CONTENT-BASED CLASSIFICATION OF AUDIO SIGNALS USING SOURCE AND STRUCTURE MODELING

Mototsugu Abe, Jun Matsumoto and Masayuki Nishiguchi

HNC Development Center, Sony Corporation
6-7-35 Kita-Shinagawa, Shinagawa-ku, Tokyo 141-0001, JAPAN
abe@av.crl.sony.co.jp

ABSTRACT

This paper describes a novel method of classifying audio signals for multimedia content-based retrieval (CBR). In order to classify complex audio streams into simple categories, two classification levels for short audio clips are introduced: one is based on the type of sound source, such as speech, music, noise etc.; and the other is based on the type of sound structure, such as the number of sources, the overlap of sources, a change of sources, etc. A classification algorithm is constructed that efficiently exploits features related to both source and structure types. To categorize audio clips, their features are classified using a simple k-means classifier. A performance evaluation of the proposed algorithm revealed the success rate to be as high as about 90%.

1. INTRODUCTION

The content-based classification, search and retrieval of video [1] and audio[2] are key issues in the handling of multimedia data on large-capacity storage media, like hard-disk recorders, which will probably replace video cassette recorders in the near future.

For audio classification, Wold[2] proposed a system for classifying audio clips a few seconds long into such categories as voice, bell, telephone, animal, etc., utilizing statistics on instantaneous features extracted from both waveform and spectrogram. Lambrou[3] devised a genre classification method for music based on statistical features in the time and wavelet domains.

Although both methods exhibit excellent performance for single sources, sources in actual TV and radio programs often overlap and alternate with other sources, for examples speech over traffic noise, speech over background music (BGM), conversation, talk during a music program, and so on. To solve the problem, Minami[4] proposed a system of classifying overlapping music and speech based solely on the stability of horizontal stripes in a spectrogram. The system exhibits good performance for some limited kinds of instrumental music.

This paper describes a new classification method based on source and structure modeling. In order to classify complex audio streams into simple categories, two classification levels for short audio clips are introduced: one is based on the type of sound source, and the other is based on the type of sound structure regardless of the source. A classification algorithm is constructed that divides a continuous stream into short clips, extracts the features that efficiently describe both the type of source and the type of structure, and then classifies the features using a simple k-means classifier.

2. FRAMEWORK

2.1. Time Block Division

A continuous audio stream is first divided into short clips. The length of a clip is determined so as to satisfy the following conditions: 1) long enough to allow identification of the source category, and 2) short enough to allow the structure of a complex stream to be simplified.

A period of 1 second was selected for the length of a block, with half-block overlaps (Fig.1), because 1) a human can identify most of the source categories in this period, and 2) in the real world, sources rarely alternate more than twice in this period.

![Figure 1: Time block division.](image)

2.2. Source Categories

Source types can be categorized into many different levels. For example, music can be thought as a single cat-
egory, as well as a combination of many subcategories of musical instruments. For simplicity, the following broad categories are used, rather than actual physical sources.

1. Speech: including male, female, child and others;
2. Music: including vocal, instrumental and other types;
3. Impact: the sound made by one object sharply striking another, such as a knock on a door, a small number of handclaps, footsteps, etc.;
4. Noise: continuous sound similar to white noise;
5. Environmental Sound: sound that does not fall into the above categories, including crowd noise, traffic noise, cheering, the sound of waves, etc.

2.3. Structural Categories

The structural classification is independent of the above source classification. The basic categories are as follows (Fig.2):

1. Silent: no sound within the entire block;
2. Single: sound from a single source;
3. Double: two overlapping sources;
4. Change: a change from one source to another;
5. Simultaneous Change: the simultaneous change of two sources;
6. Partial Change: the change of one source overlapping another continuous source.

Change, Simultaneous Change and Partial Change include the change of a source to or from Silent. To handle cases where none of the above categories is suitable, e.g. three overlapping sources, we define

7. Others: structures that do not fall into any of the above categories.

3. FEATURE EXTRACTION

Our experiments revealed that no single feature provides sufficient information to uniquely determine both the source and structural categories in all cases. That means that the determination must be based on sets of features that are statistically related to sources and structures. Twelve features (described below) extracted from amplitude, spectrogram and pitch are employed.

<table>
<thead>
<tr>
<th>Silent</th>
<th>Single</th>
<th>Double</th>
</tr>
</thead>
<tbody>
<tr>
<td>no sound</td>
<td>single source sound</td>
<td>two overlapping sources</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Change</th>
<th>Simultaneous Change</th>
<th>Partial Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>change from one source to another</td>
<td>simultaneous change of two sources</td>
<td>change of one source + continuous source</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Others</th>
<th>Others</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>two changes of source</td>
<td>3 overlapping sources</td>
<td>other complex structures</td>
</tr>
</tbody>
</table>

Figure 2: Categories for structural classification

Let $s_i(t) (0 \leq t \leq T)$ be an audio signal and let $S_i(t, \omega) (0 \leq t \leq T, 0 \leq \omega \leq \Omega)$ be the spectrogram (amplitude of STFT with a window length of 31 ms) of the $i$th block, where $T(=1s)$ is the block length and $\Omega(=8kHz)$ is the maximum frequency. For convenience of notation, we define

\[
AV[x(t)] = \frac{1}{T} \int_{0}^{T} x(t) dt
\]

(1)

to be the average and

\[
RSD[x(t)] = \frac{1}{AV[x(t)]} \sqrt{\frac{1}{T} \int_{0}^{T} (x(t) - AV[x(t)])^2 dt}
\]

(2)
to be the relative standard deviation (RSD) of the function $x(t)$.

1. **Amplitude**: Since the amplitude of speech or a knocking sound changes rapidly, while that of music and noise remains fairly constant, the average and the RSD of the amplitude is used to distinguish between these categories. They are given by

\[
P_a = AV[|s(t)|], \quad P_s = RSD[|s(t)|].
\]

(3)

2. **Spectral Width**: Because of the alternation between vowels and consonants, the spectral width of speech changes frequently, while that of noise and music is fairly constant. So, the distinctive features are the average and the RSD of the width of the spectrum whose energy exceeds a given threshold. They are given by

\[
W_a = AV[w(t)], \quad W_s = RSD[w(t)],
\]

(4)

where

\[
w(t) = \frac{1}{\Omega} \int_{\Gamma} d\omega, \quad \Gamma = \{\omega | S(t, \omega) > \text{Threshold}\}
\]

(5)
denotes the instantaneous spectral width.
3. **Low-Frequency Power**: Since the spectral components of speech are generally limited to the frequency range from 70 Hz to 4 kHz, low-frequency (below 70 Hz) components can be used to characterize non-speech sounds. The average and the RSD of the amplitude of these components are given by

\[ L_a = AV[l(t)], \quad L_s = \text{RSD}[l(t)] \]

where

\[ l(t) = \frac{\int_{\omega_0}^{\infty} S(t, \omega)d\omega}{\int_{0}^{\infty} S(t, \omega)d\omega}, \quad (\omega_0 = 70\, \text{Hz}) \]

denotes the normalized power of the low-frequency component at a given time. These features differentiate music with a bass drum from other sounds.

4. **High-Frequency Power**: Similarly, the average \( H_a \) and the RSD \( H_s \) of the amplitude of high-frequency (above 4 kHz) components can also be used to characterize non-speech sounds, in particular, music with percussion instruments.

5. **Fundamental Frequency**: The average and the RSD of the fundamental frequency is given by

\[ F_a = AV[f(t)], \quad F_s = \text{RSD}[f(t)] \]

where \( f(t) \) is the fundamental frequency calculated by Parson’s voting method[5]. These values allow us to estimate the dominant fundamental frequency in a mixture of harmonic signals.

6. **Harmonicity**: The average and the RSD of the harmonicity is given by

\[ A_a = AV[a(t)], \quad A_s = \text{RSD}[a(t)] \]

where

\[ a(t) = \frac{\int_{\Gamma} S(t, \omega)d\omega}{\int S(t, \omega)d\omega}, \]

\[ \Gamma = \{ \omega \mid n f(t) - \Delta \leq \omega \leq n f(t) + \Delta, n = 1, 2, \ldots \} \]

is the normalized energy of the harmonic components. \( \Delta \) is a small frequency (actually 15 Hz). These features are very powerful tools for discriminating source types because \( a(t) \) is high for speech due to the vowel sounds, medium for music when there are several instruments, and low for noise due to the lack of harmonicity.

Finally, all the features are combined into a 12-dimension feature vector that characterizes a clip:

\[ X_i = [P_a, P_s, W_a, W_s, \ldots, A_a, A_s] \]

In order to distinguish structures involving a change of sources, another feature vector based on the vectors for the preceding and following blocks is employed:

\[ Y_i = \frac{|X_{i+1} - X_{i-1}|}{X_{i+1} + X_{i-1}}, \]

where the division and absolute value operations are carried out on each component. A component of \( Y_i \) is 0 when the difference between the two blocks is small, and 1 when the difference is large.

4. **CLASSIFICATION**

The classification of feature vectors involves three steps (Fig.3), with the first two focusing on the structure, and the last focusing on the source. In the first step, sounds belonging to the Silent category are distinguished by using a simple thresholding method on \( P_a \). Next, the remaining sounds are broken down into two groups using the vector \( Y_i \): the Changing group includes Change, Simultaneous Change and Partial Change; and the Continuous group includes Single, Double and Others. To make it easier to identify sounds belonging to the Changing group, information from the preceding and following blocks is used. Finally, the Continuous group is broken down into the various types of sources.

Vector quantization (VQ) is employed in the second and the third steps. VQ codebooks for each category are independently learned by means of a k-means algorithm[6]. The incoming feature vector \( X_i \) (or \( Y_i \)) is classified into the category that the Mahalanobis distance between the vector and a centroid of the category is a minimum.

![Figure 3: Steps in classification procedure](image)

5. **EXPERIMENTS**

Japanese TV programs, including news, drama, baseball, commercials, etc., were recorded and a database of
2,997 clips was created. Each clip is 1 second long. All
the clips were hand-labeled according to both source and structure types. The composition of the data is shown in Table 1. Two-thirds of them were used for learning, and the rest were used for evaluation.

5.1. Silent Category (Step 1)
The threshold for detecting the Silent category was set to 40 dB below the maximum amplitude of all the clips. The overall success rate for this category was 99.7%. All Silent clips were correctly identified, but a few clips with a little white-like noise were misidentified as Silent.

5.2. Changing vs. Continuous (Step 2)
After all the Silent clips had been excluded, the remainder were broken down into Changing and Continuous groups. Four centroids were used for the VQ of each category. The classification results are shown in Table 2.

About 22% of the clips labeled Changing were wrongly classified as Continuous. A detailed analysis showed that almost all of the ones involving a change to/from Silent were classified correctly, while some clips labeled Partial Change were misidentified as Continuous.

5.3. Source Classification (Step 3)
To get an independent estimate of the success rate for source classification, all clips in the Silent category and the Changing group were excluded, and the remainder were classified. Eight centroids were used for each source category. The classification results are shown in Table 3.

The overall success rate was about 87%. As expected, the confusion rates among Speech or Music and BGM were high. Since the Noise category includes all sounds other than speech and/or music, the success rate for this category was lower than for the others. A detailed analysis showed that constant white-like noise was sometimes mistaken for Music.

6. DISCUSSION AND SUMMARY
Experimental results have demonstrated the great potential of a classification scheme based on source and structure modeling. For practical applications, the use of other features may need to be explored. In particular, the success rates for the Changing group and the identification of the BGM category should be higher. An investigation of new structural categories may also lead to an improvement in overall performance.

Table 1: Composition of database

<table>
<thead>
<tr>
<th>Structure</th>
<th>Source</th>
<th>Number of clips</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silent</td>
<td>—</td>
<td>200</td>
</tr>
<tr>
<td>Single</td>
<td>Speech</td>
<td>1034</td>
</tr>
<tr>
<td></td>
<td>Music</td>
<td>873</td>
</tr>
<tr>
<td>Noise^1</td>
<td></td>
<td>130</td>
</tr>
<tr>
<td>Double</td>
<td>BGM^2</td>
<td>446</td>
</tr>
<tr>
<td>Change^3</td>
<td>any</td>
<td>314</td>
</tr>
</tbody>
</table>

1 Noise includes Impact, Noise and Environmental Sound.
2 BGM means Speech and Music.
3 Change includes Simultaneous and Partial Change.

Table 2: Confusion matrix for Changing and Continuous groups

<table>
<thead>
<tr>
<th></th>
<th>Classification result</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Changing</td>
</tr>
<tr>
<td>Changing</td>
<td>77.8%</td>
</tr>
<tr>
<td>Continuous</td>
<td>4.8%</td>
</tr>
</tbody>
</table>

Table 3: Confusion matrix for source classification

<table>
<thead>
<tr>
<th></th>
<th>Speech</th>
<th>Music</th>
<th>Noise</th>
<th>BGM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech</td>
<td>89.7%</td>
<td>0.0%</td>
<td>0.3%</td>
<td>10.1%</td>
</tr>
<tr>
<td>Music</td>
<td>0.7%</td>
<td>87.2%</td>
<td>2.4%</td>
<td>9.7%</td>
</tr>
<tr>
<td>Noise</td>
<td>2.4%</td>
<td>9.5%</td>
<td>81.0%</td>
<td>7.1%</td>
</tr>
<tr>
<td>BGM</td>
<td>8.0%</td>
<td>4.0%</td>
<td>1.3%</td>
<td>86.8%</td>
</tr>
</tbody>
</table>

7. REFERENCES