频率大小。带通滤波器的带宽决定了时频分辨率。

本文采用的 RTFI 是基于滤波器组的 FDTF 方法，其特点是采用了最简单的一级数字复数回声滤波器，通过设定参数实现不同的时频分辨率，例如均匀分析和常数 Q 分析(constant-Q analysis)等。同时，由于采用了最低阶的滤波器组，计算复杂度很小，例如当设定为常数 Q 分析时，比传统的快速常数 Q 分析的速度更快[9]。有关 RTFI 更详细的内容，可以查阅文[10]。

1.2 参数设置

系统输入为单声道、采样率 8 kHz、采样点 16 b 的音频数据。滤波器组包括 89 个滤波器，各滤波器的中心频率按对数标度设置，相邻滤波器的中心频率差 1 个半音，分析频率范围为 25.96 Hz～4.43 kHz。输入信号通过 RTFI 分析后，得到能量谱，然后按照式(3)求平均能量谱。图 2 所示分别为音乐/语音的平均能量谱。

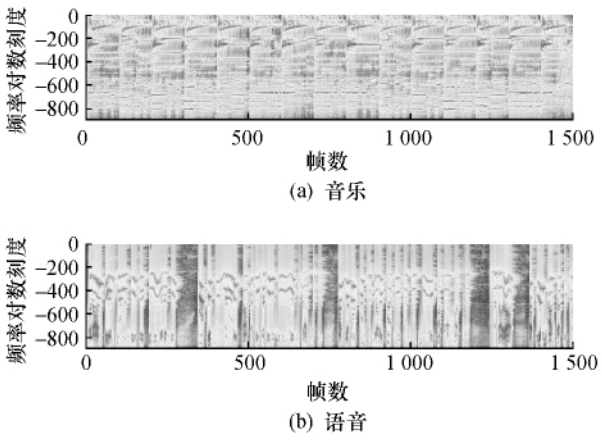

图 2 音乐/语音的平均能量谱对比

$$A(l,\omega_m)=$$

$$10\lg\left(\frac{1}{M}\sum_{n=(l-1)M+1}^{lM}|\mathrm{RTFI}(n,\omega_m)|^2\right). \quad (3)$$

其中：M 为每帧的采样点数，帧长设为 10 ms 时，M 为 80；$\mathrm{RTFI}(n,\omega_m)$ 为离散 RTFI 在采样点 n 和频率 ω_m 处的值；l 为帧索引。

由图 2 可以发现：语音由于清音、浊音及静音帧交替出现，短时能量不稳定，而音乐信号则相对较平稳，帧间能量变化不大，与语音相比，音乐的短时能量变化方差不大，低能量帧的占比较小。因此，可利用这一特点来区分音乐与语音。

2 音乐/语音分类

2.1 优化低能量比率

系统对短时低能量比率(modified short-time low energy ratio, MSTLER)[11]进行优化，

$$\mathrm{MSTLER}=$$

$$\frac{1}{2N}\sum_{n=1}^{N}[\mathrm{sign}(\mathrm{lowthres}-E(n))+1]. \quad (4)$$

$$\mathrm{lowthres}=\delta\left(\sum_{n=1}^{N}E(n)\right)/N. \quad (5)$$

其中：N 表示在一个窗长内的帧数，当设定窗长为 1 s，帧长为 10 ms 时，N 值为 100；$E(n)$ 表示第 n 帧的短时能量；lowthres 表示被认定为“低能量”的最大值；δ 为平均能量阈值系数，可以有效控制音乐/语音的分类性能。在本文系统中，δ 值设定为 0.1。从图 3 中可以看到在 $\delta=0.1$ 时音乐/语音的 MSTLER 统计结果对比。可见，音乐和语音的 MSTLER 在分布概率上存在明显的差别。MSTLER 具有很强的区分性，通过阈值设定即可直接判断片段类型。

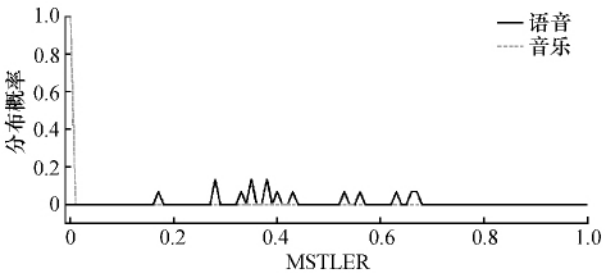

图 3 音乐/语音的 MSTLER 在 $\delta=0.1$ 的区分性能

2.2 片段分类和基于内容连续的修正

对音频片段计算 MSTLER 后，按照如下 Bayes 决策规则判定当前片段的音频类型：

$$\text{片段类型}=\begin{cases}\text{音乐}, & \text{如果 MSTLER}<\lambda;\\ \text{语音}, & \text{如果 MSTLER}\geqslant\lambda.\end{cases} \quad (6)$$

本文中 λ 设定为 0.06。同时，由于音频内容类型本身具有连续性，如果某一片段的持续时间小于阈值 T(这里 T 设定为 3 s)，且前后片段的类型一致，则将其类型修正为同前后片段类型一致。

3 片段起始点侦测及切分优化

3.1 基于包络匹配滤波器的起始点侦测

音频片段起始点的准确侦测对于音乐/语音分段也有非常关键的影响。本文采用了基于振幅包络匹配滤波器的起始点侦测算法(envelope matched filter onset detection, EMFOD)[12]对输入音频作以下处理：

1) 以 10 ms 为一帧，求每帧样点中振幅绝对值的最大值，同时为了避免振幅较小的帧在之后的处理中被当作噪声忽略，对每帧得到的振幅最大绝对值进行指数变换。即

$$B_k=A_k^{\lambda},\ 0<\lambda<1.$$

其中：B_k 为变换后的值，A_k 为每帧最初的振幅绝对值最大值，λ 在本文中设定为 0.7。

2) 按式(7)对 B_k 求卷积：

$$C_k=\sum_{\tau=0}^{11}B_{k-\tau}\cdot\mathrm{filter}[\tau]. \quad (7)$$

其中，filter 为选定的匹配滤波器，根据音频振幅在起始点的分布情况，设定为 filter[12]=[3,3,4,4,−1,−1,−2,−2,−2,−2,−2,−2]。

3) 峰值提取。当 C_{k+3} 大于某阈值且为峰值时，则将第 k 帧认定为起始点。这里阈值设定为 3.0。

4) 起始点修正。当两个起始点距离小于 1 s，即 100 帧时，忽略后者。最终，得到所有起始点位置。图 4 是该起始点侦测方法的整个过程。

3.2 音乐/语音分段起始点修正

对片段的起止点按照如下方法进行修订：1) 对于每个片段的起点，将其之前最近的起始点设定为片段起点；2) 对于每个片段的终点，将其之后最近的起始点设定为该片段终点。

4 实验结果及分析

本实验所用数据一部分来自多国语言的电视电台录音，另一部分来自国内的电话录音，包括彩铃、音乐及人声。测试数据共 50 个音频(40 个电视电台录音+10 个电话录音)，总长约 130 min，共 220 个片段。其中：语音片段 133 个，音乐片段 87 个。评测指标选定为：精确度(Precision)，召回率(Recall)，FMeasure

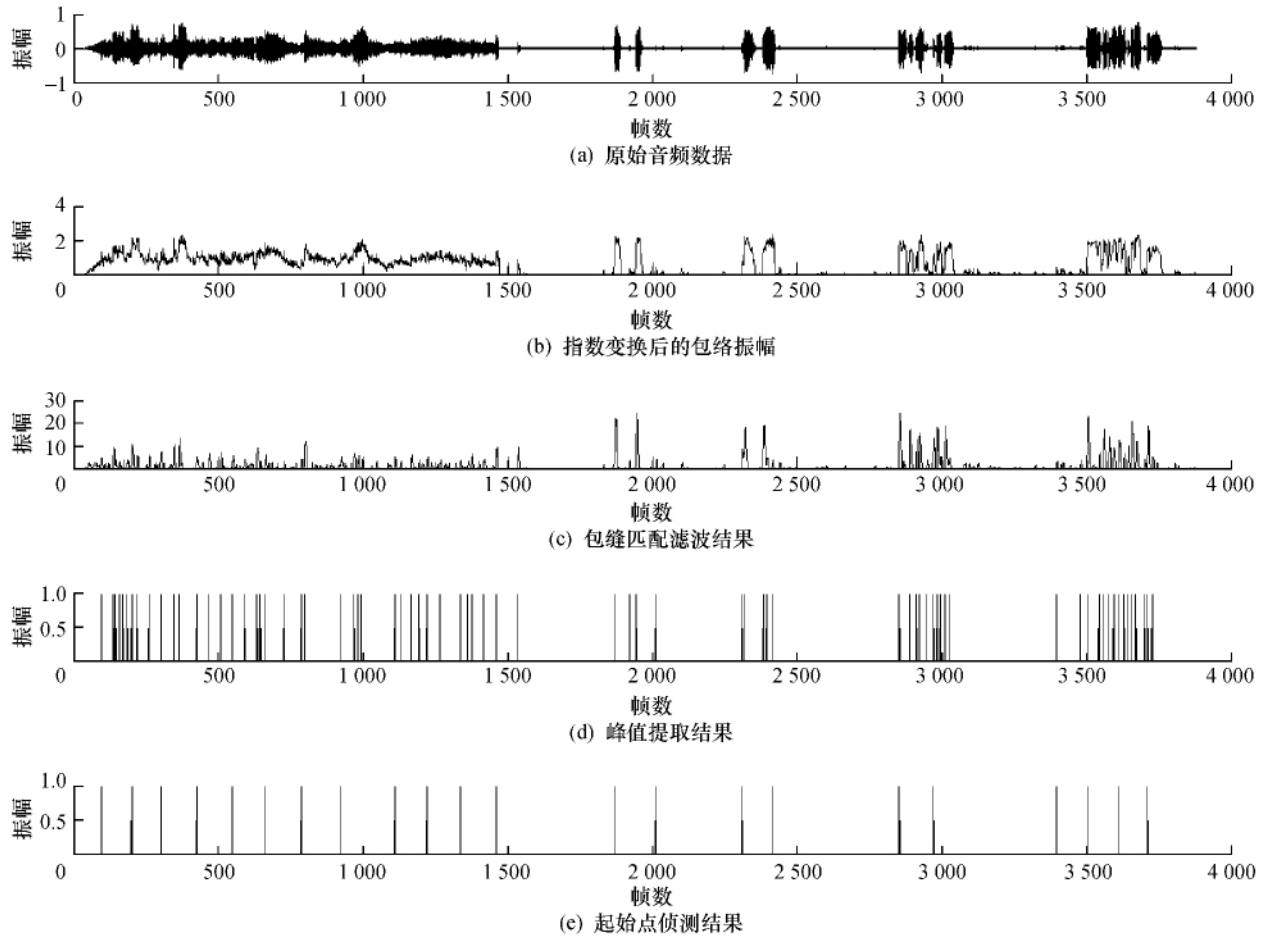

图 4 音频片段起始点侦测过程

及运行时间(running time，RT，每处理 20 s 数据的时间)。$\mathrm{FMeasure}=\frac{2\times\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}$。

在对起始点侦测算法性能进行评估时，当侦测到的起始点与正确起始点相差 60 ms 以内时，视为正确，否则为错误；将结果同短时能量法(short-time energy，STE)和帧间差分法(difference of magnitude，DM)的结果进行了对比，对比结果如表 1 所示。可见，本文的起始点侦测无论在性能和效率上均比前两者要好很多。

表 1 起始点侦测结果对比

分段方法	精确度 %	召回率 %	FMeasure	运行时间 s
STE	78	91	0.84	0.65
DM	80	94	0.86	3.43
本文	98	97	0.97	0.21

在性能评估中，当片段类型判定正确且起始点正确时认定为正确分段。将分段结果分别与基于 STFT 的两种方法进行了对比：1) 使用未优化的短时低能量比率(STLER)作为区分特征，2) 使用 MSTLER。对比结果如表 2 所示。可以看到，本文提出的音乐/语音分段方法在无论是精确度上还是召回率上都有明显提升。

表 2 片段切分结果评估

分段方法	精确度 %	召回率 %	FMeasure
STFT+STLER	81.5	91	0.860
STFT+MSTLER	91.3	94	0.926
本文	98.5	99	0.987

5 总 结

本文提出了一种快速准确的音乐/语音分段方法，该方法采用了遵循人耳感知规律的回声器时频分析方法、基于振幅包络匹配滤波器的起始点侦测法以及基于内容延续性的修正方法。实验结果表明，该方法有效地提升了音频分段的准确性和效率，可以满足很多音频处理应用的需求。

参考文献 (References)

[1] LEI Yun, Hansen J H L. Dialect classification via text-independent training and testing for Arabic, Spanish, and Chinese [J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2011, **19**(1): 85-96.

[2] Kim H G, Jang G J, Park J S, et al. Speech segregation based on pitch track correction and music-speech classification [J]. *Adv Electr Comput Eng*, 2012, **12**: 15-20.

[3] Saunders J. Real-time discrimination of broadcast speech/music [C]// IEEE International Conference on Acoustics, Speech, and Signal Processing. Atlanta, GA, USA: IEEE, 1996: 993-996.

[4] Scheirer E, Slaney M. Construction and evaluation of a robust multifeature speech/music discriminator [C]// IEEE International Conference on Acoustics, Speech, and Signal Processing. Munich, Germany: IEEE, 1997: 1331-1334.

[5] Carey M J, Parris E S, Lloyd-Thomas H. A comparison of features for speech, music discrimination [C]// IEEE International Conference on Acoustics, Speech, and Signal Processing. Phoenix, AZ: IEEE, 1999: 149-152.

[6] El-Maleh K, Klein M, Petrucci G, et al. Speech/music discrimination for multimedia applications [C]// IEEE International Conference on Acoustics, Speech, and Signal Processing. Istanbul, Turkey: IEEE, 2000: 2445-2448.

[7] Karneback S. Discrimination between speech and music based on a low frequency modulation feature [C]// Proc Eurospeech. Aalborg, Denmark, 2001: 1891-1894.

[8] Pinquier J, Sénac C, André-Obrecht R. Speech and music classification in audio documents [C]// IEEE International Conference on Acoustics, Speech, and Signal Processing. Orlando, FL, USA: IEEE, 2002: 4164-4168.

[9] ZHOU Ruohua, Reiss J D, Mattavelli M, et al. A computationally efficient method for polyphonic pitch estimation [J]. *EURASIP Journal on Advances in Signal Processing*, 2009, **2009**: 1155-1165.

[10] ZHOU Ruohua. Feature Extraction of Musical Content for Automatic Music Transcription [D]. Lausanne, Switzerland: Swiss Federal Institute of Technology, 2006.

[11] Wang W Q, Gao W, Ying D W. A fast and robust speech/music discrimination approach [C]// International Conference on Information, Communications and Signal Processing. Hohhot, China: IEEE, 2003: 1325-1329.

[12] Ding J J, Tseng C J, Hu C M, et al. Improved onset detection algorithm based on fractional power envelope match filter [C]// 19th European Signal Processing Conference. Barcelona, Spain: EUSIPCO, 2011: 709-713.

(上接第 859 页)

[7] **陈顺强, 苏连科. 彝语口传文化数字化采集方法及其保护与传承研究: 以毕摩、苏尼、口弦、阿都高腔为例** [J]. **西南民族大学学报: 人文社会科学版**, 2012(11): 47-51.
CHEN Shunqiang, SU Lianke. Yi oral cultures digital acquisition method and its protection and heritage studies: A case study of Bimo, Suni, Jew's harps and Adu high-pitched tune [J]. *Journal of Southwest University for Nationalities: Humanities and Social Science*, 2012(11): 47-51. (in Chinese)

[8] **鲍怀翘, 吕士楠. 蒙古语察哈尔话元音松紧的声学分析** [J]. **民族语文**, 1992(1): 61-68.
BAO Huaiqiao, LÜ Shinan. The acoustic analysis of vowel lax/tension characters in Mongolian Čaqar dialect [J]. *Minority Languages of China*, 1992(1): 61-68. (in Chinese)

[9] Childers D G, Hicks D M, Moore G P, et al. Electrogolottography and vocal fold physiology [J]. *Journal of Speech and Hearing Research*, 1990(33): 245-254.

[10] **孔江平. 论语言的发声** [M]. **北京: 中央民族大学出版社**, 2001.
KONG Jiangping. On Language Phonation [M]. Beijing: Minzu University of China Press, 2001. (in Chinese)

[11] **于善英. 不同歌唱类型歌手共振峰特征及音色形成的机理研究** [J]. **音乐研究**, 2010(3): 74-80.
YU Shanying. The research of different types of singing formant characteristics and the tone formation mechanism [J]. *Music Research*, 2010(3): 74-80. (in Chinese)

[12] Sundberg J. Articulatory interpretation of the singing formant [J]. *The Journal of the Acoustic Society of America*, 1974, **55**(4): 838-844.