SELF-OPTIMIZED SPECTRAL CORRELATION METHOD FOR BACKGROUND MUSIC IDENTIFICATION

Mototsugu Abe and Masayuki Nishiguchi

S&S Architecture Center, Sony Corporation
6-7-35 Kita-Shinagawa, Shinagawa-ku, Tokyo 141-0001, JAPAN
{abe, nishi}@av.crl.sony.co.jp

ABSTRACT

This paper proposes a new method of detecting a known reference signal in an input signal highly corrupted by other sounds. One major application of the method is the identification of broadcast background music corrupted by speech. In this method, the reference signal is first decomposed into a number of small time-frequency components, and the maximum similarity between each component and the input is calculated. The similarities for all the components are then integrated by a voting method. Finally, the result is used to determine whether or not the reference exists in the input; and if it exists, to determine its position. Experiments on the identification of background music and the classification of similar TV commercials have shown that this method can identify 100% of target signals with an SNR of -10dB.

1. INTRODUCTION

The content-based classification, search and retrieval of audio[1, 2] are key issues in the handling of multimedia data on large-capacity storage media. Among the many approaches to accomplishing these tasks, some previous works address the issue of the exact matching of sounds, in which a known reference signal is searched for in a long, unknown, and slightly distorted input signal[3, 4]. The major applications of these methods are the identification of the title of broadcast music and the monitoring of commercials on FM radio and TV.

Although the search accuracy is almost 100% for a slightly distorted sound (SNR $>$ 30dB[3]), these methods cannot be used to identify background music with foreground speech, because the interfering speech is usually much louder than the target music (that is, the SNR is typically from -5 to -20dB).

This paper describes a new method of searching for a known reference signal in a mixture of sounds. Utilizing the time-varying nature of each individual sound, a very small target signal in a highly corrupted input can be detected. The reference signal is first decomposed into a number of small time-frequency components. For each component, the similarity value, which depends on the time shift between the component and the input and their amplitude scaling parameters, is calculated. Then, the time shift and scaling that yield the maximum similarity value are selected. A voting method is used to determine the maximum similarities of all the components along the time shift and scaling axes. Finally, the voting results are used to determine whether or not the target exists in the input; and if it exists, to determine its position. This paper also presents some experimental results on the identification of background music and the classification of similar TV commercials.

2. SELF-OPTIMIZED SPECTRAL CORRELATION METHOD

2.1. Characteristics of background music

Figure 1 shows spectrograms of (a) reference music and (b) music mixed with foreground speech that was used as input. The SNR[1] of the input was set to -10dB. Figure 2 shows the correlation between (a) and (b), in which the boxed region in (a) is used as the reference.

The correct peak (0.542 at 5.0s) obtained by conventional correlation matching is so small that it cannot be distinguished from the other peaks. This means that correlation matching cannot detect a target signal with such a low SNR. Although previous studies[3, 4] have improved the robustness for slight distortion, this method is still unsuitable when the SNR is very low. This is because the foreground speech is much stronger than the background music.

However, some spectral components of the music clearly appear in some areas, even though the other components are completely masked by the strong speech, because the amplitudes of the speech and music components vary in different ways over time.

[1] S is the music and N is the speech.

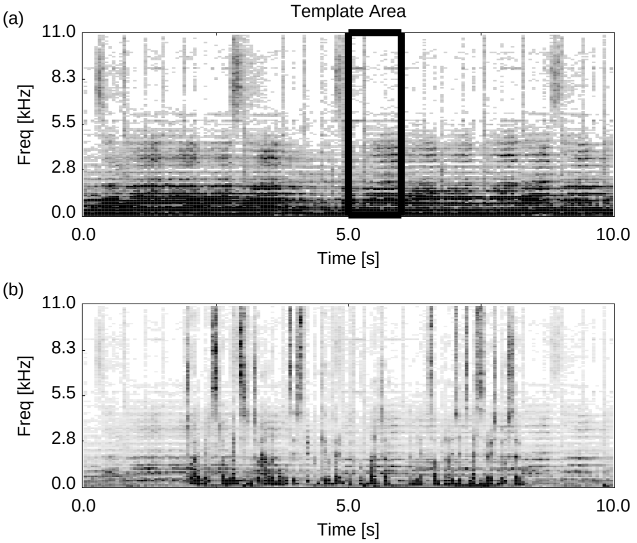

Figure 1: Examples of background music: (a) reference music and (b) music with foreground speech used as input.

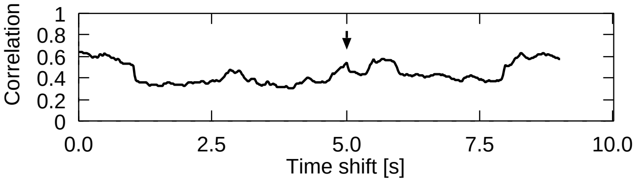

Figure 2: Results for conventional correlation method

2.2. Problem

Let $f(t,\omega)$ and $g(t,\omega)$ be spectrograms of a reference signal and an input signal, respectively, where t is time and ω is frequency. Then the relationship between them can be written as

$$g(t,\omega) = \alpha f(t-\tau,\omega) + n(t,\omega), \tag{1}$$

where α is an amplitude scaling parameter, τ is a time shift parameter, and $n(t,\omega)$ is an error component that includes foreground sound, noise and signal distortion.

Generally, if there exists a pair of parameters (α,τ) that makes $n(t,\omega)$ much smaller than $g(t,\omega)$, then the reference and the input can be said to be similar. But when the SNR is low, the foreground sound prevents $n(t,\omega)$ from becoming sufficiently small.

Assuming that the foreground sound has a time-varying nature, Eq.(1) can be rewritten as

$$g(t,\omega) \approx \begin{cases} \alpha f(t-\tau,\omega) & ((t,\omega) \in S_A) \\ n(t,\omega) & (\text{others}), \end{cases} \tag{2}$$

where S_A is a region in which the target component is much stronger than the error component. Note that only $f(t,\omega)$ and $g(t,\omega)$ are known, while α, τ, S_A and $n(t,\omega)$ are unknown.

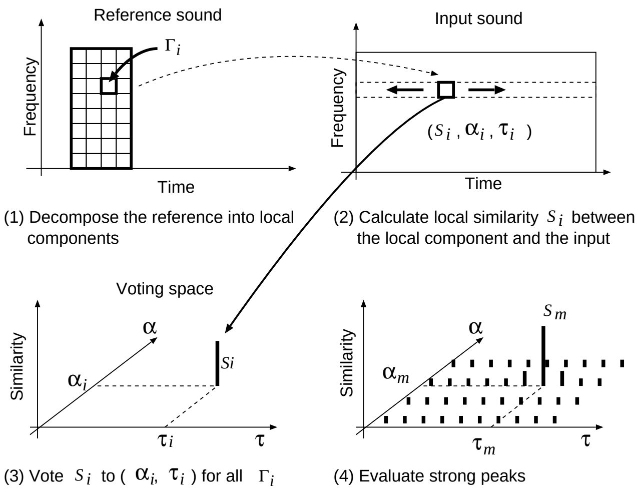
Reference sound
Γ_i
Frequency
Time
(1) Decompose the reference into local components
Input sound
Frequency
(s_i, α_i, τ_i)
Time
(2) Calculate local similarity s_i between the local component and the input
Voting space
Similarity
α
α_i
s_i
τ_i
τ
(3) Vote s_i to (α_i, τ_i) for all Γ_i
Similarity
α
s_m
α_m
τ_m
τ
(4) Evaluate strong peaks

Figure 3: Overview of algorithm

The problem can thus be stated: “Evaluate the similarity between $f(t,\omega)$ and $g(t,\omega)$, by determining α, τ, S_A and $n(t,\omega)$. ”

2.3. Algorithm

Step 1: Decomposition:

The reference is decomposed into local components as follows:

$$f_i(t,\omega) = \begin{cases} f(t,\omega) & ((t,\omega) \in \Gamma_i) \\ 0 & (\text{ others }), \end{cases} \tag{3}$$

$$\Gamma_i = [t_i - \Delta t \le t < t_i + \Delta t] \times [\omega_i - \Delta\omega \le \omega < \omega_i + \Delta\omega], \tag{4}$$

where $i = 1, ..., I$ is an index for the regions; t_i and ω_i are the center time and frequency of the ith region, respectively; and Δt and $\Delta\omega$ are a short time interval and a short frequency interval, respectively (Fig. 3(1)).

Step 2: Calculation of local similarity:

To calculate the local similarity, which is the similarity between the local component and the input (Fig. 3(2)), let

$$J(\alpha,\tau) = \frac{\int_{\Gamma_i} (g(t+\tau,\omega) - \alpha f_i(t,\omega))^2 dt d\omega}{\int_{\Gamma_i} g^2(t+\tau,\omega) dt d\omega} \tag{5}$$

be an evaluation function for the local similarity. It is a second-order error function normalized by the energy of the input. Then, the parameters at the most similar position are obtained by minimizing $J(\alpha,\tau)$, which results in

$$\tau_i = \underset{\tau}{\text{argmax}} \left[\frac{\{\int_{\Gamma_i} f_i(t,\omega) g(t+\tau,\omega) dt d\omega\}^2}{\int_{\Gamma_i} f_i^2(t,\omega) dt d\omega \int_{\Gamma_i} g^2(t+\tau,\omega) dt d\omega} \right], \tag{6}$$

$$\alpha_i = \frac{\int_{\Gamma_i} f_i(t,\omega) g(t+\tau_i) dt d\omega}{\int_{\Gamma_i} f_i^2(t,\omega) dt d\omega}. \tag{7}$$

The local similarity is defined to be

$$s_i \equiv 1 - J(\alpha_i, \tau_i) = \frac{\{\int_{\Gamma_i} f_i(t,\omega) g(t+\tau_i,\omega) dt d\omega\}^2}{\int_{\Gamma_i} f_i^2(t,\omega) dt d\omega \int_{\Gamma_i} g^2(t+\tau_i,\omega) dt d\omega}. \tag{8}$$

Clearly, the similarity is the maximum of the square of the correlation between $f_i(t,\omega)$ and $g(t,\omega)$; and τ_i and α_i are the time shift and scaling parameter at the maximum, respectively. From Eq.(2), it can be seen that the local similarity between a local component in the region S_A and the input should be relatively large when they have the same time shift and scaling, whereas it should be relatively small for other time shifts and scalings.

Step 3: Integration:

All the local similarities are integrated by a voting method (Fig.3(3)). A histogram-like distribution is created in a voting space as follows:

$$H(\alpha,\tau) = \frac{1}{I} \sum_{i=1}^{I} s_i \delta(\alpha - \alpha_i, \tau - \tau_i), \tag{9}$$

where the 2-D delta function $\delta(\alpha,\tau)$ equals 1 when $\alpha = 0$ and $\tau = 0$, and 0 when either α or τ is non-zero.

Since local similarities with the same time shift and scaling are voted to the same position, they form a strong peak in the distribution; whereas those with other time shifts and scalings are scattered. Figure 4 shows the result of voting for the sounds shown in Fig.1. Note that there is a single dominant peak at 5.0s.

Step 4: Evaluation:

To evaluate the distribution in the voting space (Fig.3(4)), the similarity, scaling, and time shift are first estimated from the highest peak of $H(\alpha,\tau)$ as follows:

$$s_m = \max_{\alpha,\tau} H(\alpha,\tau), \quad (\alpha_m,\tau_m) = \operatorname*{argmax}_{\alpha,\tau} H(\alpha,\tau). \tag{10}$$

A comparison of s_m to a predetermined threshold, s_{thsd}, yields the degree of similarity between the reference $f(t,\omega)$ and the input $g(t,\omega)$. For example, the values estimated from the peak in Fig.4 are: $s_m = 0.332$, $\tau_m = 5.0$ s and $\alpha = 0.2$.

Step 5: Determination of similar region:

The region S_A where the components of the reference and the input are similar is determined as follows:

$$S_A = \bigcup_i \Gamma_i \quad \text{s. t.} \quad (\alpha_i,\tau_i) \approx (\alpha_m,\tau_m). \tag{11}$$

Figure 5 shows the selected region of the input shown in Fig.1. Dissimilar regions are masked in black.

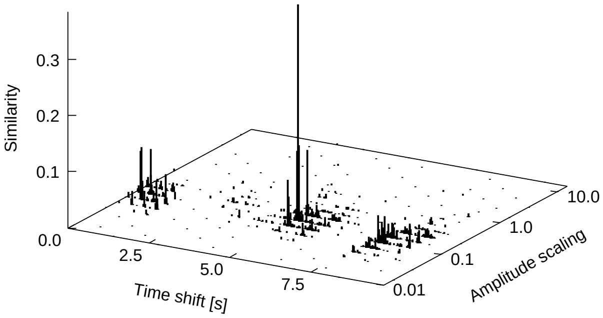

Figure 4: Voting results

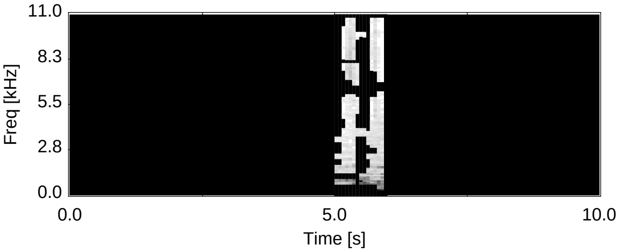

Figure 5: Similar components in the input signal.

3. EXPERIMENTS

3.1. Identification of background music

The first experiment was an attempt to identify background music mixed with foreground speech. Forty music samples and 40 speech samples, all 10 seconds long and sampled at 22 kHz, were selected; and 4 sets, whose SNRs are -10dB, -20dB, -25dB and -30dB, of 40 input sounds were created by mixing them on a computer. The reference sounds consisted of 1-second-long segments were selected at random times from the music samples. An example of an input and a reference are shown in Fig.1.

The size of the decomposition region (Γ_i) was set to 100ms × 720Hz, and the threshold for evaluating the degree of similarity was set to 0.04. The task was to check the similarities (s_m) of all combinations (40 inputs × 40 references for each SNR) and to identify the sounds for which s_m was larger than the threshold.

The results are shown in Table 1. When the SNR was -10dB, all the target pieces were correctly identified. When it was -20dB, 10% of target pieces were not identified though no misidentification occurred. The maximum similarity between false pairs was 0.024.

3.2. Classification of similar TV commercials

In order to make their TV commercials immediately identifiable to viewers and present a consistent theme, some companies put the same type of sounds into all their commercials. An example is shown in Fig.6, in which similar spec-

Table 1: Experimental results 1

SNR of input (dB)	-10	-20	-25	-30
Number of true positives	40	36	28	25
Number of false negatives	0	4	12	15
Number of false positives	0	0	0	0

Table 2: Experimental results 2

Number of true positives	152
Number of false negatives	1
Number of false positives	0

tral components are corrupted by other components.

The second experiment involved the classification of commercials on the basis of common sounds; that is, those from the same company were to be identified. The sound tracks of 62 TV commercials sampled at 12kHz were recorded, and the last 2 seconds of them were manually extracted. The commercials were paired into 1891 sets, 153 of which contained common sounds. The decomposition size was 200 ms × 300 Hz, and the similarity threshold was 0.1. The task was to identify the pairs containing similar sounds.

The results are shown in Table 2. Except for one case, all the pairs containing similar sounds were correctly identified; and no misidentification occurred. The reason for the one false negative was that the lengths of the similar components were different (although they sounded similar to the human ear) and the input was highly distorted. Figure 7 shows the components of the commercials shown in Fig.6 that were determined to be similar. Dissimilar regions are masked in black.

4. DISCUSSION AND SUMMARY

A new method of detecting a reference signal strongly corrupted by other sounds has been devised. The key points are the calculation of local spectral similarities by a correlation method and the integration of local similarities by a voting method. The experimental results have shown that the method can be used not only to identify a clean reference in corrupted input, but also to evaluate the similarity between two corrupted signals.

The voting method explained in this paper yields only the maximum similarity in Eq.(8). However, it can easily be extended to yield more than one similar region in the input signal by using the second (or more) maximum similarity. This allows the detection of multiple occurrences of the same instrument in a single music track. Similar concepts of the voting are seen in the Hough transform for image processing and the computational implementation of the auditory scene analysis[5].

One drawback of the proposed algorithm is that it is computationally very intensive. Using a Sun Workstation (UltraSPARC 2, 300 MHz), about 0.4s was needed to compare just one pair of commercials in the second experiment. For practical use, some scheme for reducing the computational load will be needed to handle large amounts of data.

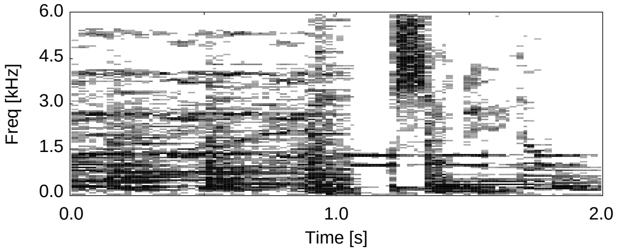

Figure 6: Spectrograms of last 2 seconds of different TV commercials from the same company.

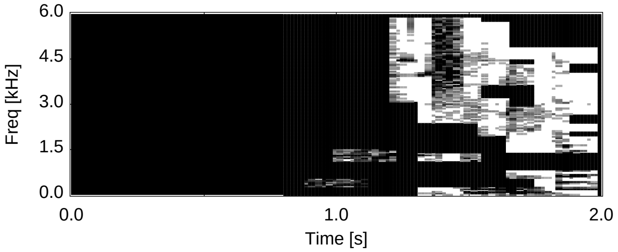

Figure 7: Similar components of commercials.

5. REFERENCES

[1] E. Wold et. al.: "Content-Based Classification, Search and Retrieval of Audio," IEEE Multimedia, 3, 3, 27/36 (1996).

[2] M. Abe et. al: "Content-Based Classification of Audio Signals Using Source and Structure Modeling,", Proc. the 1st IEEE Pacific-Rim Conf. on Multimedia, 280/283 (2000).

[3] K. Kashino et. al.: "Time-Series Active Search for Quick Retrieval of Audio and Video," Proc. ICASSP'99, 2993/2996 (1999).

[4] S. E. Johnson et. al.: "A Method for Direct Audio Search with Applications to Indexing and Retrieval," Proc. ICASSP'00, 1427/1430 (2000).

[5] M. Abe et. al.: "Auditory Scene Analysis Based on Time-Frequency Integration of Shared FM and AM," Proc. IEEE ICASSP'98, 2421/2424 (1998).