J. Acoust. Soc. Jpn. (*E*) **21**, 4 (2000)

ACOUSTICAL LETTER

Quick audio retrieval based on histogram feature sequences

Kunio Kashino, Gavin Smith, and Hiroshi Murase
NTT Communication Science Laboratories,
3-1, Morinosato-Wakamiya, Atsugi, 243-0198 Japan
(*Received 22 November 1999*)

Keywords: Time-series search, Audio search, Audio retrieval, Active search, Histogram modeling
PACS number: 43. 60. −c

1. Introduction

The human auditory system is often considered to have better ability than computers, especially in the context of machine-listening research. It is true that humans seem to surpass computer systems in selective listening in a noisy environment. However, humans have difficulty hearing sounds played faster than real time. This means, for example, that a human searching for a specific theme song in a six-hour video recording would have to listen to the recording in real time, which makes the task very time-consuming. For such a task, computers should be utilized. Therefore, this paper proposes a method to quickly detect and locate a known sound (called a *reference signal*) in a long recording (called an *input signal*).

With multimedia information processing becoming a major research topic, the importance of acoustic information is even more highlighted.[1] In fact, many researchers have tried to utilize sounds for information retrieval from multimedia databases. Most existing approaches aim at content-based retrieval,[2] and includes word-spotting, sound clustering, and indexing.[3] The task discussed in the present paper, however, is different in that we assume that a reference signal is known. We further assume that the amount of spectral distortion is small when the reference signal is included in the input signal.

Apparently, existing techniques such as spectral or waveform matching are applicable to this task. It is difficult, however, for such methods to search through a long-running audio signal without errors in a reasonable processing time, because of the enormous amount of computation required. Of course, the amount of computation can be reduced by introducing heuristic skips in the matching stage, but this inevitably results in omissions.

2. Time-series active search

The active search method, proposed by Vinod and Murase,[4] is a visual search method that detects a region of an input image that matches a reference image using color histograms. In the method, prunning based on a property of histograms reduces the amount of computation compared with exhaustive matching without compromising accuracy. The time-series active search presented here applies the same type of prunning to audio signals.[5]

Figure 1 illustrates the processing flow of the time-series active search. Firstly, a series of feature vectors is extracted from the reference signal and input signal. Secondly, a time window is applied for both signals, and a histogram of feature vectors in the window is created for each. Then, the presence of the reference signal in the current section of the input signal is determined based on whether the histogram similarity exceeds a predefined value (termed a threshold). Here, as explained later, the skip width is calculated from the similarity and the threshold. Thus the search proceeds with the time window for the input signal shifted by that width.

The power spectrum is here employed as the feature vector. The reference and input signals are analyzed by an N-channel band-passfilter bank. Every M input-samples, the square of the output waveform for each band-pass filter is averaged over the samples, and then the ratio of the averaged value with respect to the frequency channels forms an N-dimensional feature vector.

The histogram is created by partitioning each element value of the feature vector into a certain number of bins. Letting b denote the number of bins for each frequency channel, the total number of bins L is given as $L=b^N$. The similarity S between the reference histogram H^R and the input histogram H^I is defined as:

$$S=\frac{1}{D}\sum_{l=1}^{L}\min\left(h_l^R, h_l^I\right), \tag{1}$$

where h_l denotes the number of the feature vectors falling into the l-th bin in the histogram H, and D is the window length (*i.e.* the number of feature vectors in H). S can be viewed as the intersection rate of the histograms.

The skip width w is given by:

$$w=\begin{cases}\mathrm{floor}(P(\theta-S))+1 & (S<\theta),\\ 1 & (\text{otherwise}),\end{cases} \tag{2}$$

where the unit of w is the number of feature vectors, floor $(\cdot)$ denotes rounding down, and θ is the threshold. Equation (2) means that if $S<\theta$ at the present window position, then S can never exceed θ while the window moves to the position which is $w-1$ apart. This can be easily understood if one considers the similarity upper bound; that is, the case where all outgoing feature vectors are irrelevant and all incoming ones are relevant to the histogram intersection as the

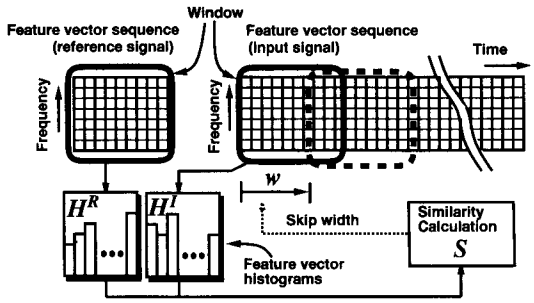

Fig. 1 Overview of the proposed processing.

window moves forward. When $S \geq \theta$, the window does not skip ($w=1$) to find local peaks of the similarity.

In the description so far, the histogram is composed of all of the feature vectors in the window. However, it is possible to incorporate time-order information by dividing the window into a certain number of sub-windows. In such cases, it is natural that the similarity for the whole window be defined as the minimum value among the similarities for each sub-window. In the searching process, it is not always necessary to compute similarities for all sub-windows; for example, when a sub-window has similarity that is less than the threshold, the window can be immediately moved. In such cases, the skip width should be the maximum value of the skip widths for the sub-windows calculated at the current window position.

3. Experiments

3.1 Search time

The time-series active search was implemented on a workstation (SGI O_2) and the search time was measured using an audio signal from television broadcasts. Figure 2 shows a similarity pattern when the sampling frequency was 11.025 kHz, $N=7$, $M=128$, $b=3$, and the number of window divisions was 2 (the duration of reference signal: 15 s). The time required for the search comprises the feature extraction time and the search time based on the extracted feature vectors. The former was approximately 3 min for the case in Fig. 2 (where a 6-h signal and a 15-s signal were processed). The latter depends on particular reference/input signals, the number of window divisions, and the threshold, but for the case in Fig. 2, it took approximately 1 s to search through 6 h (the times are not CPU times but real times).

The ratio of the search time (based on the extracted feature vectors) with the proposed method in comparison with the exhaustive search (w fixed to 1) also depends on the particular reference/input signals, the number of window divisions, and the threshold. In the case in Fig. 2, the proposed method is about 40 times faster than exhaustive search.

In the meantime, it took about 20 min (about 1,000 times slower than the proposed method) for conventional spectral matching based on the feature vectors' inner product to search using the same feature vectors. For waveform matching, the amount of computation will be greater even if the signal is down-sampled to the sampling frequency of several-hundred Hz.

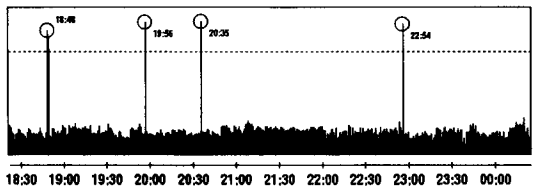

Fig. 2 An example of audio retrieval with the proposed method.

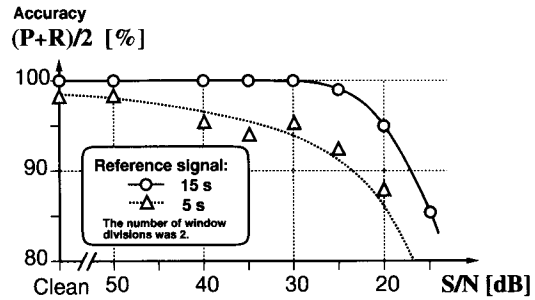

Fig. 3 Accuracy of the proposed method.

3.2 Search accuracy

The search accuracy was evaluated using another television broadcast recording. A number of different television commercials were recorded and edited to 20 min using a video recorder. The 20-min signal was captured on a workstation twice; once as a reference signal source and the second time as an input signal. In each trial, a 5-s segment and a 15-s segment were randomly chosen from the reference signal source and the input signal was searched through. The trial was repeated 100 times. Here the white Gaussian noise was added to the input signal. The search parameters were the same as in the previous section. The experimental results are shown in Fig. 3. The ordinate in the figure corresponds to the average of the precision rate and the recall rate maximized by changing the threshold. It is shown that the search was perfectly carried out when the duration of the reference signal was 15 s even against the noise addition down to a 30 dB signal-to-noise ratio by choosing an appropriate threshold value.

4. Conclusion

This paper has proposed a method of detecting and locating a known reference signal in a long audio signal. An experimental system implemented on a workstation can find a 15-s reference signal in a 6-h television broadcast recording in about 1 s, once the feature vectors are calculated in advance. In addition,

when the duration of the reference signal is 15 s, the search is robust against white Gaussian noise addition down to a 30 dB signal-to-noise ratio. The authors plan to extend the present method to multimedia searches combined with visual information, searches dealing with sound mixture, and searches allowing greater feature fluctuations.

Acknowledgement

The authors wish to thank Dr. Yo'ichi Tohkura, Dr. Ken'ichiro Ishii, Dr. Norihiro Hagita, and our research group members for their help and encouragement.

References

1) B. L. Vercoe, W. G. Gardner, and E. D. Scheirer, "Structured audio: Creation, transmission, and rendering of parametric sound representations," Proc. IEEE **86**, 922–940 (1998).
2) E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-based classification, search, and retrieval of audio," IEEE Multimedia, **3**(3), 27–36 (1996).
3) S. J. Young, M. G. Brown, J. T. Foote, G. J. F. Jones, and J. K. Sparck, "Acoustic indexing for multimedia retrieval and browsing," Proc. ICASSP 97, Vol. 1, 199–202 (1997).
4) V. V. Vinod and H. Murase, "Focused color intersection with efficient searching for object extraction," Pattern Recognition **30**, 1787–1797 (1997).
5) G. Smith, H. Murase, and K. Kashino, "Quick audio retrieval using active search," Proc. ICASSP 98, Vol. 6, 3777–3780 (1998).