Popular Music Retrieval by Detecting Mood

Yazhong Feng Yueting Zhuang Yunhe Pan

College of Computer Science, Zhejiang University, Hangzhou, China

fengyz_zju@263.net yzhuang@cs.zju.edu.cn panyh@sun.zju.edu.cn

Categories and Subject Descriptors

H.3.3 [**Information Storage and retrieval**]: Information Search and Retrieval - *retrieval models*, *search process*, *selection process*.

General Terms

Algorithms, Performance, Design, Experimentation

Keywords

Music information retrieval, Mood detection, Tempo, Articulation, Visualization

1. INTRODUCTION

In the community of music information retrieval, researchers developed methods to retrieval music with a particular melody, they also developed methods to retrieval music by similarity. We aim to retrieval music by mood, which is sometimes the exclusive manner people select music to enjoy, for example, when someone is sad for some reason, she or he wants to listen to a piece of music that can cheer her or him up, at this moment she or he will search music segment by mood no matter what the melody sounds and which the piece of music is similar to. In the essence, the difficulty for music retrieval by mood stems from the gap between the rich meanings that users want when they query and the shallowness of content descriptor we can actually compute. Recently, a new approach, Computational Media Aesthetics (CMA) is presented aiming to fill this gap, the core trait of it is that data is interpreted with its maker's eye [1]. Composers choreograph the expectation to arise emotion, and performers convert the musical intention into music language to arise emotion, which inspires us to analysis music mood on the viewpoint of how music is made, i.e. the viewpoint of CMA avoiding coping with ambiguity of emotion audience arise when confronted with music. Juslin found that two music dimensions could explain the transfer of emotional content from performer to audience: tempo and articulation, tempos were either fast or slow while articulations were either staccato or legato [2], Table 1 depicts the relationship of music mood with tempo and articulation.

In our scheme, music database is indexed on four labels of music mood, concretely "happiness", "sadness", "anger" and "fear"; three features, relative tempo, the mean and standard deviation of average silence ratio, are used to classify mood, the classifier is a BP neural network. When user's query is accepted, the system displays the corresponding region of music database by visualizing the feature space, users then browse to select music piece. The hypothesis we make in our music mood detection scheme is that tempo and articulation are invariable in the concerned music segment.

Copyright is held by the author/owner(s).

SIGIR'03, July 28-August 1, 2003, Toronto, Canada.

ACM 1-58113-646-3/03/0007.

Table 1. The relationship of music mood with tempo and articulation

MOOD	TEMPO		ARTICULATION	
	fast	**slow**	**staccato**	**legato**
happiness	yes	no	yes	no
sadness	no	yes	no	yes
anger	yes	no	no	yes
fear	no	yes	yes	no

2. TEMPO DETECTION

We define a real-valued interval $[s, f]$ representing the tempo boundary of most music pieces, s is the lowest tempo and f is the fast tempo, s and f are statistically derived in Section 4, a feature called relative tempo is defined on this interval:

$$rTEP = \frac{TEP}{1/s + 1/f} \tag{1}$$

We adopt the approach in [3] to detect music tempo, it does not use any priori knowledge such as style, time signature or approximate tempo about music, nor does it model the cognition mechanisms involving human rhythm perception. It is robust in various music styles and computation coat is moderate. Deducted tempo TEP is expressed by *inter-beat interval* (seconds) in this approach.

3. COMPUTATIONAL ARTICULATION MODEL

A time-domain feature called Average Silence Ratio (ASR) is employed to model articulation, whose definition is:

$$ASR = \frac{1}{2N}\sum_{n=0}^{N-1}\left(1 - \mathrm{sgn}\left(STE(n) - \rho \times avgSTE\right)\right) \tag{2}$$

$$avgSTE = \frac{1}{N}\sum_{n=0}^{N-1} STE(n) \tag{3}$$

$$STE(n) = \sum_{k=n}^{n+m-1} a^2(k) \tag{4}$$

Where N is the total frame number in one second analyzing time window, $STE(n)$ is the short-time energy of frame n , ρ is an experimental parameter so that $\rho \times avgSTE$ acts as the reference to determine if the short-time energy of frame n is low enough to reveal a silence, m is the frame size, $a(k)$ is the signal amplitude at time point k in frame n .

This feature indicates that a frame is regarded as silence if its energy is lower than ρ percent of the average energy in the one-second time window, and ASR is a counter of silence frame in this time window. The lower ASR means fewer silence frames present in music piece, or legato in articulation, and the higher ASR means more silence frames present in music piece, or staccato in articulation. Figure 1 shows that ASR is a good discriminator of different moods in the facet of music articulation. In our system, with regard to a piece of music, we use the mean and standard deviation of its ASR sequence as two simple music articulation features, which are denoted as $mASR$ and $vASR$, we believe that them give us more information about the articulation character of music piece then ASR does because they are the global statistics of articulation.

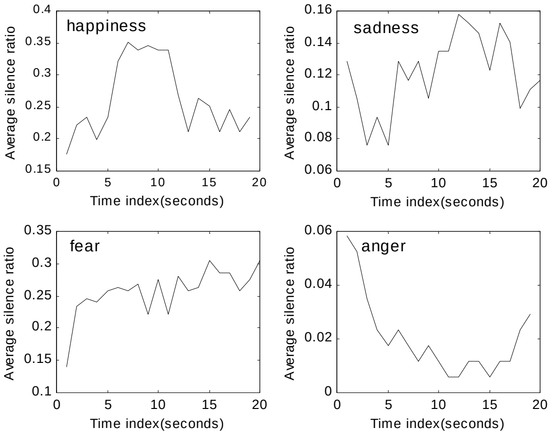

Figure 1. The average silence ratio of four segments of 20s music in different moods: happiness, fear, sadness and anger. The *ASR*s of "happiness" and "fear" segment is relatively high with the value higher than 0.15, the *ASR*s of "sadness" and "anger" segment is relatively low with the value less than 0.15. $\rho = 0.25$.

4. EXPERIMETAL RESULTS

We collect 223 pieces of modern popular music containing multiple instruments and vocal singing form Internet and personal CD repository, 200 pieces are used as training data, 23 pieces as testing data. Each music piece is converted to 22050Hz/mono/16bit raw audio signal. The structure of neural network classifier in our experiment is three input nodes, ten hidden nodes and four output nodes, to train the neural network, [$rTEP$, $mASR$, $vASR$] of music piece is used as the input of neural network classifier, the target output are scores of music mood.

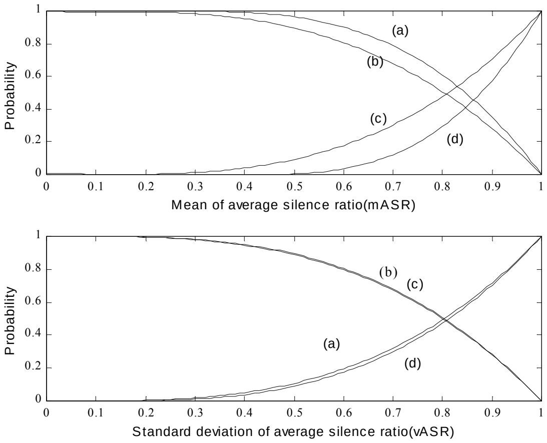

Figure 2. Probability distributions of $mASR$ and $vASR$, the curves labeled by (a) (b) (c) (d) represent the probability distributions of "sadness", "anger", "happiness" and "fear" music piece, respectively.

We also plot the probability distributions of $mASR$ and $vASR$ for the experimental corpus (illustrated in Figure 2), it is obviously that music in different moods in differentiable by the mean of average silence ratio. We experiment on different ρ in ASR , for example, $\rho = 0.25, 0.5$, and find almost no difference in ASR of staccato and legato articulation, concretely, $ASR < 0.15$ for legato and $ASR > 0.15$ for staccato. We calculate the *precision* and *recall* to evaluate the retrieval performance (Table 2). The total *precision* is 67% and the total *recall* is 66%.

Table 2. Retrieval performance

Percent(%)	HAPPINESS	SADNESS	ANGER	FEAR
PRECISION	86	75	83	25
RECALL	57	38	100	67

Experimental results show that music mood is computable. Our test corpus is by no means large enough; if possible, we will test our scheme on a commonly recognized evaluation corpus in the future. For our computational articulation model, we only use time domain energy of music signal to calculate articulation feature, we will try other features in future.

5. REFERENCES

[1] Dorai, C. and Venkatesh, S., Computational Media Aesthetics: Finding meaning beautiful. IEEE Multimedia, 8(4), October-December, 2001, 10-12.

[2] Juslin, P.N., Cue utilization in communication of emotion in music performance: Relating performance to perception. J. Experimental Psychology, 26, 2000, 1797-1813.

[3] Dixon, S. A lightweight multi-agent musical beat tracking system. In proceedings of the Pacific Rim International Conference on Artificial Intelligence, Melbourne, Australia, 2000, 778-788.