A BACKGROUND MUSIC DETECTION METHOD BASED ON ROBUST FEATURE EXTRACTION

Tomonori Izumitani, Ryo Mukai, and Kunio Kashino

NTT Communication Science Laboratories
3-1, Morinosato Wakamiya, Atsugi-shi, Kanagawa 243-0198, Japan
{izumi, ryo, kunio}@eye.brl.ntt.co.jp

ABSTRACT

We propose a music segment detection method for audio signals. Unlike many existing methods, ours specifically focuses on a background-music detection task, that is, detecting music used in *background* of main sounds. This task is important because music is almost always overlapped by speech or other environmental sounds in visual materials such as TV programs. Our method consists of feature extraction, dimension reduction, and statistical discrimination steps. For each step, we analyzed a set of methods to maximize the detection accuracy. With a simple post processing step, we achieved a framewise error rate as low as 8 % even when the mixed speech was louder than the target music by 10dB.

Index Terms— Background music detection, Gaussian mixture model, k-nearest neighbor method, feature selection

1. INTRODUCTION

Automatically detecting music parts from audio signals in TV or radio broadcasts is becoming a basic and important task to meet the increasing demands for multi-media indexing systems and music copyright management systems. In such audio signals, music is often overlapped by narration, conversation, or other environmental sounds. We call the task of detecting segments containing music on audio signals "background music detection" in this paper.

For a similar purpose, many speech/music discrimination methods have been proposed in terms of feature extraction and discrimination algorithms. They mainly discriminate pure speech or music from each other.

Saunders proposed a speech/music discrimination system based on the zero-crossing rate [1]. Then, Scheirer et al. proposed various features and developed a system based on a Gaussian mixture model (GMM) and the k-nearest neighbor method (k-NN). Some of the features, namely, the spectral centroid, the spectral roll-off, and the spectral flux, are related to the form of a sound spectrum and commonly used as basic features for other speech/music discrimination systems [2][3] or for other tasks, such as musical genre classification [4].

Some methods adopt features commonly used for audio processing, such as Mel-frequency cepstral coefficients (MFCCs) with discrimination algorithms based on a hidden Markov model (HMM) [5] or a support vector machine (SVM) [6].

Several multi-media indexing systems include a function for detecting music overlapped by other sounds. They are based on empirical features such as edge intensity on a spectrogram [7] or the level of harmonicity [8]. They assume musical sounds comprising harmonic components with a stable frequency and are therefore not always effective.

However, it is still unclear whether these features work efficiently for background music detection because most features are designed to discriminate speech from music.

In this paper, we investigate the properties of various features, including the empirical features and the MFCCs used in previous works, using statistical learning methods, GMM and k-NN, when music is overlapped by speech with various amplitude ratios. Then, we test two schemes for dimension reduction: feature selection by Fisher's criterion and a principal component analysis. These analysis show that the use of the spectral powers along the Mel-frequency scale, dimension reduction with PCA, and frame-by-frame discrimination based on GMM yields the best accuracy.

2. METHODS

2.1. Feature extraction

To investigate features suitable for background music detection, we consider four feature sets: (1) a set composed of 14 empirical features developed in previous works, (2) MFCC, (3) the spectral powers with the linear-scaled frequencies, and (4) the spectral powers with the Mel-frequencies.

Empirical features

Most speech/music discrimination systems adopt features that are empirically designed to emphasize differences between speech and music sounds. We call them "empirical features" in this paper.

Here, we focus on six features from the seven used by Scheirer et al. [9]: 4-Hz modulation energy, the percentage of low-energy frames, the spectral centroid, the spectral roll-off point, the spectral flux, and the zero-crossing rate. Some of these features also used in musical genre classification systems [4].

In addition, we adopt a feature proposed by Minami et al. that represents the intensity of the edge in the time direction of a sound spectrogram [7]. We call this feature the "spectral edge" in this study.

The first six features are based on the definition by Scheirer et al. [9] and Tzanetakis et al. [4], and the last one is based on that by Minami et al. [7], with minor modifications.

- 4-Hz modulation energy: 4-Hz periodicity of each frequency channel in the spectrum. First, for each frequency channel, a bandpass filter bank with center frequency of 4 Hz is applied. Next, powers of all channels are averaged through frames within a window around the focusing frame and the value is normalized using the average power of the spectrum within the window.
- Percentage of low energy frames: The proportion of frames where the square root of averaged signal power (RMS power) is less than 50 % of RMS power within a window around the frame of interest.

- Spectral centroid: The spectral centroid denotes the frequency index located on the center of gravity of the spectrum.
- Spectral roll-off point: The spectral roll-off point is a frequency index that denotes the 95% point of the magnitude of the spectrum.
- Spectral Flux: The spectral flux denotes the difference in the magnitude of the spectrum from the preceding frame.
- Zero-crossing rate: The zero-crossing rate denotes how often the original audio signal changes its sign within the frame.
- Spectral edge: The spectral edge at t-th frame E_t denotes the intensity of the edge on a spectrogram:

$$E_t = \sum_{i=1}^{N} \left| \sum_{j=t-w}^{t+w} (S_j[i-1] - 2S_j[i] + S_j[i+1]) \right|, \quad (1)$$

where $S_j[i]$ denotes the spectral power of the i-th frequency channel at the j-th time frame and w determines the window size.

Finally, we generate a 14-dimensional feature vector by adding a variance of each feature calculated from frames within a window around the frame of interest.

Mel-frequency cepstral coefficients

A MFCC is a feature related to the perceptual scale of pithes (Mel-scale). This feature is commonly used in speech recognition systems and also in some speech/music discrimination systems [5][6].

Typically, low-order coefficients are used for discrimination [4]. However, it is not known which coefficients are efficient for background music detection. In this step, we first use 80 coefficients calculated from 80 filter banks for the feature set. In the dimension reduction scheme in the next step, a small number of components efficient for the discrimination are extracted.

For all MFCCs, we calculate the variances using frames within a window around the focusing frame and add them to the MFCC feature set. For the MFCC feature set, a 160-dimensional vector is generated.

Linear frequency spectral powers

The last two feature sets are based on spectral powers. The first set is calculated using frequency bands equally spaced in the frequency direction. To obtain this set, a power spectrum is calculated by short-time Fourier transform (STFT). Then, values within each frequency band are averaged and a 80-dimensional feature vector is again generated.

Mel-frequency spectral powers

In the second set, another 80 frequency bands are located in the Mel-frequency scale. It is intermediately produced in the MFCC calculation. The frequency bands are closely allocated in the low-frequency region and sparsely allocated in the high-frequency region.

Then, we generate a 160-dimensional vector by adding the variance of each frequency band power as well as the MFCC feature set.

2.2. Dimension reduction by Fisher's criterion and PCA

Each frame of an audio signal is represented by a high-dimensional vector, especially when MFCC or spectral powers are used. Using all elements for statistical learning, the classification accuracy may be degraded due to so called "curse of dimensionality", such as over-training for complicated models. To avoid this problem, a dimension reduction process is executed before applying statistical learning methods.

We implemented the process using two approaches. The first is the feature selection by which significant features for the classifications are selected. For this purpose, a method based on Fisher's criterion [10] or one that finds effective features by repeating classification trials by adding features recursively [11] have been proposed.

Here, we adopt a feature selection method based on Fisher's criterion for low computational cost. Fisher's criterion represents the degree of separation of two classes by using a single feature and is defined as

$$F = \frac{(\bar{x}_1 - \bar{x}_0)^2}{\sigma_1^2 + \sigma_0^2}, \quad (2)$$

where $\bar{x}_c$ and σ_c are the mean value and variance for class c of the focusing feature.

The second approach is the principal component analysis (PCA), which is commonly used for dimension reduction. The PCA maps high-dimensional vectors onto a low-dimensional subspace according to the distribution.

2.3. Discrimination by statistical learning methods

We tested the GMM and k-NN to discriminate frames containing music from those not containing music. They classify the t-th frame, using feature vector $\mathbf{x}_t$, into two classes, ω_1 and ω_0, which represent whether a frame contains musics or not, respectively.

The GMM represents the probability density of each class as a weighting addition of multiple Gaussian distributions. The GMM probability density for a class ω_c is defined as

$$p(\mathbf{x}|\omega_c) = \sum_{l=1}^{K} \pi_l \mathcal{N}(\mathbf{x}|\mu_l^c, \mathbf{\Sigma}_l^c), \quad (3)$$

where K is the number of Gaussian distributions composing the mixture. $\mathcal{N}(\mathbf{x}|\mu_l^c, \mathbf{\Sigma}_l^c)$ denotes a Gaussian distribution with a mean vector μ_l^c and a covariance matrix $\mathbf{\Sigma}_l^c$. The value of π_l is the mixing coefficient of l-th Gaussian distribution and π_l satisfies $0 \leq \pi_l \leq 1$ and $\sum_{l=1}^{K} \pi_l = 1$.

When the number of mixtures K is given as a hyper parameter, μ_l^c and Σ_l^c can be estimated by the expectation maximization (EM) algorithm. In the classification stage, the system discriminate the music frame according to $p(\mathbf{x}|\omega_c)$.

The k-NN is a method that calculates the distances between the test sample and every training sample and determines the class by majority vote using the classes of the nearest k samples. We simply use the Euclidian distance as the distance measure in this study.

3. EXPERIMENTS

3.1. Preparing musical audio overlapped by speech

To carry out the examination systematically, we prepared musical audio signal overlapped by speech signal with various amplitudes and also prepared pure speech signal. The system discriminates music-containing parts from pure speech parts.

For music data, we used 30 musical pieces in the RWC Music Database: Music Genre (RWC-MDB-G-2001) [12]. The pieces were selected from various genres, such as popular, rock, dance music, jazz, classical music, and world music. We extracted a 30-second-long audio signal for each of the 30 pieces.

For speech data, we used ten conversation sessions from the Corpus of Spontaneous Japanese (CSJ) [13]. It includes the voices of seven females and five males. Five conversation sessions were used

for the pure speech part. We extracted 900-second-long speech audio signals from these sessions. The remaining five conversation sessions were used to overlap the music parts.

Ten data variations were prepared according to the ratio of the amplitude of music signals to the speech, namely, –30, –20, –10, –3, 0, +3, +10, +20, +30, and $+\infty$ dB. We call this ratio the "music/speech ratio" in this study. Then, "$+\infty$ dB" means that the music parts consist of pure music and do not include speech. This is the same condition as in previous works for speech/music discrimination.

All audio signals were resampled at 12 kHz and normalized using mean and variance for every musical piece or conversation session. Each conversation session was divided into 30-sec-long segments. The divided conversation sessions and the 30-sec-long musical pieces were arranged alternately.

Then, we generated the four feature sets determined in the previous section: the spectral powers from the linear-scaled frequency (SPEC), the spectral powers with the Mel-frequencies (SPMF), MFCCs, and the empirical features (EMPR). All features were extracted at every 50 msec-long frame. A 2.0-sec-long window around each frame was used for variance calculation.

To obtain SPEC, we used 1,024-sample-long window for the calculation of FFT and the spectral power in dB was averaged every six frequency bands. The first 80 elements were used for SPEC.

For the calculations of the 4-Hz modulation energy, the percentage of low-energy frames, and the spectral edge in the empirical features, a 1.0-sec-long window is applied.

Using Fisher's criterion or PCA, the dimension was reduced to 3, 5, 10, and 20.

3.2. Discriminating music frames

We evaluated the music detection accuracy for all combinations of the four feature sets and discrimination methods. We examined k = 1, 3, 5, and 7 for the k-NN and the number of Gaussian mixtures K = 1, 3, 5, and 7 for he GMM. For one Gaussian model (K = 1), we used mean vectors and covariance matrices directly calculated from vectors in training data.

When applying the k-NN, every feature was normalized using mean and variance in the time direction to adjust the scale of features.

The evaluation was performed using a five-fold cross validation. In the data, the test set does not include frames from the same musical pieces or conversation sessions as the training set. Additionally, we randomly selected a music/speech ratio at each frame in the music part of training data, considering a practical use.

We used all frames in the training set for GMM learning, and used 4,000 frames randomly selected from the training set for k-NN to reduce computational cost.

Figure 1 shows the error rates against the music/speech ratio using (A) GMM and (B) k-NN. For each feature set, the best combination of parameters, namely, the number of reduced dimension, feature reduction method, and the number of Gaussian mixture K or nearest neighbor k, are investigated.

Using GMM, every feature set, especially the EMPR, achieves a very low error rate when the music/speech ratio is higher than 0 dB. However, when the music/speech ratio falls under 0 dB, the EMPR gets degraded rapidly. This indicates that a lot of characteristics based on the difference between music and speech audio signals disappears when speech and music are mixed.

On the contrary, simple spectral powers based on the Mel-frequency scale give very low error rates through whole music/speech ratio. The MFCC did not improve the accuracy.

When k-NN is applied, the SPMF yielded the lowest error rate with GMM as well. And the SPEC and the EMPR did not work well.

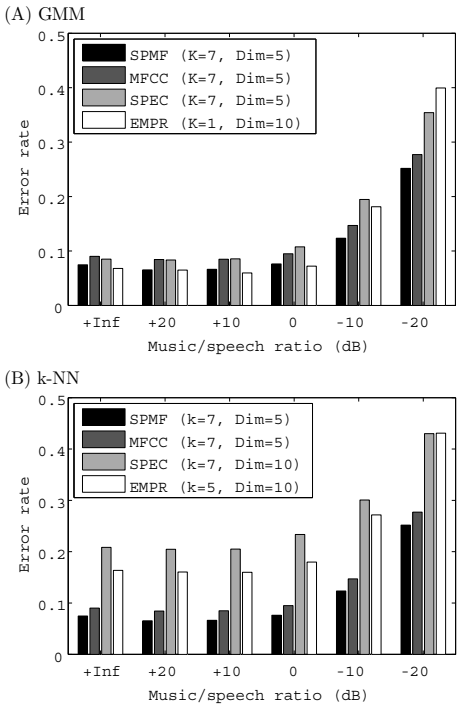

Fig. 1. Error rate of frame classification using (A) GMM and (B) k-NN. "Dim" denotes the number of dimensions extracted. For each feature set, the best parameter combination through the music/speech ratio is used. Fisher's criterion is used for SPEC and EMPR with k-NN and the PCA is used for all remaining sets.

In the experiment, the accuracy of k-NN was worse than GMM. It may be improved by using more training frames or refining the scaling.

On the whole, a complicated learning model (e.g. K=7 of the GMM) with fewer dimension achieves good classifiers.

3.3. Effect of dimension reduction

Table 1 shows effects of the dimension reduction process with GMM (K = 3) using a data set of a –10 dB music/speech ratio. For every feature set, the accuracy is improved by reducing the dimension using both Fisher's criterion and PCA.

When feature selections are performed using Fisher's criterion, three dimensions yield the best accuracy for all data sets. This indicates that we can construct a discrimination system using only a few informative features.

Also, when other K values are used, similar tendencies as those shown in Table 1 were found. On the whole, the model becomes complicated, with larger K, the effect of the dimension reduction increases.

Table 2 shows the top seven features that have high Fisher's criterion values from EMPR. The values in the table are averaged Fisher's criterion values calculated from frames used as training data of five-fold cross-validation.

In the case of pure speech/music discrimination ($+\infty$ dB), variance of spectral flux, percentage of low-frequency energy, and 4-Hz modulation energy yield a high Fisher's criterion. This is consistent with the results shown by Scheier et al. [9] However, when speech overlaps the music with large amplitude, some features, such as per-

Table 1. Error rates for the number of dimensions reduced by Fisher's criterion and PCA, using the –10 dB test set. This table shows the effect of dimension reduction for each feature set. The best number of dimensions for each feature set is indicated by bold font.

Feature set	Reduction method	# of dimension				
		3	5	10	20	All
SPMF	Fisher	**0.15**	0.16	0.18	0.23	0.27
	PCA	0.14	**0.13**	0.14	0.17	
MFCC	Fisher	**0.16**	0.16	0.18	0.23	0.27
	PCA	0.16	0.15	**0.15**	0.17	
SPEC	Fisher	**0.19**	0.23	0.29	0.28	0.50
	PCA	0.22	**0.22**	0.25	0.30	
EMPR	Fisher	**0.19**	0.24	0.24	–	0.34
	PCA	0.24	0.24	**0.23**	–	

Table 2. Fisher's criterion of the empirical features [F in eq. (2)]. Top seven features from 14 empirical features of training data are shown. "Mixed" indicates training data that includes every music/speech ratio and "Var." means variance.

music/music ratio	$+\infty$ dB	–20 dB	Mixed
Var. spec. flux	**1.40**	**0.28**	**1.08**
Spectral edge	**2.90**	**0.16**	**1.07**
Low-energy frame	**1.57**	0.00	**0.72**
4-Hz modul. energy	0.11	**0.21**	0.40
Var. spec. roll-off	0.52	0.14	0.37
Var. spec. cent.	0.48	0.00	0.25
Var. low-energy frame	0.15	0.02	0.17

centage of low-frequency energy and variance of spectral centroid, are significantly degraded. This explains why empirical features fails under the low music/speech ratio conditions.

For other feature sets, when Fisher's criterion is used, variances of spectral powers within lower frequency channels (SPMF and SPEC) and variances of low-order coefficients (MFCC) are extracted as top features.

3.4. Effect of post-processing

The accuracy can be improved by applying a post-processing to the frame-by-frame discrimination results as shown in previous works for speech/music discrimination [9].

We simply employed a smoothing process as a post-processing. A 4-sec-long window was applied to each frame. Then, the existence of music was determined by the proportion of frames within the window that were classified into music in the previous discrimination step. A threshold $\Theta = 0.4$ was used.

Table 3 shows error rates with or without the post-processing for each feature set. It indicates that the accuracy of background music detection can be improved by about four to six percent, for every feature set, with the post-processing based on a simple smoothing scheme.

4. CONCLUSION

We have proposed a music detection method that is effective even when loud interfering sounds simultaneously exist. The robustness was obtained by the use of the spectral powers with Mel-frequency scale with a GMM-based discriminator. We showed that the accuracy of the proposed method is better than the case where the empirically designed features are used. We also found that the PCA-based dimension reduction effectively works and the number of dimensions can be reduced down to five. With a simple post processing, we achieved a framewise error rate as low as 8 % even when the music/speech ratio is –10 dB, which is the case where the interfering speech is significantly louder than the target music. Future work will include refining the learning and post-processing methods to further improve the accuracy.

Table 3. Error rates with and without the post-processing. Test data sets with –10 dB music/speech ratio and the same parameters as Fig. 1 are used.

	SPMF	MFCC	SPEC	EMPR
With post-processing	0.08	0.09	0.18	0.12
W/o post-processing	0.12	0.15	0.22	0.18

5. ACKNOWLEDGEMENT

We would like to thank Drs. Shoji Makino and Junji Yamato for their continuous encouragement.

6. REFERENCES

[1] J. Saunders, "Real-time discrimination of broadcast speech/music," *Proc. ICASSP*, vol. 2, pp. 993–996, 1996.

[2] M. J. Carey, E. S. Parris, and H. Lloyd-Thomas, "A comparison of features for speech, music discrimination," *Proc. ICASSP*, vol. 1, pp. 149–152, 1999.

[3] T. Giannakopoulos, A. Pikrakis, and S. Theodoridis, "A speech/music discriminator for radio recordings using bayesian network," *Porc. ICASSP*, pp. 809–812, 2006.

[4] G. Tzanetakis, "Musical genre classification of audio signals," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[5] J. Ajmera, I. McCowan, and H. Bourlanrd, "Speech/music segmentation using entropy and dynamism features in a HMM classification framework," *Speech Communication*, vol. 40, pp. 351–363, 2002.

[6] L. Lu, S. Z. Li, and H.-J. Zhang, "Content-based audio segmentation using support vector machines," *Proc. ICME*, vol. 40, pp. 749–752, 2001.

[7] K. Minami, A. Akutsu, and H. Hamada, "A sound-based approach to video indexing and its application," *Trans. IEICE D-II (in Japanese)*, vol. J81-D-II, no. 3, pp. 529–537, 1998.

[8] T. Zhang and C. C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 4, pp. 441–457, 2001.

[9] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," *Proc. ICASSP*, vol. 2, pp. 1331–1334, 1997.

[10] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[11] M. Liu and C. Wan, "Feature selection for automatic classification of musical instrument sounds," *Proc. ICDL*, pp. 247–248, 2001.

[12] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," *Proc. of ISMIR*, pp. 229–230, 2003.

[13] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of japanese," *Proc. of LREC*, pp. 947–952, 2000.