A FAST SEARCH ALGORITHM FOR BACKGROUND MUSIC SIGNALS BASED ON THE SEARCH FOR NUMEROUS SMALL SIGNAL COMPONENTS

Hidehisa Nagano, Kunio Kashino and Hiroshi Murase

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

ABSTRACT

This paper proposes a method for detecting and locating a known music signal in a long audio stream. Unlike existing methods, ours assumes that the music is used as background music (BGM) and overlapped by another sound such as speech and that the interfering sound is typically louder than the target music. The proposed method is based on Time-series Active Search, which is a quick signal search method reported earlier. To realize the BGM search, however, a novel extension is introduced. That is, the music signal is firstly decomposed into a number of small time-frequency regions, and the search is carried out for each of those components. The results of the search are then integrated based on a voting scheme to find the target music locations. Experiments show that accurate search is possible when SNR is -5 dB and that the search completes in about 8 s for a 30-m stored signal.

1. INTRODUCTION

We have been tackling retrieval tasks for segments similar to a given specific audio/video signal (reference signal) from the long audio/video signal (stored signal). In this paper, we propose a fast search method for retrieving music signals used as background music (BGM) in a long audio signal (stored signal), where music is overlapped by another signal such as speech. In this retrieval, a music signal is assumed to be given as a reference signal and every segment including that music signal as BGM must be detected from the stored signal. Hereafter, we call this retrieval BGM retrieval. One major application of BGM retrieval would be in checking the use of music in radio and TV. In broadcast programs of radio and TV, music is often used as BGM and overlapped by speech. Another is video retrieval by query of the music used as BGM. In these cases, we use music CDs for the reference signal and retrieve the segments including the reference signal.

Music retrieval is an important application of information retrieval and several search methods for it have been proposed [1, 2, 3, 4, 5]. These methods can be divided into two groups according to their objectives. One aims at retrieving music "similar" to the query in some musical way, such as the same melody or a rearrangement of the music [1, 2, 3, 4]. The other group aims at retrieving segments almost the same as the query (i.e., reference signal) by signal level comparison. We call the latter *exact retrieval*. Exact retrieval can be accomplished by an exhaustive search carried out by calculating the similarities between the reference signal and the segment in the search window sliding on the stored signal. However, this method requires a huge amount of search time. To overcome this problem, we proposed a fast method for exact retrieval, which is called *Time-series Active Search (TAS)* [5]. TAS has been applied for music information retrieval systems and commercial monitoring systems on TV and radio. In addition, a feature fluctuation absorption technique for TAS has also been proposed [6, 7]. However, these methods are for exact retrieval, where it is assumed that the segments to be detected are almost the same as the reference signal. They cannot be applied to BGM retrieval. In BGM retrieval, even the segments in the stored signal, including the reference signal, considerably differ from the reference signal due to the overlap of other audio signals such as voices. In addition, the voices are often louder than the BGM.

Recently, the self-optimized spectral correlation method (SSC method) has been proposed for the classification of BGM [8]. This method can be applied to BGM retrieval by combining it with the above exhaustive search. In the SSC method, the reference signal is first decomposed into a number of small time-frequency components, and the local similarities of each decomposed component to the corresponding components of the stored signal with the same size in the time-frequency domain are calculated. In this calculation, the local similarity is maximized by optimizing the scale parameter for each pair of a reference signal component and a stored signal component. After the calculation of the local similarities for all pairs, the local similarities are integrated by the votes of the similarities in the time-position scale-parameter domain. In the voting, the local similarities with the same time position and scale parameter are summed up as the total similarity at that time position and scale parameter. Finally, the time positions whose total similarities are large enough at a certain scale parameter are detected. Although this method can be applied to BGM

0-7803-7663-3/03/$17.00 ©2003 IEEE

This paper was originally published in the Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, & Signal Processing, April 6-10, 2003, Hong Kong (cancelled). Reprinted with permission.

Table 1. Classification of previous methods and ours.

	exact retrieval	BGM retrieval
slow	exhaustive search (correlation method/SSC method)	
fast	TAS	**proposed method**

retrieval, it is impractical in terms of the retrieval speed due to its computational complexity.

Here, we propose extending TAS to BGM retrieval. This extension also utilizes the signal decomposition technique. However, we search for only the small components of the stored signal that are similar to the components of the reference signal by TAS before integrating the local similarities. By utilizing TAS and searching for only the similar components, the entire search can be performed faster than exhaustive search. Table 1 shows the relation among the existing methods and the proposed method.

The rest of this paper is organized as follows. Section 2 describes the proposed method. Section 3 then shows some experimental results. Finally, Section 4 concludes this paper.

2. PROPOSED METHOD

The aim of our BGM retrieval is to detect every segment of the stored signal that includes the reference signal as BGM as shown in Figure 1. Each segment is identified by the point in time at which the segment begins.

2.1. Algorithm

Figure 2 overviews the proposed search method.

Step 1 As shown in Figure 2(a), the frequency-power spectra of the reference signal and the stored signal are extracted, and the spectrum of the reference signal is decomposed into a number of small time-frequency components of uniform size. Here, let F_{t_i,w_m} be the component of the reference signal obtained by the decomposition, where t_i is time when F_{t_i,w_m} starts and w_m is the frequency band of F_{t_i,w_m}. In the same way, let G_{t,w_m} be the component of the same size at time t and the frequency band w_m of the stored signal. Here, let $T_R = \{t_1, t_2, \dots\}$ be the entire set of t_i of decomposed components of the reference signal and $W = \{w_1, w_2, \dots\}$ be the entire set of frequency bands. The power values of spectra are normalized in each component[1].

Step 2 For each F_{t_i,w_m}, detect every component at w_m of the stored signal similar to F_{t_i,w_m}, as shown in Figure 2(b). This search is performed by TAS for each F_{t_i,w_m}. Here, for each F_{t_i,w_m}, every component with a similarity greater than the threshold value s^p_{th} is detected. We call the similarity between components the local similarity.

Step 3 The local similarities of the components of the stored signal detected by the above process are integrated and the total similarity for each segment of the stored signal is calculated. The total similarity of the segment at the time t, referred to as $S(t)$, is calculated as

$$S(t) = \frac{1}{|T_R|} \sum_{t_i \in T_R} \left(\max_{w_m \in W} (s^p(F_{t_i,w_m}, G_{t+t_i,w_m})) \right), \quad (1)$$

where $s^p(F_{t_i,w_m}, G_{t+t_i,w_m})$ is the local similarity between F_{t_i,w_m} and G_{t+t_i,w_m}. Note that if G_{t+t_i,w_m} is not detected as a component similar to F_{t_i,w_m}, i.e., if $s^p(F_{t_i,w_m}, G_{t+t_i,w_m}) \leq s^p_{th}$, then $s^p(F_{t_i,w_m}, G_{t+t_i,w_m})$ is set to 0 in the calculation of Eq. (1). That is, in the calculation of Eq. 1, $s^p(F_{t_i,w_m}, G_{t+t_i,w_m})$ is evaluated only if G_{t+t_i,w_m} is detected as a component similar to F_{t_i,w_m} in Step 2. Note too that in Eq. 1 the maximum local similarity is accumulated at each t_i. This is done to chose the frequency band in which the influence of the foreground sound is relatively small.

Step 4 The segments whose total similarities are larger than the search threshold S_{th} are determined to include the reference signal.

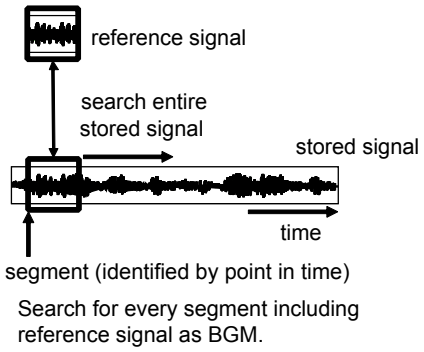


Fig. 1. Overview of BGM retrieval.

2.2. Characteristics of Proposed Method

The proposed method searches for only the components of the stored signal that are similar to the components of the reference signal quickly by using TAS. In BGM retrieval, the overlap of foreground sound corrupts the stored signal extremely. Almost all of the local similarities between the components of the reference signal and the components of the stored signal are small. Therefore, we search for only the pairs of components with large local similarities and integrate the large local similarities that influence the total similarities. Following this strategy, we can accelerate the retrieval. In the proposed method, the search for similar components of each F_{t_i,w_m} is performed faster by TAS than by exhaustive search. In TAS, when the search window is moved on the stored signal, the matchings of F_{t_i,w_m} to the components of the stored signal that are not similar to F_{t_i,w_m} are skipped, and every component similar to F_{t_i,w_m} is detected. On the other hand, in exhaustive search, the matchings are performed at each point in time while the search window is moved. The search result of TAS is the same as that of exhaustive search, but TAS is faster. Exper-

[1] The power values at the same time in a component are normalized by their average value.

imental results in Section 3 show the efficiency of this. For the details of TAS, see [5].

3. EXPERIMENTS

Here, we show two experimental results of BGM retrieval using the proposed method. One result concerns the search accuracy, and the other the search speed.

3.1. Experimental Conditions

For the experiments, we prepared two 30-m audio signals. One is an audio signal of pop and rock music, and the other is one of speech without any blank segments longer than 1 s. The speech is from a narration of a TV music program. The sampling rate of the audio signals is 44.1 kHz and the quantization level is 16 bits. For the stored signals, we mixed the audio signals of music and speech with the SNR of -5, -10 and -15 dB, where the SNR is defined as

$$10\log_{10}\left(\frac{\text{average power of music}}{\text{average power of speech}}\right) \quad \text{(dB)}. \qquad (2)$$

The negative SNR values mean that the speech is louder than the music. The above SNR values are realistic in TV or radio. For the reference signals, we randomly chose 15 segments of 15 s from that 30-m audio signal of music. Trial retrievals of these 15 reference signals from the stored signal in each case of the SNR were performed. For each reference signal, the segment in the stored signal to be detected is at the time when the reference signal was chosen. We call the segment to be detected *the target segment*, and we call a segment detected by the search procedure *a retrieved segment*. In the experiments, if there was a retrieved segment t and its total similarity was the greatest among the total similarities of the segments retrieved within 15 s before or after t, it was assumed that only t was detected in that time (Other retrieved segments were neglected). After that, a retrieved segment was considered correct if there was a target segment within 15 s before or after that retrieved segment. In the same way, a target segment was defined as retrieved if a retrieved segment existed within 15 s before or after the target segment.

For the extraction of the frequency-power spectra, we processed audio files through a band-pass-filter bank. The bank consists of 28 filters equally placed in the log-frequency axis from 525 to 2000 Hz. The analysis frame length used in these experiments was 1 ms with a shift of 0.5 ms.

The parameters of the decomposition of the spectra used in the experiments were as follows. The length of the component was 50 ms, and the 28 filters were divided into four bands equally in the order of their frequency. Namely, each component consisted of outputs of seven filters in 50 ms. We obtained such components from the reference signal at intervals of 0.6 s. From a 15-s reference signal, totally 100 components were obtained (four components in the frequency axis and 25 in the time axis).

The threshold value s^p_{th} for the local similarity was set to 0.6 throughout the experiments. The threshold value S_{th} for the total similarity was changed for each SNR, but it was fixed through the 15 trials for a single SNR.

Table 2. Search Accuracy.

SNR	accuracy
-5 dB	100%
-10 dB	93.3%
-15 dB	79.6%

3.2. Search Accuracy

The average precision rate and the average recall rate of 15 trials were adjusted to be the same by choosing the value of S_{th} for each SNR, and we used that rate to evaluate the search accuracy. Here, the precision rate and recall rate are defined as #(correct retrieved segments)/#(retrieved segments) and #(retrieved target segments)/#(target segments), respectively.

Table 2 shows the results for search accuracy. The accuracy was 100% when the SNR was -5 dB. This means that neither false retrievals nor retrieval misses occurred. When the SNR was -10 or -15 dB, the accuracy degraded. In such cases, i.e., when the BGM is very soft, finer decomposition of the reference signal will be needed.

3.3. Search Speed

The characteristic of the proposed search method is quickness, which is achieved by utilizing TAS for the search process for the decomposed components in Step 2 (Section 2.1). By utilizing TAS, the number of matchings of components in Step 2 is reduced, whereas in exhaustive search, every decomposed component of the reference signal is matched at every point in time in the stored signal.

Table 3 shows the average numbers of matchings of components performed in Step 2 for a retrieval of a reference signal. The matching number is invariant with the SNR when exhaustive search is used. For every SNR, the number of matchings in the proposed method was about 2% of that of exhaustive search. Table 4 shows the average search time for retrieval of a reference signal. According to the reduction of the number of matchings, the search times of the proposed method were reduced to about 4% of those of exhaustive search. Here, the search times were evaluated on a PC (CPU: Pentium 4 at 2 GHz with a 256 KB L2 cache, OS: Red Hat Linux 7.2). The time for the extraction of the spectra from the audio files is not included in the search time.

4. CONCLUSIONS

We have proposed a search method for BGM retrieval. The method is an extension of TAS and achieves a quick search in comparison with exhaustive search. Experiments have shown that the 15-s music signal can be detected in approximately 8 s from a 30-m audio signal overlapped by speech. We plan to further accelerate the search and improve the search accuracy by optimizing the parameters, such as the number and size of the decomposed components of the reference signal. We also plan to investigate a method of inte-

Table 3. Number of matchings.

SNR	exhaustive search	proposed method (rate)
-5 dB	3.3×10^8	6.5×10^6 (0.02)
-10 dB	-	6.5×10^6 (0.02)
-15 dB	-	6.5×10^6 (0.02)

Table 4. Search time (CPU time).

SNR	exhaustive search	proposed method (rate)
-5 dB	207.3 s	8.3 s (0.04)
-10 dB	218.9 s	8.5 s (0.04)
-15 dB	218.3 s	8.4 s (0.04)

grating local similarities that is more robust to the overlap of speech.

ACKNOWLEDGEMENTS

The authors thank Dr. Kenichiro Ishii and Dr. Noboru Sugamura for their help and encouragement.

5. REFERENCES

[1] A. Ghias, J. Logan, D. Chamberlin, and B. C. Smith, "Query By Humming: Musical Information Retrieval in An Audio Database," in *ACM Multimedia '95*, 1995.

[2] L. A. Smith, R. J. McNab, and I. H. Witten, "Sequence-Based Melodic Comparison: A Dynamic Programming Approach," in *Melodic Similarity : Concepts, Procedures, and Applications*. MIT Press, 1998.

[3] N. Kosugi, Y. Nishihara, S. Kon'ya, M. Yamamuro, and K. Kushima, "Music Retrieval by Humming: Using Similarity Retrieval over High Dimensional Feature Vector Space," in *Proc. of IEEE PACRIM'99*, 1999.

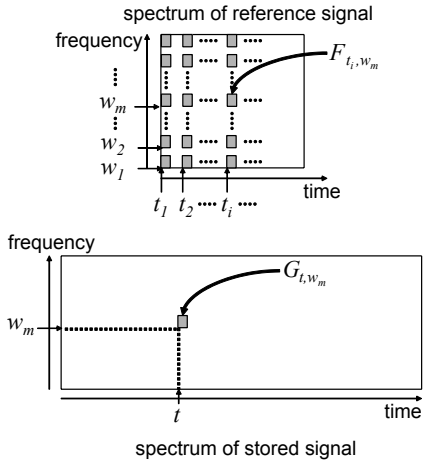
[4] H. Nagano, K. Kashino, and H. Murase, "Fast Music Retrieval Using Polyphonic Binary Feature Vectors," in *Proc. of ICME2002*, vol. I, 2002.

[5] K. Kashino, G. Smith, and H. Murase, "Time-series Active Search for Quick Retrieval of Audio and Video," in *Proc. of ICASSP-99*, vol. VI, 1999.

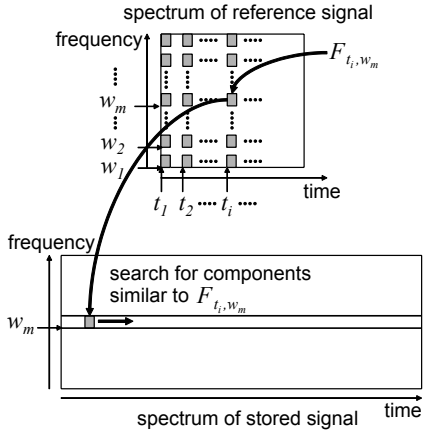
[6] K. Kashino, T. Kurozumi, and H. Murase, "Feature Fluctuation Absorption for a Quick Audio Retrieval from Long Recordings," in *Proc. of ICPR2000*, vol. 3, 2000.

[7] T. Kurozumi, K. Kashino, and H. Murase, "A Method for Robust and Quick Video Searching Using Probabilistic Dither-Voting," in *Proc. of ICIP-2001*, vol. 2, 2001.

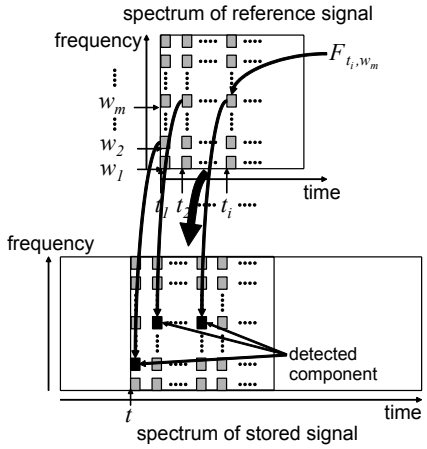
[8] M. Abe and M. Nishiguchi, "Self-optimized Spectral Correlation Method for Background Music Identification," in *Proc. of ICME2002*, vol. I, 2002.



(a) Extract spectra and decompose the spectrum of reference signal into small components.



(b) Detect every similar component for each F_{t_i,w_m}.



(c) Calculate total similarity at t from local similarities of detected components.

Fig. 2. Overview of proposed search method.