# A Robust Audio Searching Method for Cellular-Phone-Based Music Information Retrieval

Takayuki Kurozumi, Kunio Kashino and Hiroshi Murase
NTT Communication Science Laboratories, NTT Corporation
3-1, Morinosato Wakamiya, Atsugi-shi, 243-0198, JAPAN
E-mail: {kurozumi, kunio, murase}@eye.brl.ntt.co.jp

## Abstract

*We propose a search method for detecting a query audio signal fragment in long audio recordings. The query signal is assumed to be captured by a portable terminal, such as a cellular phone, in the real world. A major problem in this kind of search is that the features of the query sound may include distortions due to terminal characteristics or environment noise. The method proposed here comprises local time-frequency-region normalization and robust subspace spanning. The former is used to make features invariant to additive noise and frequency characteristics, and the latter to choose frequency bands that minimize the effect of feature distortions. Experiments using cellular phones in the real world show the proposed method is effective.*

## 1. Introduction

This paper proposes a similarity-based search method for an audio signal database. Specifically, we assume that the database stores music signals and information such as the title and artist, and then people use a cellular phone to look up this information when they hear something they like, and also that the query for the database is a music fragment captured by the cellular phone. We call this scheme cellular-phone-based music information retrieval.

In such retrieval, it is desired that the query signal be a short (e.g. 5 s in length) segment at any arbitrary location in a music piece. Therefore, our approach is based on signal matching and time-shifting, as shown in Figure 1. In the preparation stage, feature vectors are calculated from the stored signal, which is the music signal database. In the search stage, feature vectors are calculated from a given query signal, and the window is applied to the stored feature vectors. The window size is the same as the query signal length. Then, the similarity between the query feature vectors and those in the window is calculated. If the similarity exceeds a threshold value, the query signal is considered to be detected and located. Then, the window is shifted forward in time on the stored signal and the search proceeds until all of the stored signals are scanned. Note that the search here is based on the audio signal rather than symbolic representation [2] such as "notes", because it is still difficult to precisely extract notes in musical pieces.

Assuming that hundreds of thousand of music titles are stored for practical use, and a query signal is captured in a noisy environment, there are apparently two problems. One is the computational cost of searching, and the other is feature fluctuation. For the former, however, a very quick method called Time-series Active Search (TAS) has already been developed [1]. Therefore, this paper focuses on the latter problem.

The feature fluctuation problem has been widely discussed in the literature. For example, in the research aiming at speech recognition in the real world [5], major approaches include microphone arrays, spectral subtraction [4], various noise filtering techniques based on noise modeling, and noise addition to the recognition dictionaries. However, a method applicable to the present task, music search by a cellular phone, has not yet been fully investigated. The essential problem is that in cellular-phone-based music information retrieval the noise characteristics greatly vary with the user's environment. For example, as shown in Figure 2, the power spectrum of a specific music segment varies significantly depending on the recording situation and devices used. Our task here is to recognize the signal as the same one regardless of spectral variations, and at the same time, to distinguish different sections of music. For this purpose, we propose a robust search method that features two techniques to make features invariant to the terminal characteristics and environment noises. One technique, local time-frequency-region normalization, is used to absorb the additive noises and the frequency characteristics. The other, robust subspace spanning, is used to choose the frequency bands that minimizes the effect of feature distortions.

1051-4651/02 $17.00 (c) 2002 IEEE

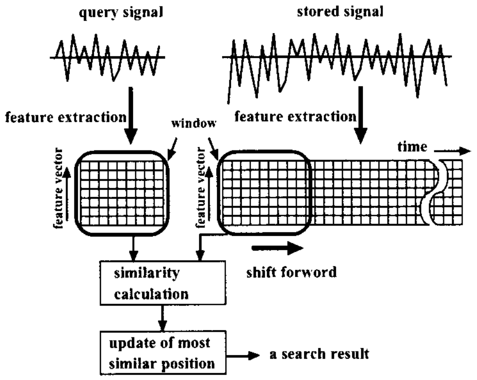

**Figure 1. Overview of time-series search**

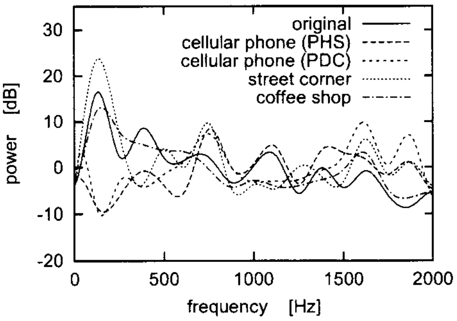

**Figure 2. A segment of a music piece**

This paper is organized as follows: Section 2 describes our methods. Section 3 evaluates the search accuracy, and, finally, Section 4 gives some conclusions.

## 2. Method

### 2.1. Feature extraction

As feature vectors, the zero crossing rates, short-time power spectrum, LPC cepstrum, and MFCC (Mel frequency cepstral coefficients) can be considered [3]. In this paper, however, we simply use the FFT-based short-time power spectrum, as our focus is on normalization and subspace spanning. Let $P(t, k)$ be the short-time power spectrum of the audio waveform $x(t)$ at time $t$ and frequency $k$, and $\boldsymbol{P}(t)$ be the vector whose elements are $P(t, k)$. Then, the $i$-th frequency feature vector $\boldsymbol{Q}(i)$ is defined as

$$\boldsymbol{Q}(i) = \boldsymbol{P}(si) \qquad (1)$$

where $s$ is the amount of the window shift for the analysis. In the following experiments, the sampling frequency was 8000 Hz, the FFT window length was 8192, and the parameter $s = 512$.

### 2.2. Local time-frequency-region normalization

One of the keys to robust music retrieval is feature normalization. Usually, normalization is done by using the average (or maximum) power in an analysis window. However, it is obvious that features extracted by such a conventional scheme is not invariant to noises or frequency characteristics effectively.

In our approach, normalization is done with regard to the local time-frequency region. The $k$-th normalized frequency feature $\boldsymbol{y}(i)$ is defined as

$$y(i, k) = \frac{1}{\sigma(i, k)}(Q(i, k) - m(i, k)) \qquad (2)$$

where

$$m(i, k) = \frac{1}{2M}\sum_{i=-M}^{M-1} Q(i, k) \qquad (3)$$

$$\sigma(i, k)^2 = \frac{1}{2M}\sum_{i=-M}^{M-1} (Q(i, k) - m(i, k))^2. \qquad (4)$$

The $M$ is half of window size for calculation of the mean value and the standard deviation. We chose $M = 16$ in the following experiments.

The idea of subtracting the mean value $m(i, k)$ is similar to that of CMN (cepstrum mean normalization), which is often employed in real-world speech recognition systems [5]. This subtraction is expected to cancel the additive feature fluctuations. Here, we further introduce the normalization by $\sigma(i, k)$, in order to make the features more invariant to the complicated frequency characteristics of cellular phone terminals.

### 2.3. Robust subspace spanning

The normalized feature vectors are mapped to a subspace. The subspace is created so that the feature vector variation due to the feature distortion becomes small, and that due to the audio content becomes great, as a result of the mapping. This is basically done by Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA), and our method is based on PCA. However, instead of performing PCA for all learning vectors together, we calculate the mean vector over the various distortion types for each original (non-distorted) vector. By doing this noise-averaging calculation, the subspace is expected to be more robust than one created only with simple PCA or LDA.

The procedure of our subspace spanning is almost the same as the standard PCA, except for the noise-averaging calculation. First, $L$ original signal segments are extracted from signals such as music CDs. Then, $C$ signals are prepared for each segment. They include the original signal and the distorted signals. The distorted signals are obtained, for example, by playing the original signal in various environments using a loudspeaker whose characteristics are known, and then capturing the sound by cellular phones. The resulting distorted frequency features are written as $y_{lc}$, where $l$ is the order of segments, and $c$ is the distortion type.

Then we calculate the scatter matrix $R$:

$$R = \frac{1}{L}\sum_{l=1}^{L}(\overline{y}_l - \overline{y})(\overline{y}_l - \overline{y})^t, \qquad (5)$$

where the mean value $\overline{y}_l$ for each class is

$$\overline{y}_l = \frac{1}{C}\sum_{c=1}^{C} y_{lc}, \qquad (6)$$

and the mean value for all classes is

$$\overline{y} = \frac{1}{L}\sum_{l=1}^{L}\overline{y}_l. \qquad (7)$$

Finally, the eigenvectors of $R$ are calculated. The subspace is spanned using the eigenvectors $\phi_k$. Let $z$ stand for the feature vector used for searching. Then, the $k$-th element of the vector is written as

$$z_k = y\phi_k, \qquad (8)$$

where $y$ is a normalized frequency feature as described in Section 2.2.

Note that the proposed method is quicker than using PCA or LDA without distortion average calculation, when the same number of learning vectors are used.

## 3. Experiments

The proposed method was tested with regard to search accuracy. Figure 3 shows the experiment setup[1]. We prepared a set of query signals captured by five devices in seven places in real environments (Table 1), and a database of 200 music pieces (Table 2).

Firstly, we performed two experiments to evaluate the advantages of the proposed method using limited data sets: one to evaluate the effect of the proposed normalization (exp. 1), and the other to evaluate the subspace spanning method (exp. 2). We then tested the proposed method under more realistic circumstances using full data sets (exp. 3).

All experiments were done in the nearest-neighbor scheme. The correct retrieval result was defined as the retrieval of the correct position rather than just the music title. The retrieval position was judged correct if the location error was less than 5 s.

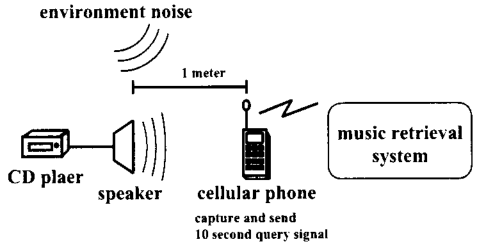

**Figure 3. Experiment setting**

**Table 1. Signals captured for queries**

| | |
|---|---|
| contents | 20 minutes (34 titles including rock, pop, jazz and classical music)<br>10 minutes for learning and<br>the other 10 minutes for the tests |
| places | 4 places (noise-level 1)<br>an office room, an idling car,<br>a convenience store, a karaoke bar<br>3 places (noise-level 2)<br>a crowded street corner, a noisy coffee shop,<br>a busy-traffic intersection |
| devices | 2 cellular phones (PHS),<br>3 cellular phones (PDC) |

**Table 2. Test database**

| | |
|---|---|
| contents | 13 hours (200 titles) |

### 3.1. Effect of local time-frequency-region normalization (exp. 1)

First, we performed a search accuracy test using local time-frequency-region normalization of each local time-frequency area. We used 10-m original signal including 17 titles in the database and 200 query signals of a 10-s interval selected at different positions. We prepared 8 kinds of query signals, recorded using 2 PHS terminals at the four noise-level-1 places in Table 1. The total number of search tests was 1,600. A 0-2000 Hz band was divided into 8 sub-bands, and 8-dimensional feature vectors were calculated by the mean power of each sub-band. Then, 80-dimensional vectors composed of the elements of ten feature vectors extracted every second was used for the search test. We used the Euclid distance as the similarity measure. Figure 4 shows an example of search. Table 3 shows the search results. The CCR, cumulative classification rate, was calculated by choosing the five most similar segments. We compared the proposed method with a method using feature vectors normalized by using the power of the 0-2000 band. The proposed normalization achieves higher search accu-

[1] PHS (Personal Handy-phone System) is a cellular phone service in Japan, based on 32 kbps ADPCM encoding. PDC (Personal Digital Cellular system) is a cellular phone service based on 6.7 kbps VSELP encoding.

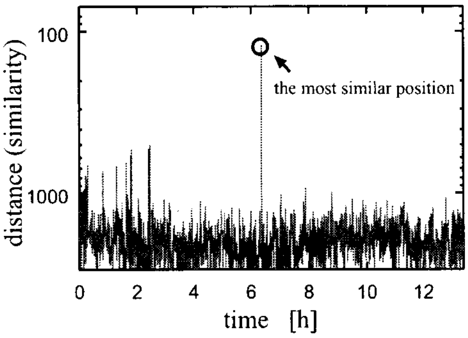

**Figure 4. An example of search**

**Table 3. Effect of local time-frequency-region normalization (exp. 1)**

| | accuracy | CCR |
|---|---|---|
| normalization using power only | 29.7% | 41.9% |
| proposed normalization | 74.4% | 84.4% |

CCR: cumulative classification rate

racy than the normalization using the power; the accuracy improved from 29.7% to 74.4%.

### 3.2. Effect of robust subspace spanning (exp. 2)

Next, we performed a search accuracy test using subspace spanning. We used the 10-m contents for learning in Table 1. 300 samples for learning were selected on different positions. We performed a search accuracy test using the other 10-m contents in Table 1. We compared the proposed method with three methods: a method using feature vectors calculated by the mean power of each sub-band as described in Section 3.1, PCA, and LDA. The proposed method achieves higher search accuracy than the other methods; the accuracy improved from 74.4% to 79.1%.

**Table 4. Effect of robust subspace spanning (exp. 2)**

| | accuracy | CCR |
|---|---|---|
| mean power | 74.4% | 84.4% |
| PCA | 73.3% | 84.9% |
| LDA | 67.1% | 78.8% |
| proposed | 79.1% | 87.8% |

**Table 5. Search Accuracy (exp. 3)**

| noise | level 1 | level 2 |
|---|---|---|
| cellular phone (PHS) | 83.4% | 44.0% |
| cellular phone (PDC) | 63.6% | 32.0% |

### 3.3. Search accuracy (exp. 3)

Finally, we performed a search accuracy test using a database of 200 music pieces. We used all of the signals in Table 1 as query signals. In this experiment, 32-dimensional feature vectors were used. As shown in Table 5, the accuracy was 83.4% when the query signals were captured by cellular phones (PHS) in noise-level-1 places.

## 4. Conclusion

This paper has described an audio search method to retrieve music information by query signals captured with cellular phones. Specifically, we have proposed local time-frequency-region normalization to make features invariant to additive noise and frequency characteristics, and robust subspace spanning to choose frequency bands that minimize the effect of feature distortions. Experiments using audio signals received in the real world prove the effectiveness of the proposed method. In a test under realistic circumstances, for example, the search accuracy was 83.4% when a 13-h audio recording was searched by query signals captured by cellular phones (PHS) in the real world. We consider the results promising for realizing such a music information retrieval system. Our future work includes an investigation of non-stationary noise absorption as well as enlargement of the database scale.

**Acknowledgments:** The authors thank Dr. Ken'ichiro Ishii and Dr. Noboru Sugamura for their help and encouragement.

## References

[1] K. Kashino, G. Smith and H. Murase. "Time-series Active Search for Quick Retrieval of Audio and Video". *Proc. of ICASSP'99*, 6:2993–2996, Mar. 1999.

[2] K. Lemstrom and S. Perttu. "SEMEX - An efficient Music Retrieval Prototype". *MIR 2000*.

[3] L. R. Rabiner and B. H. Juang. *Fundamentals of speech recognition*. Prentice Hall, 1993.

[4] S. F. Boll. "Suppression of acoustic noise in speech using spectral subtraction". *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-27(2):113–120, 1979.

[5] S. Furui. "Cepstral analysis technique for automatic speaker verification". *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-29(2):254–272, 1981.