

Using Structure Patterns of Temporal and Spectral Feature in Audio Similarity Measure¹

Rui Cai

Department of Computer Science
and Technology, Tsinghua Univ.
Beijing, 100084, China

cairui01@mails.tsinghua.edu.cn

Lie Lu, Hong-Jiang Zhang

Microsoft Research Asia
No. 49 Zhichun Road
Beijing, 100080, China

{llu, hjzhang}@microsoft.com

Lian-Hong Cai

Department of Computer Science
and Technology, Tsinghua Univ.
Beijing, 100084, China

clh-dcs@tsinghua.edu.cn

ABSTRACT

Although statistical characteristics of audio features are widely used for similarity measure in most of current audio analysis systems and have been proved to be effective, they only utilized the averaged feature variations over time, and thus lead to inaccuracy in some cases. In this paper, structure pattern, which describes the representative structure characteristics of both temporal and spectral features, is proposed to improve the similarity measure for audio effects. Three kind structure patterns are proposed and utilized in current work, including energy contour pattern, harmonicity pattern and pitch contour pattern. Evaluations on a content-based audio retrieval system indicate that structure patterns can improve the performance pretty much.

Categories and Subject Descriptors

H.5.5 [Information Interfaces and Presentation]: Sound and Music Computing - *signal analysis, synthesis and processing*;

I.5.1 [Pattern Recognition]: Models - *structural*

General Terms

Algorithms, Measurement, Design, Experimentation, Theory

Keywords

Audio similarity measure, structure pattern, audio retrieval

1. INTRODUCTION

Similarity measure is a fundamental step in content-based audio analysis, such as audio classification [4], audio retrieval [2] and audio scene analysis [6]. In most of current audio analysis systems [2][6], similarity measure is based on statistical characteristics of the temporal and spectral features of each frame; and the statistics, including mean, standard deviation or covariance, are used to describe the property of an audio clip. These statistical features have proved their effectivity in many previous works. However, they only utilized the averaged feature variations over time, but ignored the detail status in each time slot or frequency band and the variation trend of each feature, and thus lead to inaccurate similarity measure in some cases.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MM'03, November 2-8, 2003, Berkeley, California, USA.
Copyright 2003 ACM 1-58113-722-2/03/0011...\$5.00.

For instance, Figure 1(a) illustrates two different sounds which have a close similarity based on their statistical characteristics only. The left part is spectrogram and energy envelope of a sound of "car crash", which is a sudden bang followed by a series of decrescendo effects of things broken. The right part is a sound of "surf", which describes a gradually approaching ocean wave which finally impacts the coast. Although these two sounds are absolutely different from human perception, they have very similar statistical characteristics of both temporal and spectral features, such as short time energy, zero crossing rate, and spectral centroid. On the other side, using statistical features only may also make two audio clips of the same sound different. Figure 1(b) shows such an example, where both the left and right are sounds of "jet plane", which describe a jet plane flying over the heads. The statistical features of them are not as similar as expected. For example, the left one has higher spectrum energy, especially in high-frequency band; and its energy is more concentrated in temporal domain, which makes the derivation much larger. It results in a relatively far distance in calculating their similarity.

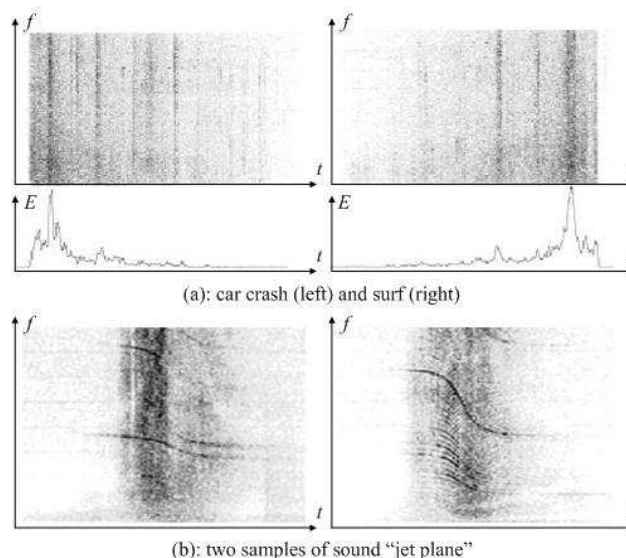


Figure 1. Illustration of some sound effects (a) different sounds with similar statistical features; (b) similar sounds with different statistical features

To complement the disadvantages that statistical features only represent the average information, feature structure pattern is

¹This work was performed when the first author was a visiting student in Media Computing Group, Microsoft Research Asia

proposed in this paper to improve the similarity measurement. Feature structure pattern means the representative pattern which describes the structure characteristics of both temporal and spectral features, such as the energy envelope and pitch contour pattern. Some psychophysical researches [1] have indicated that these patterns act an important role in human perception of sound objects. For instance, the pattern of energy envelope can help to distinguish sounds illustrated in Figure 1(a), where the envelope of “car crash” has a rapid attack and gradual decay, while that of the “surf” rise slowly but decay rapidly; for the similar sounds in Figure 1(b), both of them have a ‘z’ shape pattern in spectrum. In this paper, several basic and typical structure patterns of both temporal and spectral features are presented.

The rest of this paper is organized as follows. Temporal and spectral structure patterns are described in Section 2. The corresponding similarity measurements are presented in detail in Section 3. In Section 4, experiments and evaluations are given.

2. STRUCTURE PATTERNS

Gygi [1] deeply investigated the acoustic factors involved in the identification of sound effects. His perception experiments revealed that structure of the envelope and harmonicity are important for sound effects identification. In this section, the structure pattern of energy envelope, harmonicity and pitch contour are defined respectively.

2.1 Energy Envelope Pattern

As mentioned above, human’s auditory system is sensitive to the sound’s energy envelop, which represents the development procedure of a sound effect. In our approach, a polynomial curve fitting based clustering method is proposed to extract representative energy envelope shapes, which are defined as the energy envelope structure pattern.

Firstly, the energy envelopes of all audio clips in our database are extracted and normalized into a same length. For each normalized envelop, a polynomial $p(x)$ of degree n is used to fit it:

$$p(x) = p_1x^n + p_2x^{n-1} + \dots + p_nx + p_{n+1} \quad (1)$$

All the $n+1$ coefficients of $p(x)$ are estimated by solving the least-squares problem. Thus, each envelope curve can be represented by a vector of $(n+1)$ dimension. In our experiments, n is empirically chosen as 10, since too small polynomial orders give poor simulation to some detail, while larger orders will lead to numerical instability in fitting processing.

Unsupervised k -means clustering algorithm is then performed on all the envelope vectors to find the cluster number with the minimum clustering errors. Finally, five representative energy envelope patterns are obtained, as illustrated in Figure 2. The essence of these patterns is mainly characterized by the *attack*, *sustain* and *decay* phases of a sound.

- Pattern *a*. This pattern is characterized by a large increase in energy with a brief attack time, which is followed by an extended sustained period, and subsequently by a gradual decay. One example is a sound of car crash.
- Pattern *b*. It can be considered as the symmetrical pattern of pattern *a*, with a gradual increase, followed by a period of sustain and a rapid decrease.

- Pattern *c*. Sounds in this pattern often have a long time sustain, such as applause.
- Pattern *d*. It has a quick increase followed by a quick decay, almost without any sustain, such as the sound of gun shot.
- Pattern *e*. It is with a gradual increase, followed by a brief sustain and subsequently by a gradual decrease.

Thus, the energy envelope structure pattern of each audio clip can be denoted as:

$$Eng = [e_1], \quad e_1 \in \{a, b, c, d, e\} \quad (2)$$

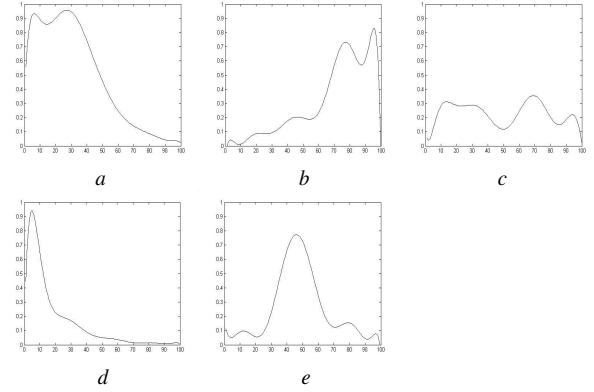


Figure 2. Five typical energy envelope patterns

2.2 Spectral Structure Pattern

To represent spectral structure, two kind patterns, which are on harmonicity and pitch contour, are extracted from a narrow-band spectrogram. Considering that only the main sketch of spectrogram is meaningful for patterns extraction, the region whose value is lower than the average spectral intensity is set to zero, in order to remove the noise in the spectrogram.

2.2.1 Harmonicity Pattern

Unlike the harmonicity ratio defined in [3], which describes the percent of harmonic frames over time, the harmonicity pattern is designed to detect harmonic status in each frequency sub-band and represent harmonicity distribution in spectral domain. In our approach, six sub-bands are used, including $[0, \frac{\omega_0}{2^6})$, $[\frac{\omega_0}{2^6}, \frac{\omega_0}{2^5})$,

... $[\frac{\omega_0}{2^2}, \frac{\omega_0}{2^1}]$, where ω_0 is the sample rate of audio clip.

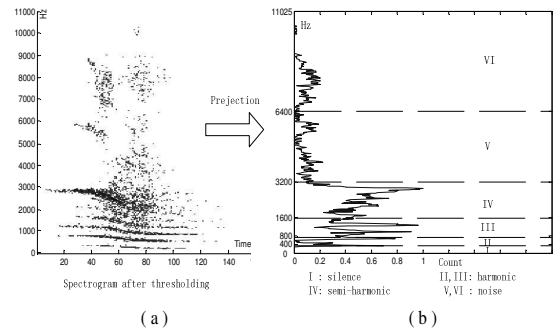


Figure 3. Definition of harmonicity pattern

The harmonic status in each sub-band is detected and assigned to one of the following patterns: *silence*, *noise*, *semi-harmonic* and *harmonic*, where “*silence*” means that there is little energy in the sub-band; “*noise*” denotes the non-harmonic region; “*harmonic*” represents the sub-band with a clear harmonic structure; and “*semi-harmonic*” is a combination of noise and harmonic. Figure 3 illustrates an example spectrogram and the corresponding harmonicity patterns in different sub-bands. Thus, the harmonicity pattern of an audio clip can be represented as a vector:

$$\begin{aligned} Har &= [h_1, h_2, \dots, h_6] \\ h_i &\in \{silence, noise, semi-harmonic, harmonic\}, 1 \leq i \leq 6 \end{aligned} \quad (3)$$

The harmonicity pattern detection is based on the projection of spectrum on frequency axis, which is normalized to $[0, 1]$ by divided by the maximum value in the projection curve, as illustrated in Figure 3(b). The detail pattern detection procedure is illustrated in Figure 4. For each sub-band, the average spectrum energy and the number of prominent peak in the curve are first calculated, where prominent peak is defined as a peak whose value is larger than a pre-defined threshold, which is set as 0.2 in our approach. For those sub bands whose prominent peak number is larger than zero, if energy is less than a threshold E_H , the sub-band is classified into *harmonic*; otherwise it is assigned with the label *semi-harmonic*. For those sub-bands having no prominent peaks, if energy is less than another threshold E_L , the sub-band is classified into *silence*; otherwise, it is assigned with label *noise*. In our experiments, the thresholds are set as:

$$\begin{aligned} E_L &= \mu - \delta \\ E_H &= \mu + 2\delta \end{aligned} \quad (4)$$

where μ and δ are the mean and standard deviation of the projection curve respectively.

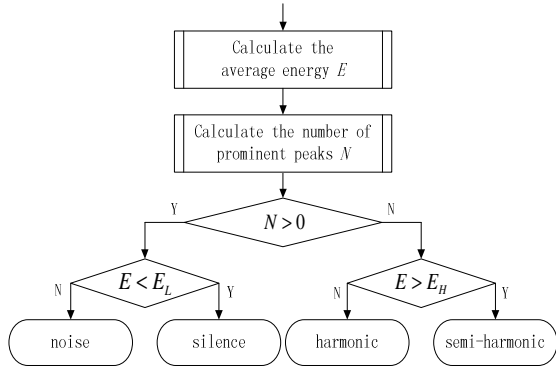


Figure 4. Flow diagram for harmonicity pattern assignment

2.2.2 Pitch Contour Pattern

Pitch contour pattern is designed to represent the shape of fundamental frequency along the time. An efficient pitch-tracking algorithm [5] is performed by using a combination of Fourier transforms. The pitch contour is then divided into M parts in temporal domain and each part is assigned with one semantic label to indicate its frequency contour shape. The available labels including *absence*, *increase*, *decrease* and *sustain*, which indicate no pitch, pitch rising, pitch falling and pitch unchanged, respectively. Thus the vector denoting the pitch contour pattern of an audio clip is:

$$\begin{aligned} Frq &= [f_1, f_2, \dots, f_M] \\ f_i &\in \{absence, increase, decrease, sustain\}, 1 \leq i \leq M \end{aligned} \quad (5)$$

Since the pitch contours of most sound effects vary tardily in comparison with speech, for samples in our database, setting M as 5 is enough to describe the pitch evolution along the time. Thus, the samples of “jet plane” illustrated in Figure 1(b) have the same frequency contour pattern looks like $[absence, sustain, decrease, sustain, absence]$.

3. SIMILARITY MEASURE

In our approach, the distance between two audio clips is calculated to measure their similarity. Since all the structure patterns are represented as vectors of semantic labels, to measure the similarity between two pattern vectors, the distance between two labels s_1 and s_2 is first defined as:

$$D_{label}(s_1, s_2) = \begin{cases} 0 & (s_1 = s_2) \\ 1 & (s_1 \neq s_2) \end{cases} \quad (6)$$

Then, for two symbol vectors V_1 and V_2 of certain structure pattern, the corresponding distance is defined as:

$$D_{vec}(V_1, V_2) = \frac{1}{L} \sum_{i=1}^L D_{label}(V_1(i), V_2(i)) \quad (7)$$

where L is the vector dimension, which is 1, 6 and 5 for the structure pattern of energy envelope, harmonicity and pitch contour in our current approach, respectively.

Integrating structure patterns with statistical feature, the distance between two audio segments S_1 and S_2 can be defined as:

$$\begin{aligned} D_{seg}(S_1, S_2) &= \alpha \cdot D_{stat}(S_1, S_2) + \beta \cdot D_{vec}(En_{S_1}, En_{S_2}) + \\ &\gamma \cdot (D_{vec}(Har_{S_1}, Har_{S_2}) + D_{vec}(Frq_{S_1}, Frq_{S_2})) \end{aligned} \quad (8)$$

where $D_{stat}(S_1, S_2)$ is the distance between the statistical features of S_1 and S_2 , which is calculate by a L_2 distance, as [2] did, and then normalized to $[0, 1]$. In (8), α , β and γ are weightings and set as 1.0, 0.5 and 0.25 respectively in our experiments, simply assuming that the statistical features and structure patterns, as well as the temporal and spectral structure patterns, have equal contribution to similarity measure.

The discussion above deals with the case where each sound is a single gestalt [2]. However, many sounds are much longer and consist of several gestalts. In order to give a more precise similarity measure, an audio clip is first divided into segments according to the normalized energy contour. For example, an audio effect of several chirms is divided into segments, and each segment contains one chirm.

Finally, the distance between audio clips C_1 and C_2 is defined as:

$$D_{clip}(C_1, C_2) = \frac{1}{2} (d(C_1, C_2) + d(C_2, C_1)) \quad (9)$$

and

$$d(C_m, C_n) = \frac{1}{p} \sum_{i=1}^p \min(D_{seg}(C_m(i), C_n(j)), 1 \leq j \leq q) \quad (10)$$

where $C_m(i)$ and $C_n(j)$ represent the i^{th} and j^{th} segment in clip C_m and C_n , and there are totally p segments in C_m and q segments in C_n respectively. The final distance in (9) is such defined, since the distances calculated by (10) are not symmetrical with each other, that is, $d(C_m, C_n) \neq d(C_n, C_m)$.

4. EXPERIMENTS

A content based audio retrieval system is built in our experiments to evaluate how the structure patterns can improve the similarity measurement. The baseline retrieval system is established based on the framework of Muscle Fish [2], only using the statistics (mean, variance or covariance) of the following features [2][3][4][6]: short-time energy, low short-time energy ratio, average zero-crossing rate, high zero-crossing rate ratio, sub-band energies, brightness, bandwidth, spectrum centroid, spectrum rolloff and 8-order *MFCC*. The temporal and spectral structure patterns are then integrated in the baseline system. Similarity measurement presented in the above section is used to calculate the distance between query and clips in database.

Our testing database consists of around 600 audio clips. These sounds vary in duration from less than one second to about 30 seconds; and include kinds of sounds, such as animals, machines, vehicles, human, weapons and so on. All the sample rate of sounds in our database is 22050 Hz.

Experiment results shows that either temporal or spectral structure patterns can improve recall ratios. Table 1 illustrates the detailed results of a query of a “jet plane” sound when using different combinations of statistical features and structure patterns in similarity measure. Symbols *S*, *H* and *E* represent statistical features, spectral structure patterns (harmonicity pattern and pitch contour pattern) and energy envelope pattern respectively. There are totally 9 clips of the same kind sound in database and each row in Table 1 lists their ranks in retrieval results. From the table, it can be seen that almost each rank is improved after using temporal and spectral structure pattern. For example, the rank of the first target clip is improved from 3 to 1; the second one is improved from 8 to 3, and the third one is improved from 11 to 5, after using structure patterns.

Table 1. Retrieval for a sound effect “jet”

	1	2	3	4	5	6	7	8	9
<i>S</i>	3	8	11	12	36	100	102	140	180
<i>S+H</i>	1	5	10	12	15	70	71	99	134
<i>S+E</i>	1	4	6	9	11	39	54	98	201
<i>S+H+E</i>	1	3	5	8	10	22	37	80	169

In order to show more general performance, more sound effects clips (totally about 100), including sheep, gun, racing car, applause and so on, are used as queries in experiments. Figure 5 illustrates the comparison between the retrieval results with and without structure patterns, where average recall and precision ratio are used to evaluate the performance. From the Figure 5, both recall and precision are increased much after integrating structural patterns. For example, in the results of top 30, about 70% targets are retrieved with structure patterns, while only 40% is obtained using common statistical features; the precision is also increased by 60% after using structural patterns. It clearly indicates that the proposed structure patterns give a distinct improvement in the similarity measurement of sound effects.

5. CONCLUSION

In this paper, we proposed some structure patterns on energy contour, harmonicity and pitch contour, and integrate them into

audio similarity measure. Experiments on a content-based audio retrieval system proved that the feature structure patterns are effective supplements to similarity measure. More future works may include: (i) design more effective and representative structure patterns, which are important and relevant to auditory perception; (ii) look for a better way to integrate the structure patterns into similarity measure, with or without traditional statistical features.

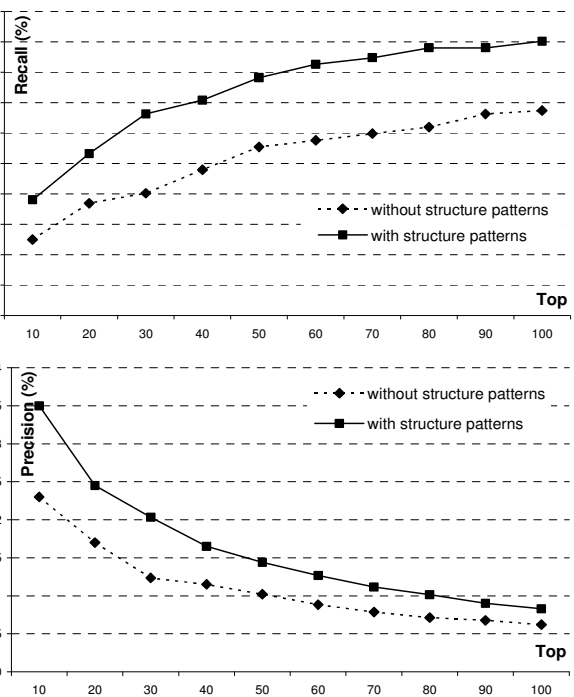


Figure 5. Comparisons of recall and precision ratios between the retrieval results with and without structural patterns

6. REFERENCES

- [1] B. Gygi, “Factors in the Identification of Environmental Sounds”, *Ph.D. Thesis*, 2001.
- [2] E. Wold, T. Blum, D. Keislar, and J. Wheaton, “Content-Based Classification, Search, and Retrieval of Audio”, *IEEE Multimedia*, Vol.3, No.3, pp. 27-36, 1996.
- [3] ISO-IEC/JTC1 SC29 WG11 Moving Pictures Expert Group. Information Technology – Multimedia Content Description Interface – Part 4: Audio. Committee Draft, 15938-4, ISO/IEC, 2000.
- [4] L. Lu, H.-J. Zhang, H. Jiang, “Content Analysis for Audio Classification and Segmentation”, *IEEE Trans. on Speech and Audio Processing*, pp.504-516, Vol.10, No.7, Oct. 2002.
- [5] S. Marchand. “An Efficient Pitch-Tracking Algorithm: Using A Combination of Fourier Transforms”, *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-01)*, Limerick, Ireland, December 6-8,2001.
- [6] Z. Liu, Y. Wang and T. Chen, “Audio Feature Extraction and Analysis for Scene Segmentation and Classification”, *J. VLSI Signal Processing Systems*, pp. 61-79, June, 1998.