ICICS-PCM 2003
15-18 December 2003
Singapore

A Fast and Robust Speech/Music Discrimination Approach

W.Q. Wang[1,2] W. Gao[1,2] D.W. Ying[1]

College of Information Sciences and Engineering, Graduate School of the Chinese Academy of Sciences [1]
Institute of Computing Technology, Chinese Academy of Sciences [2]

Abstract

This paper presents a simple and effective approach to discriminate speech and music. First, the proposed modified low energy ratio is extracted from each window-level segment as the only feature. Then the system applied the Bayes MAP classifier to decide the audio class of each segment. Last, based on the fact that the audio types of neighboring segments have very strong relevance, a novel context-based post-decision method is designed to refine the classification results. The proposed method is evaluated on about 5 hours of audio data, which involves clean and noisy speech from various speakers, as well as a wide range of musical content. The experimental results are promising, and a classification accuracy of more than 97% has been achieved despite the low computation complexity of the method.

1. Introduction

Automatic segmentation and classification of audio streams are very significant in many applications. For instance, it is desirable that the non-speech portion in audio streams are disabled to input into the speech recognizers in automatic speech recognition (ASR) systems [1]. This technique is also very useful for content-based indexing and retrieval of audiovisual data [2].

Music and speech are the two most important audio classes. Many research efforts have been reported on discriminating them. Saunders [3] proposed a real-time speech/music discriminator, which was used to automatically monitor the audio content of FM radio channels. In his system, four statistical features on the zero-crossing rate and one energy-related feature are extracted in a 2.4 second window, then a multivariate-Gaussian classifier is applied, and an accuracy of 98% is reported. Scheirer and Slaney [1] presented a complicated approach to the task. They exploited thirteen features to characterize distinct properties of speech and music signals, and examined three classification schemes, i.e. the multidimensional MAP Gaussian classifier, the GMM classifier, and the nearest-neighbor classifier. They reported an accuracy of over 90%. Carey et al [4] compared and evaluated the maximum discrimination power from four types of features, i.e. amplitudes, cepstra, pitch and zero-crossings. Their experimental results showed cepstra and delta cepstra bring the best performance on the problem. Khaled El-Malech et al [5] combined the line spectral frequencies and zero-crossing-based features for frame-level speech/music discrimination. The Gaussian classifier and the KNN classifier were evaluated in their work. After adapting their decision to the segment-level, the performance can be compared favorably with that of [1]. Stefan[6] found the low frequency modulation amplitudes over 20 critical bands and their standard deviations can form a good discriminator for the task, and the features were less sensitive to channel quality and model size than MFCC. Pinquier [7] presents an original modeling approach, called the differentiated modelling approach, to discriminate speech/music, which characterizes each class with their own feature spaces and statistical models. According to their report, their system could identify speech with an accuracy of 99.5% and music with 93%.

Compared with the previous works, this paper presents a very low computation complexity but effective approach. The approach exploits only one simple new feature, called modified low energy ratio. A novel context-based post-decision method is proposed to improve the performance. Its simplicity and robustness make its application scope very wide, especially for the applications where the low system cost is strongly demanded.

The rest of the paper is organized as follows. Section 2 presents the details of the proposed method, especially the proposed feature, and the proposed context-based post-decision method. In the Section 3, the approach is carefully evaluated on a large and unbiased data set, and the experimental results are reported. At last, related conclusions are given in section 4.

2. Speech/Music Discrimination

2.1 Modified Low Energy Ratio

In early study [3], Saunders pointed out that the energy contour of a waveform is capable to discriminate speech and music. For music, the contour tends to show little change over a period of several seconds, but the alternation between voicing and frication in speech can produce a remarkable change. Scheirer and Slaney [1] exploit the percentage of "low–energy" frames to characterize such distinction. They define the feature as the proportion of frames with RMS power less than 50% of the mean RMS power within a one-second window. The same feature is used by Lu et al [8] for speech-music discrimination, but called as Low Short-Time Energy Ratio (LSTER). They constructed the class-conditional probability density

function curves for speech and music, and derived the Bayes error rate is 8.27%.

The research experience in [9] suggests that the distinction in energy contour between speech and music is not best characterized by the feature definition from [1][8]. We present a modified version of the feature, called as *Modified Low Energy Ratio* (MLER),

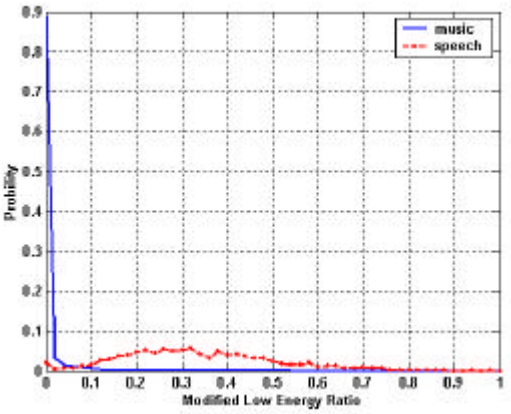
$$MLER = \frac{1}{2N}\sum_{n=1}^{N}[\mathrm{sgn}(lowthres - E(n)) + 1] \quad (1)$$

$$lowthres = d \cdot \frac{\sum_{n=1}^{N} E(n)}{N} \quad (2)$$

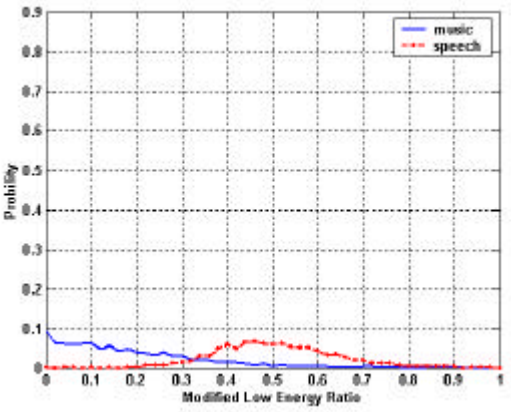
where N is the total number of frames in a window, $E(n)$ is the short time energy of the nth frame, d is a control coefficient, which decides how low the $E(n)$ needs to be so that the frame is considered as "low energy". Compared with the definition in [1][8], we introduce a coefficient d instead of 0.5. In our perspective, d is a very significant parameter, which greatly affects the feature's discrimination power for speech and music, and 0.5 is too large. Our experience suggests it is desirable d is chosen from the interval [0.05, 0.12]. In the Fig.1, an intuitive comparison is given on the discrimination power under the different values. The probability distribution function (pdf) curves in Fig.1 are calculated with 20ms frames and 1-second window, from a little less than 70 minutes' speech and music, and the data of each class exceed 34 minutes. To make the curves more typical, speech is chosen from different languages, including English, Mandarin and French, and music involves a wide range of instrumental music and songs. From the Fig. 1, we see that the pdf of speech primarily keeps the same shape when d is changed from 0.5 to 0.1, and a reasonable translation is observed. At the same time, the pdf of music changes drastically, and the probability for MLER=0 is 0.8905, which makes its curve very steep. If the Bayes' maximum a posteriori (MAP) classification approach is applied to the data, the best classification results are summarized as the Tab.1, with the assumption that the class probabilities for speech and music are equal. The average misclassification ratio under d =0.1 is 4.08% while the value under d =0.5 is 11.75%. Therefore an appropriate d value is very significant in characterizing the distinction of speech and music signals. Intuitively, d =0.1 is a good selection, which makes music signals more separable from speech signals in the MLER feature space, and d =0.5 seems a bad one.

Tab.1 The classification results using Bayes' MAP classification approach under different d

Class	Classification results (d =0.1)		Classification results (d =0.5)	
	Music	Speech	Music	Speech
Music	95.04%	4.96%	83.27%	16.73%
Speech	3.20%	96.8%	6.77%	93.23%



(a) d =0.1



(b) d =0.5

Fig. 1 The comparison of the discrimination power of MLER under d =0.1 and d =0.5

2.2 Context-based Classification and Segmentation

The framework of our system is outlined in the Fig.2. First, an audio stream is divided into a sequence of segments with window length. For each segment, its MLER feature is extracted. Then Bayes MAP decision rule is applied to decide the type of the segment, i.e., music or speech. Since the feature space is one-dimension, the process can be simplified and modeled as,

$$AudioClass = \begin{cases} Music & if\ p[i] < l \\ Speech & if\ p[i] \geq l \end{cases} \quad i = 1,2,3... \quad (3)$$

where $p[i]$ denotes the modified low energy ratio of the ith segment, and l is a threshold which corresponds to the MLER value of the cross point of the two pdf curves in Fig.1 (a). Apparently the Bayes classifier does not use the fact that the audio types of neighboring clips have very strong relevance. Therefore a context-based post-decision module is designed to further process the output sequence of the Bayes classifier to improve the accuracy.

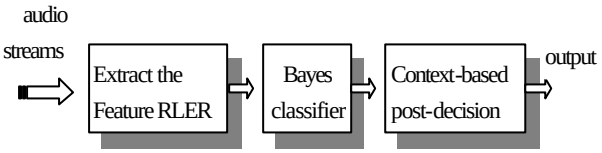


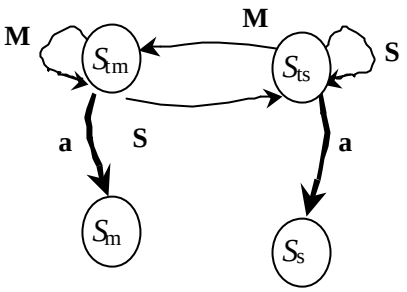
Fig.2 The system framework

The research works in [5][10] have noticed that the likelihood of class transfer is very low in a continuous audio stream, i.e. the audio clip of the same class usually lasts for quite a few seconds and it is almost impossible that the audio types switch frequently in a few seconds. The knowledge has been exploited to improve the performance. They exploit the strategy that the single speech clip between two music clips is corrected as music and vice versa in [5][10]. We call the method as Median Filter (MF) approach. Compared with the method, we propose a novel context-based post-decision approach, which can better characterize the persistence in continuous audio streams and make the system more robust.

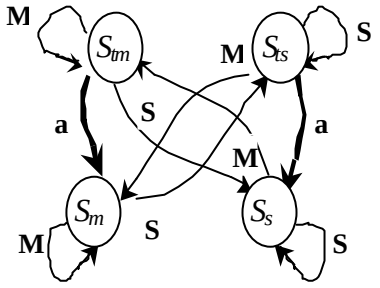
In the approach, four states are defined to help the system decide the final audio type of the current clip. The states are respectively denoted as S_m, S_s, S_{tm} and S_{ts}. The module operates like an automata. It ceaselessly receives two kinds of symbols from the output sequence of the Bayes classifier, i.e. **M** (music) and **S** (speech). When a new symbol enters, the module will keep or update its current state, as shown in Fig.3. The system experiences two stages, initialization stage and working stage, for a continuous audio stream. The initial state is S_{tm} or S_{ts}, depending on the audio type of the first segment. If it is music, the initial state will be S_{tm}, else it is S_{ts}.

In the initialization stage, Fig.2 (a) gives a complete description of state transitions and the output of audio class of the corresponding segment. For S_{tm} and S_{ts}, each of them has a maximum stay time T and a stack, denoted by L_{tm} and L_{ts} respectively. If the system stays in the S_{tm} for more than T, an automatic state transition to S_m occurs and the initialization stage ends. At the same time, all the segments in the stack L_{tm} and L_{ts} are labeled as music, then two stacks are emptied. The similar process occurs for the state S_{ts}. Let t denote the time when the system stays in the current state. When $t < T$, if the symbol **S** is met in the state S_{ts}, the current state will not change; if the symbol is **M**, the current state becomes S_{tm}. Analogously, the transition function in the S_{tm} can be derived. While the system still stays in the S_{tm} or S_{ts}, the audio class of the current segment will be postponed to declare.

Once the system enters the working stage, the state transition function and the output of audio class are defined as Fig.2 (b). Compared with the initialization stage, the behavior of S_m and S_s are defined. If the current state is S_m and the current symbol is **M**, the module will stay in the state S_m and declare the audio type of the current segment is music immediately. If the symbol **S** is met in the state S_m, the system will transit to the state S_{ts} and store the identifier of the current segment into the stack L_{ts}. In the scenaria, the segment's audio type cannot be declared immediately, since the system needs more information from the proceeding symbols. If the symbol **M** is met in the state S_{ts}, the system will transit to the state S_m, all the segments in the stack L_{ts} will be declared as music, and the stack L_{ts} will be emptied. If the symbol **S** is met in the state S_{ts}, the system stays in the state S_{ts}, the identifier of the current segment is pushed in the stack L_{ts}. When the system stays in the state S_{ts} for a long interval T, it automatically transits to the state S_s, all the segments in the stack L_{ts} are declared as speech, and the stack L_{ts} will be emptied. Due to the symmetry property, the similar definition for the state S_s and S_{tm} can be derived from that for S_m and S_{ts}.



(a) Initialization stage



(b) Working stage

Fig.3 The state transition graph of our context-based post-process approach

3. Experiment Evaluation

To evaluate the proposed algorithm, we carefully prepare the audio test database. The speech data come from news program of radio and TV stations, talks, as well as dialogs in movies, and the languages involve English, French and Chinese with different levels of noise, especially in news programs. The speakers involve male, female with different ages. The length of the whole speech data is 2 hours and 40 minutes. The music consists of songs and instrumental music. The songs covers as more styles as possible, such as rock, pop, folk, classic, and are sung by male and female in English and Chinese. The instrumental music chosen by us

covers different instruments and different styles. For example, various instruments include piano, violin, cello, clarino, piper, and electronic instruments, etc. Some music pieces in movies are also included, which are played by multiple different instruments. The length of the whole music data is totally 2 hours and 23 minutes.

On the whole test database, the discrimination power of MLER (d =0.1) and LER (d =0.5) are compared. Both features are calculated with 20ms frames and 1 second window. Based on the statistics on the 70 minutes' speech and music data in section 2.1, we choose 0.06 for l under d =0.1, and 0.32 under d =0.5. The experimental results are summarized in the Tab.3. The misclassification rate for MLER decreases by 13.7% for music, and 4.9% for speech. From the results, we can see the MLER has stronger discrimination power for music and speech. At the same time, the comparison result shows us another advantage of the MLER that the misclassification rate for each class is well balanced. The point is very significant, since it means the performance of the system is insensitive to the mixture ratio of two class data. Compared with MLER, besides the obvious increase of the misclassification rate, the LER seems to have a strong preference to speech. Such unbalance is very harmful for robustness. Therefore we say the MLER outperforms the LER.

Tab.2 The classification results using Bayes' MAP classification approach under different d

Class	Classification results (d =0.1)		Classification results (d =0.5)	
	Music	Speech	Music	Speech
Music	90.2%	9.8%	76.5%	23.5%
Speech	8.6%	91.4%	13.5%	86.5%

In the test experiments, we also pay attention to the adaptive capability of the MLER. For music, we found the MLER can achieve the classification accuracy of more than 95% for most styles. Some exceptions include the songs with very strong rhythm created by percussion instruments and the instrumental music played only by wind instrument, such as trumpet, piper. The average classification accuracy rate of them can only reach 76.2%. For speech, the MLER has very excellent performance for pure speech. Most classification errors come from noisy speech in news program, but they are common in the news programs in our test database. More than 60% of test speech is news program.

On the basis of the experiment aforementioned, we further integrate the proposed context-based post-decision technique into the system to see how much gain it brings. At the same time, the strategy in the [5][10] is also implemented and evaluated on the same input, i.e. the output of Bayes MAP classifier under d =0.1. The Tab.3 tabulates the corresponding experimental results of the two approaches. In our proposed approach, we choose 4 seconds for the maximum stay time T. According to the Tab.3, the proposed method makes the classification accuracy of the system increase by more than 6.5% while the MF method brings only an increase of less than 4.5%. The experimental results demonstrate the effectiveness of the context-based approach.

Tab.3 The comparison of classification results using different post-decision methods

Class	Our method		MF method	
	Music	Speech	Music	Speech
Music	97.0%	3.0%	94.0%	6.0%
Speech	1.6%	98.4%	4.3%	95.7%

4. Conclusions

This paper presents a very simple but robust approach to discriminate speech and music. The method exploits the modified version of low energy ratio, and a novel context-based post-procession technique. The careful experiment evaluation shows the modification of the audio feature LER can improve the discrimination power greatly and more effective. The original definition of the audio feature seems to hide the potential. The audio types of neighboring segments in an audio stream have very strong relevance. We propose a novel context-based post-decision method to characterize the persistence. The comparison experiment shows it has better performance than the counterpart in the previous work. The experimental results also demonstrate the robustness of the system. The classification accuracy achieves more than 97% on a 5 hours' audio data with a wide range of styles. At the same time, its simplicity brings obvious advantages in constructing low cost system.

References

[3] C.D. Smith and E.F. Jones, "Title of paper," Proceedings of the IEEE Conference on ..., pp. 320 - 324, City, State, Country, June 1985.

[1] E. Scheirer and M. Slaney, " Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator", *Proc. ICASSP'97*, Vol. II, pp. 1331-1334, Munich, Germany, April 1997.

[2] T. Zhang, C. -C. Jay Kuo, " Audio Content analysis for Online Audiovisual Data Segmentation and Classification", IEEE Transaction on Speech and Audio Processing, Vol.9, No. 4, May 2001.

[3] J. Saunders, " Real-Time Discrimination of Broadcast Speech/Music", *Proc ICASSP'96*, Vol.II, pp. 993-996, Atlanta, May 1996.

[4] Michael J. Carey, Eluned S. Parris and Harvey Lloyd-Thomas, "A Comparison of Features for Speech, Music Discrimination", ICASSP1999, pp1432-1435

[5] K.El-Maleh, M.Klein, G. Petrucci, and P. Kabal, " Speech/music discrimination for multimedia applications", *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing* (Istanbul), pp. 2445-2448, June 2000.

[6] *Stefan Karnebäck*," Discrimination between speech and music based on a low frequency modulation feature", European Conference on Speech Communication and Technology, September 3-7, 2001, Allborg, Denmark, pp.1891-1894,

[7] *Julien Pinquier, Christine Sénac and Régine André-Obrecht, "* Speech and Music Classification in Audio Documents", ICASSP 2002

[8] Lie Lu, Hao Jiang, and HongJiang Zhang, "*A robust audio classification and segmentation method*," ACM Multimedia 2001, Ottawa, Canada, September 30 - October 5, 2001

[9] W.Q. Wang, W. Gao, "Automatic Segmentation of News Items Based on Video and Audio Features", The Second IEEE Pacific-Rim Conference on Multimedia 2001,Oct,24-26,2001, Beijing, China.

[10] Jarina R, O'Connor N, Marlow S, Murphy N, "Rhythm Detection for Speech-Music Discrimination in MPEG Compressed Domain", *the 14th International Conference on Digital Signal Processing (DSP 2002)*, Santorini, Greece, 1-3 July 2002.