工作日誌

- **找尋研究方向**

  因本身對音樂這領域有興趣，目前構想的研究主題想主要和音樂有關，又剛好 Machine Learning 應用到音樂領域也是非常熱門，所以想做這方面的研究。

- **初步研究題目**

  (1) 音樂風格轉換 Music Style Transfer

  (2) 音樂情緒分類 Music Emotion Classification

  - 目前做情緒分類大部分都是以四個象限四個情緒為主，是否可以進一步分析更多種情緒？
  - 音樂轉圖像？(透過情緒對應)

文獻探討

[1] "Style" Transfer for Musical Audio Using Multiple Time-Frequency Representations, ICLR, 2018

[2] E. Grinstein, N. Q. K. Duong, A. Ozerov, and P. Perez, “Audio style transfer,”, arXiv: 1710.11385, 2017.

[3] M. B. Mokhsin, N. B. Rosli, W. A. W. Adnan, and N. A. Manaf, “Automatic Music Emotion Classification Using Artificial Neural Network Based on Vocal and Instrumental Sound Timbres,” New Trends in Software Methodologies, Tools, and Techniques, 2014, pp. 3–14

[4] C. Lin, M. Liu, W. Hsiung and J. Jhang, "Music emotion recognition based on two-level support vector classification," 2016 International Conference on Machine Learning and Cybernetics (ICMLC), Jeju, 2016, pp. 375-389.doi: 10.1109/ICMLC.2016.7860930

[5] Chia-Hung Yeh & Wen-Yu Tseng & Chia-Yen Chen & Yu-Dun Lin & Yi-Ren Tsai & Hsuan-I Bi & Yu-Ching Lin & Ho-Yi Lin, Popular music representation: chorus detection & emotion recognition, Springer Science + Business Media, Multimedia Tools and Applications, 2014, Volume 73, Issue 3, pp. 2103–2128

論文名稱	"Style" Transfer for Musical Audio Using Multiple Time-Frequency Representations
摘要	神經風格轉移（Gatys et al，2016）已成為使用卷積神經網絡生成不同藝術風格圖像的流行技術。最近在圖像樣式轉換方面的成功提出了一個問題，即是否可以利用類似的方法來改變音樂音頻的「風格」。在這項工作中，我們嘗試在時域中進行長時間高質量的音頻轉換和紋理合成，抓取與音樂風格相關的旋律，節奏和音色的元素，使用具有不同長度和音樂鍵作為例子。我們展示了使用隨機初始化卷積神經網絡將音樂風格的這些方面從一個片段轉移到另一個片段的能力，使用 3 種不同的音頻表示：短時傅立葉變換（STFT）的對數幅度，Mel 頻譜圖和 CQT 轉換頻譜圖，使用這些表示作為產生和修改音樂音頻內容的重要特徵的方式。我們透過仔細設計與音樂音訊的本質互補的神經網路結構，來展示每個表示法的缺點和優勢。最後，我們展示了最引人注目的「風格」轉換例子，利用這些表示的集合來說明捕捉音訊信號的不同期望特徵。
優點	1. 比較 3 種音頻的表示方式，發現使用 Mel 頻譜圖和 CQT 轉換頻譜圖可改善先前的方法，能抓取到有意義的樣式資訊。 2. 成功嘗試完全在時域中執行風格轉移。
缺點	1. 本篇在 style loss 和 content loss 沒有將計算結果列出來。
自評	1. 認為自己在專有名詞上還需多加了解，以方便了解流程圖的內容。

論文名稱	Audio Style Transfer
摘要	圖像之間的「風格轉移」最近成為一個非常活躍的研究課題，由卷積神經網絡（CNN）的力量推動，並且已成為社交媒體中非常流行的技術。本文研究了音頻領域中的類似問題：如何將參考音頻信號的風格轉換為目標音頻內容？我們提出了一個靈活的任務框架，它使用聲音紋理模型來提取表徵參考音頻風格的統計數據，然後是基於優化的音頻紋理合成來修改目標內容。與基於主流優化的視覺傳遞方法相比，所提出的過程由目標內容而不是隨機噪聲初始化，優化的損失僅僅是紋理而不是結構。這些差異被證明是我們實驗中音頻風格轉移的關鍵。為了提取感興趣的特徵，我們研究了不同的體系結構，無論是在其他任務上預先訓練，如在圖像樣式轉移中完成，還是基於人類聽覺系統設計。對不同類型的音頻信號的實驗結果證實了所提出的方法的潛力。
優點	1. 使用 4 種不同的模型(VGG-19、SoundNet、Wide-Shallow-Random network、McDermott)來比較，分別探討何者的成效較好。
缺點	1. 本文未計算出參考樣式音頻和輸出音頻之間的 Style loss，僅提供聲音檔和頻譜圖來判斷差異。 2. 實驗數據的部分較薄弱，無法明確知道該方法是否能真正能達到預期的效果。 3. 作為實驗的音頻數量不多，應該透過更多組音頻來比較測試。
自評	1. 如果未來要繼續做風格轉換這區塊，可以參考本篇的架構，針對風格特徵提取和風格轉換這兩部分。

論文名稱	Automatic music emotion classification using artificial neural network based on vocal and instrumental sound timbres
摘要	檢測歌曲中的情感特徵仍然是各種研究領域的挑戰，尤其是在音樂情感分類 (MEC) 中。為了將所選擇的歌曲分類為具有特定的情緒或情緒，機器學習的算法必須足夠智能以學習數據特徵以相應地將特徵與精確情緒相匹配。到目前為止，只有少數關於 MEC 的研究利用了歌曲的聲樂部分和歌曲的樂器部分結合的音頻音色特徵。音色特徵是音樂的特性或聲音，它區分人類聲音和樂器中不同類型的聲音產生，如弦樂器，管樂器和打擊樂器。 MEC 中的大多數現有作品都是通過查看音頻，歌詞，社交標籤或兩個或更多類的組合來完成的。問題是從聲樂和樂器聲音特徵中利用音色特徵是否有助於在 MEC 中產生積極效果？因此，本研究目的在於利用人工神經網絡通過從聲樂和樂器聲音片段中提取音頻音色特徵來檢測馬來流行音樂中的情感特徵。該研究的結果將基於對聲樂和樂器音色特徵的操縱來共同改進 MEC，並且有助於音樂信息檢索，情感計算和心理學的文獻。
優點	1. 選擇 ANN 分類器，有別於以往使用 SVM 分類器。 2. 以往的有歌詞的歌曲大多是利用歌詞來判斷情緒，而本篇使用唱歌者的聲音和樂器伴奏的音色來作為分類依據。
缺點	1. 在音頻的部分只能使用 WAV 格式的音檔。 2. 分類器所選擇的分類依據只有音色。 3. 本篇目前只針對馬來西亞的音樂做為研究。
自評	1. 未來可加入更多的音樂元素作為分類的依據，畢竟影響音樂情緒的因素有很多種。 2. 本篇的準確率有機會再提升。 3. 增加訓練樣本的多樣性。

論文名稱	Music emotion recognition based on two-level support vector classification
摘要	音樂情感識別 (MER) 可以檢測音樂片段中人們固有的情感表達。 MER 有助於多媒體理解，音樂檢索和其他與音樂相關的應用。隨著近年來在線音樂內容的數量迅速擴大，最近出現了對情感檢索的需求。以計算方式確定音樂的情感內容是一項跨學科研究，不僅涉及信號處理和機器學習，還涉及對聽覺，心理學，認知科學和音樂學的理解。評估自動音樂情感檢測的一個挑戰是，目前還沒有完善的音樂情感描述情感模型。此外，由於基於聲學特徵的音樂情感識別器的透射率低，因此難以解釋由該機制產生的數據。在這項研究中，提出了一個基於領域知識預先描述的音樂流派和音樂特徵的兩級分類系統。該框架具有利用最合適的聲學信息的優點。實驗將通過衡量不同情緒表達和各種音樂線索之間的相關性來進行。為了驗證整體系統的性能，還將基於音樂特徵與地面真實情感之間的一致性來評估提議模型。
優點	1. 在特徵提取方面，使用到音樂的元素(節奏、音色、音調、動態)，透過這些元素能表達一首歌的情感。 2. 使用到特徵加權的工具(RReliefF)。 3. 採用雙層的 SVM。
缺點	1. 實驗的音頻未提供。 2. 情緒的類別在本篇只分成四種，或許可以增加情緒的多元性。
自評	1. 本文使用到特徵加權的工具(RReliefF)，畢竟音樂元素眾多，可以挑選一兩個當作主要特徵，其餘輔佐用，這樣就能更確定某一特徵的成效性或影響性。

論文名稱	Popular music representation：chorus detection & emotion recognition
摘要	本文提出了一種基於歌曲情感的流行音樂表現策略。首先，通過所提出的合唱檢測算法將一段流行音樂分解為合唱和詩歌片段。從結構化片段中提取三個描述特徵：強度，頻帶和節奏規律性，用於情緒檢測。採用分級 Adaboost 分類器來識別一首流行音樂的情感。音樂的一般情緒根據 Thayer 的模型分為四種情緒：快樂，憤怒，沮喪和放鬆。在 350 個流行音樂數據庫上進行的實驗表明，我們提出的合唱檢測的平均召回率和精確度分別約為 95%和 84%;情緒檢測的平均準確率為 92%。對具有不同歌詞和語言的封面版本的歌曲進行附加測試，結果精確率為 90%。提議方法已經由專業在線音樂公司 KKBOX Inc.測試和驗證，並且顯示出有效且有效地識別各種流行音樂的情緒的有希望的表現。
優點	1. 本篇使用自己的 database 和 MIREX 2009 的 database 來做比較，以證明自己的 database 比較好。 2. 本篇 3 個特徵提取的部份都能得到很好的結果。 3. 階層式分類器在此篇能有很精準的結果。
缺點	1. 有些地方矛盾(前面 XY 軸屬性和後面說特徵值都使用 arousal 有關) 2. Database 的內容應該要針對音樂類型有所挑選(像是舞曲部份，節奏過於相似，無法突顯特別結構) 3. 文中有些公式的參數錯誤。
自評	1. 使用音樂的元素作為分類依據，有別於以往單純只使用以文字為基礎來分類。 2. 我認為在副歌偵測和情緒偵測的準確率有機會再提升。 3. 加大資料集的規模、修改資料集內容。 4. 自己定義的資料集沒有公開內容。 5. 音頻轉成圖片的應用(音頻情緒對應到該情緒的圖片)

Popular music representation: chorus detection & emotion recognition

Source : Multimedia Tools and Applications December 2014, Volume 73, Issue 3, pp. 2103–2128

Authors : Yeh, CH (Yeh, Chia-Hung) ; Tseng, WY (Tseng, Wen-Yu) ; Chen, CY (Chen, Chia-Yen) ; Lin, YD (Lin, Yu-Dun) ; Tsai, YR (Tsai, Yi-Ren) ; Bi, HI (Bi, Hsuan-I); Lin, YC (Lin, Yu-Ching); Lin, HY (Lin, Ho-Yi)

Impact Factor : 1.541

Speaker : Ching-Yi, Chiu

Date : 2018/10/1

Outline

- Abstract
- Introduction
- Related Works
- Proposed Method
- Experimental Results
- Conclusion
- Comment

Abstract

- Two tasks

Tasks	Content
chorus detection	• decomposed into chorus and verse segments
emotion detection	• Three descriptive features: Intensity , Rhythm regularity , Frequency band • hierarchical Adaboost classifier : recognize the emotion • Thayer's model

- Using 350-popular-music database to experiment, both of them have good average recall and precision.

- It have been tested and proven by the professional online music company, KKBOX Inc.

Introduction

- Popular music differs from others music, it consists of a variety of genres, but they often have similar structures.

- Popular songs consist of five basic parts:

- Verse → repeats two or more times within a song with different lyrics
- Chorus → repeated several times with a song with same lyrics song

Introduction

- Musical elements affect emotions (timbre, rhythm, pitch, harmony…)

- faster tempo and slightly higher pitch → happiness or cheerfulness

  slower tempo → solemnity and gravity

- This paper consists of two phases :

  A. analysis of popular music structure (color representation method)

  B. the emotion model of music (Thayer‘s 2-D emotion model + Adaboost classifier)

Related Works

A. Music retrieval systems

Text-based
- artist
- title
- lyrics

Users may lack specific music information.
It is time-consuming and tedious.

Content-based
- beat : utilize beat features
- note : detect pitch contour
- melody : retrieve a song through the humming, singing
  more flexible way to find music
- timbre : analyze tonal characteristics(Mel-frequency Cepstral Coefficients, MFCCs)

Related Works

B. Song Structure Analyze

Using similarity matrix to analyze the structure of a song.
Bartsch et al.[1] , Cooper et al. [2] , Foote et al. [3] , Yu et al. [4]

novel approach

chorus detection via color representation.

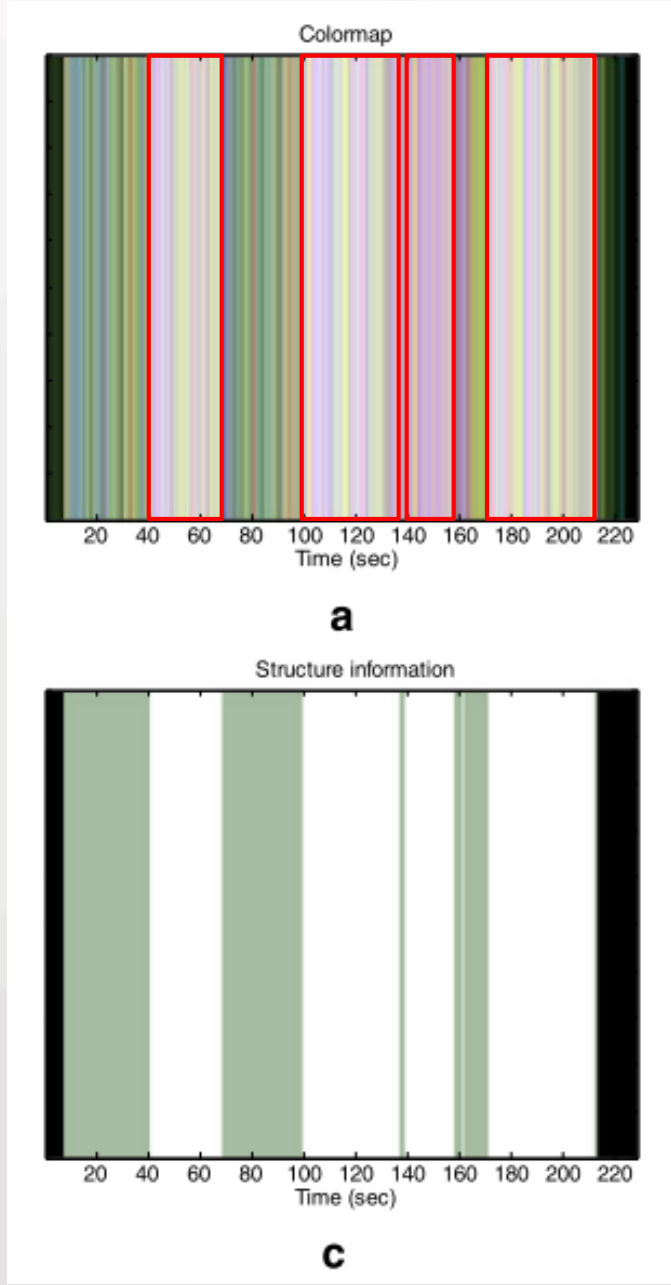


Fig. 1 Colormap & structure information

Related Works

C. Thayer's model

categorizes emotions and defines emotions using a two-dimensional model

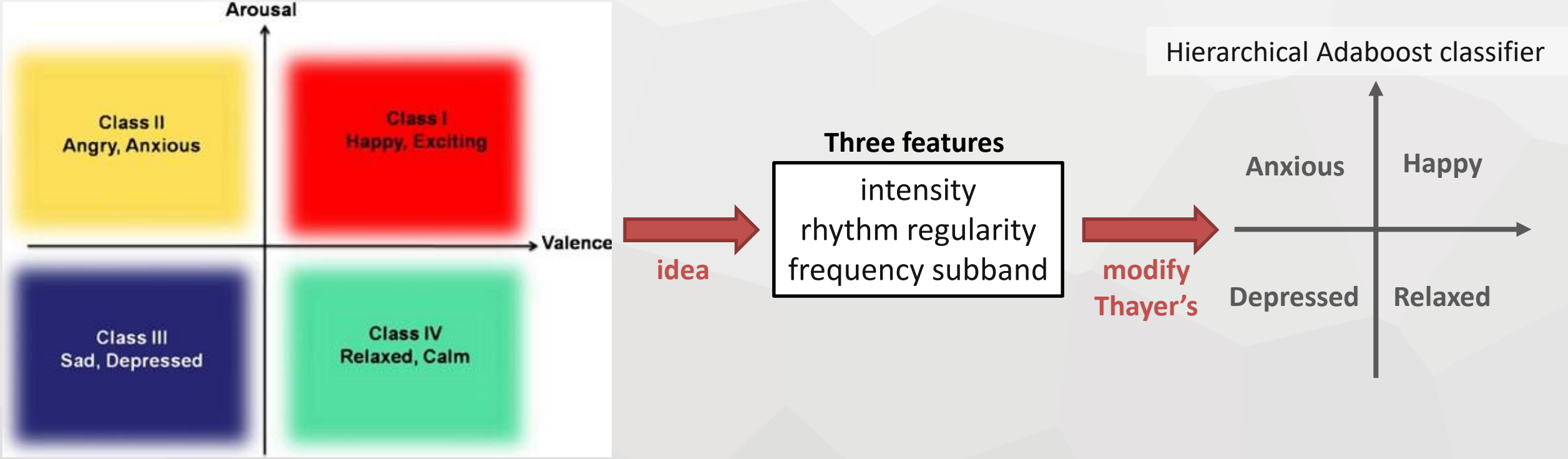


Fig. 2 Illustration of Thayer's emotion model

Proposed Method

A. Chorus Detection

Calculating three features

Feature vectors to the R, G, B color space
Cluster the region with similar color distribution

Extract MFCCs from the colormap

Classify verse or chorus

Exclude unreliable regions

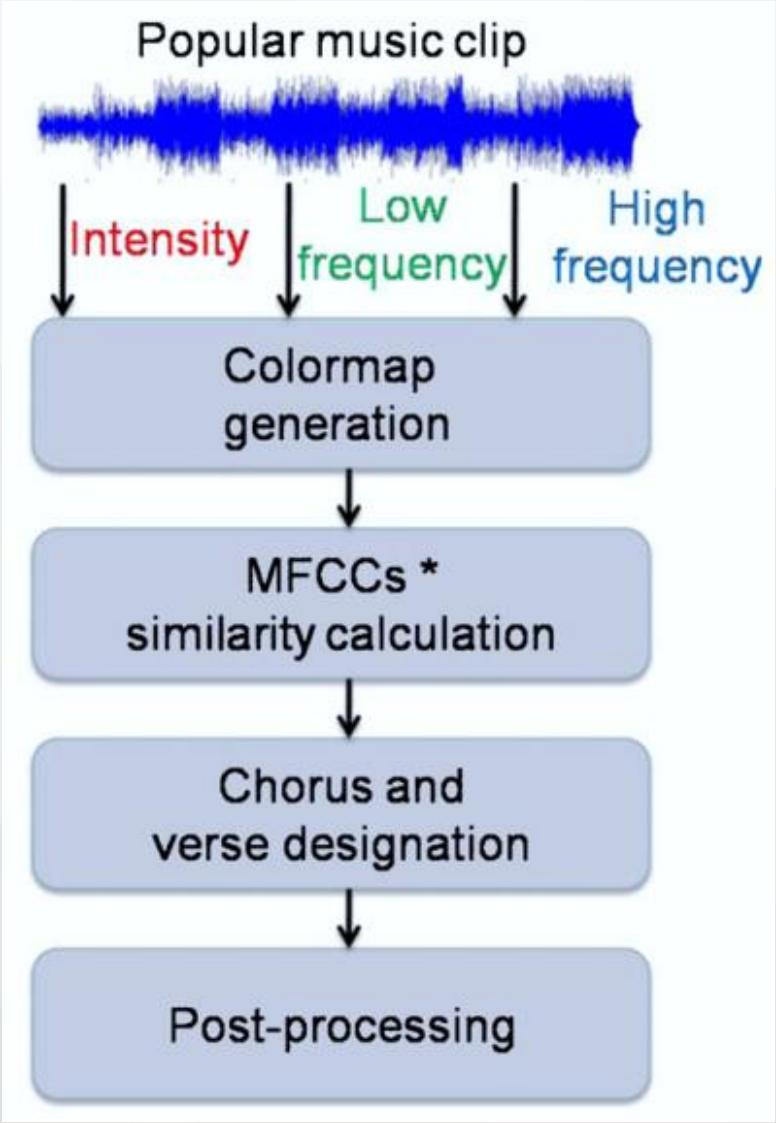


Fig. 3 Flowchart of chorus detection

Proposed Method

a) Colormap generation (FFT)

Fast Fourier transform (FFT)

$$X_k = \sum_{n=0}^{N-1} X_n \mathrm{e}^{-2\pi k \frac{n}{N}}$$

F_1 : The summation of the energy of a song

$$F_1(k) = \sum_{h=1}^{44100} |X_k(h)|$$

F_2 : Sum up FFT coefficients of low frequency bands

$$F_2(k) = \sum_{h=1}^{2048} |X_k(h)| \times w(h)$$

F_3 : Sum up FFT coefficients of high frequency bands

$$F_3(k) = \sum_{h=2049}^{22050} |X_k(h)| \times w(h)$$

- X_k is the k^{th} frame of audio data.
- $w(h)$ is a window function at h Hz.

Proposed Method

b) Colormap generation (RGB)

$$R(k) = \frac{F_1(k) - \min\left(F_1(K)\right)}{\max\left(F_1(k)\right) - \min\left(F_1(k)\right)}$$

$$G(k) = \frac{F_2(k) - \min\left(F_2(K)\right)}{\max\left(F_2(k)\right) - \min\left(F_2(k)\right)}$$

$$B(k) = \frac{F_3(k) - \min\left(F_3(K)\right)}{\max\left(F_3(k)\right) - \min\left(F_3(k)\right)}$$

Adaptive clustering method (RPCL) : find similar segments.

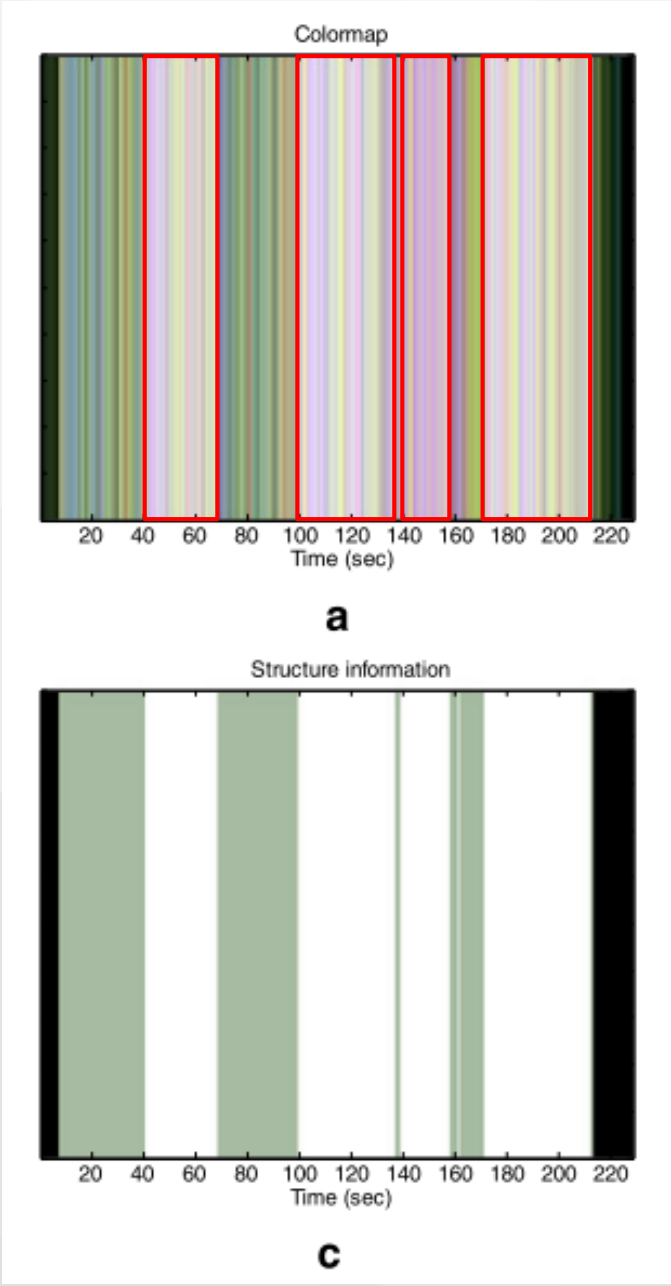


Fig. 4 Colormap & structure information

Proposed Method

c) Chorus and verse designation (MFCCs + similarity)

$$C_s(k) = \frac{1}{N}\sum_{l=1}^{N} Y(l)\cos\left(l\frac{\pi}{N}(s-0.5)\right)$$

- $C_s(k)$ is the s^{th} coefficient of the k^{th} frame.
- $Y(l)$ is a l^{th} filter bank.

The similarity between cluster i and j is calculated using the similarity matrix.

$$S_{(i,j)} = \frac{v_i \cdot v_j}{|v_i||v_j|}$$

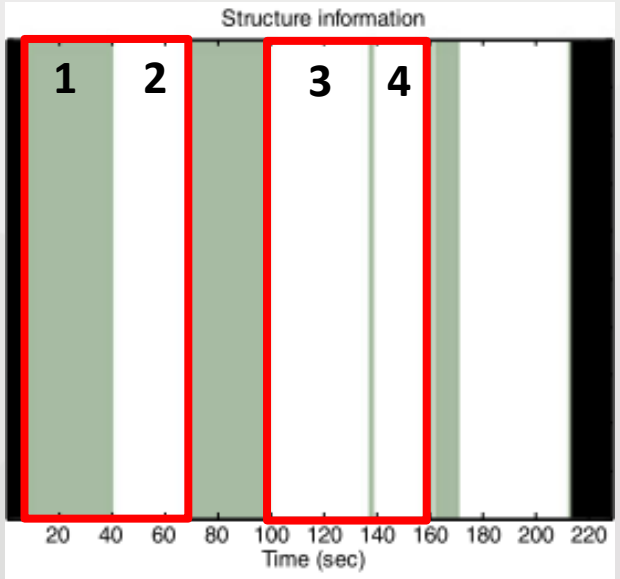


Fig. 5 Structure Information

Proposed Method

d) Chorus and verse designation (time-constraint)

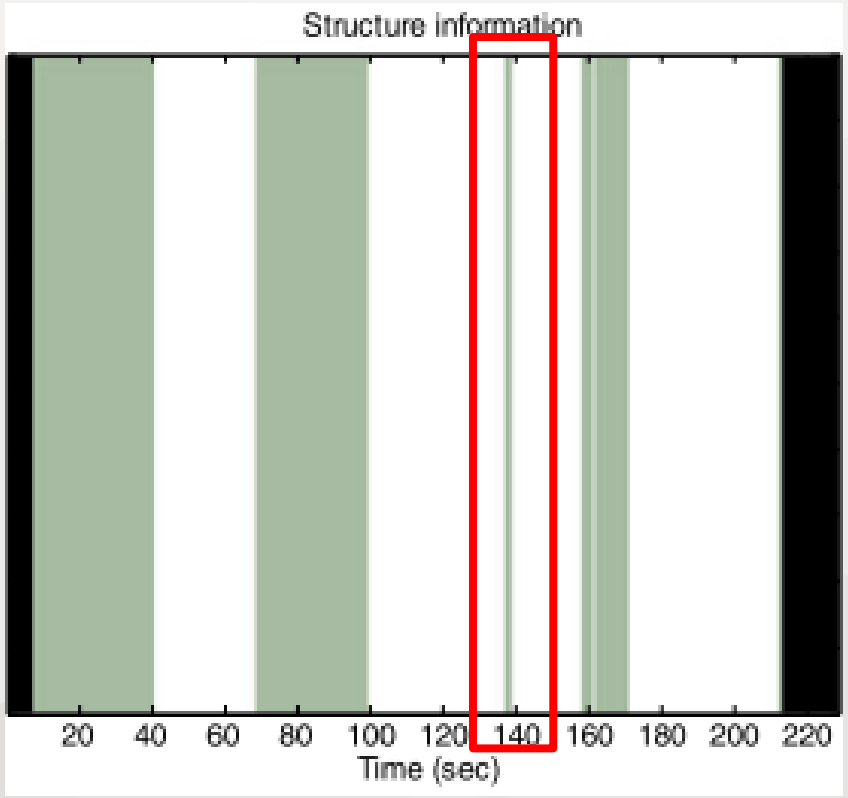


Fig. 6 Colormap & structure information

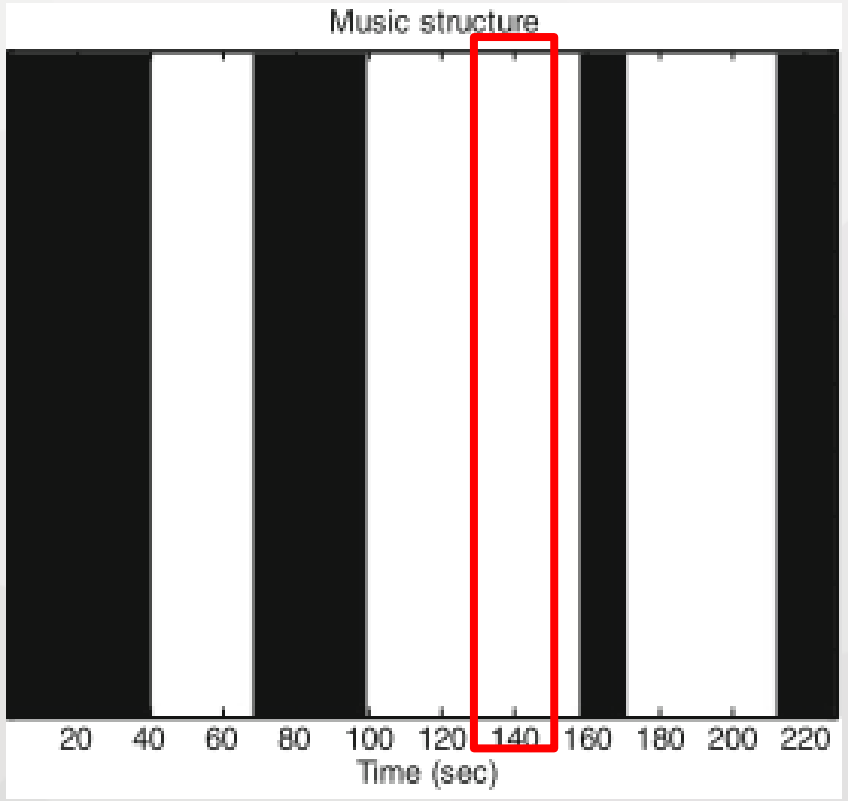
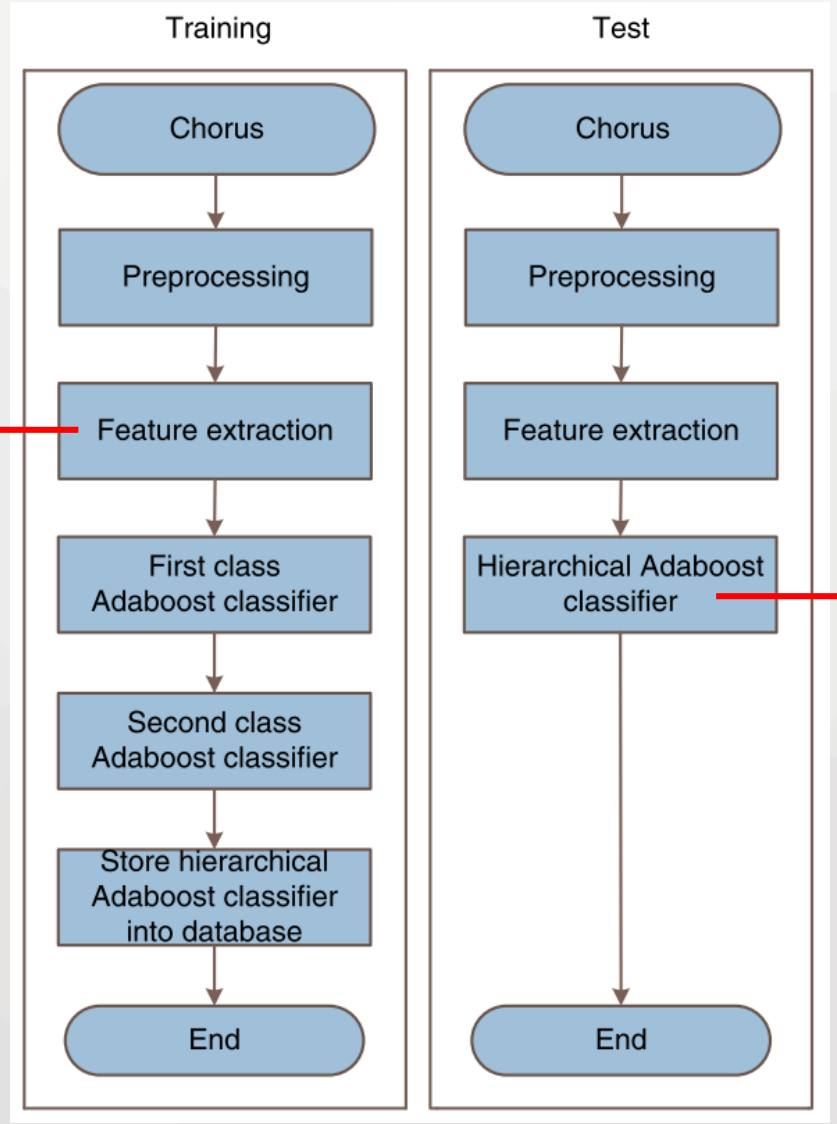


Fig. 7 Results of chorus detection

Proposed Method

B. Emotion Recognition

Three features
1. Intensity
2. Rhythm regularity
3. Frequency subband



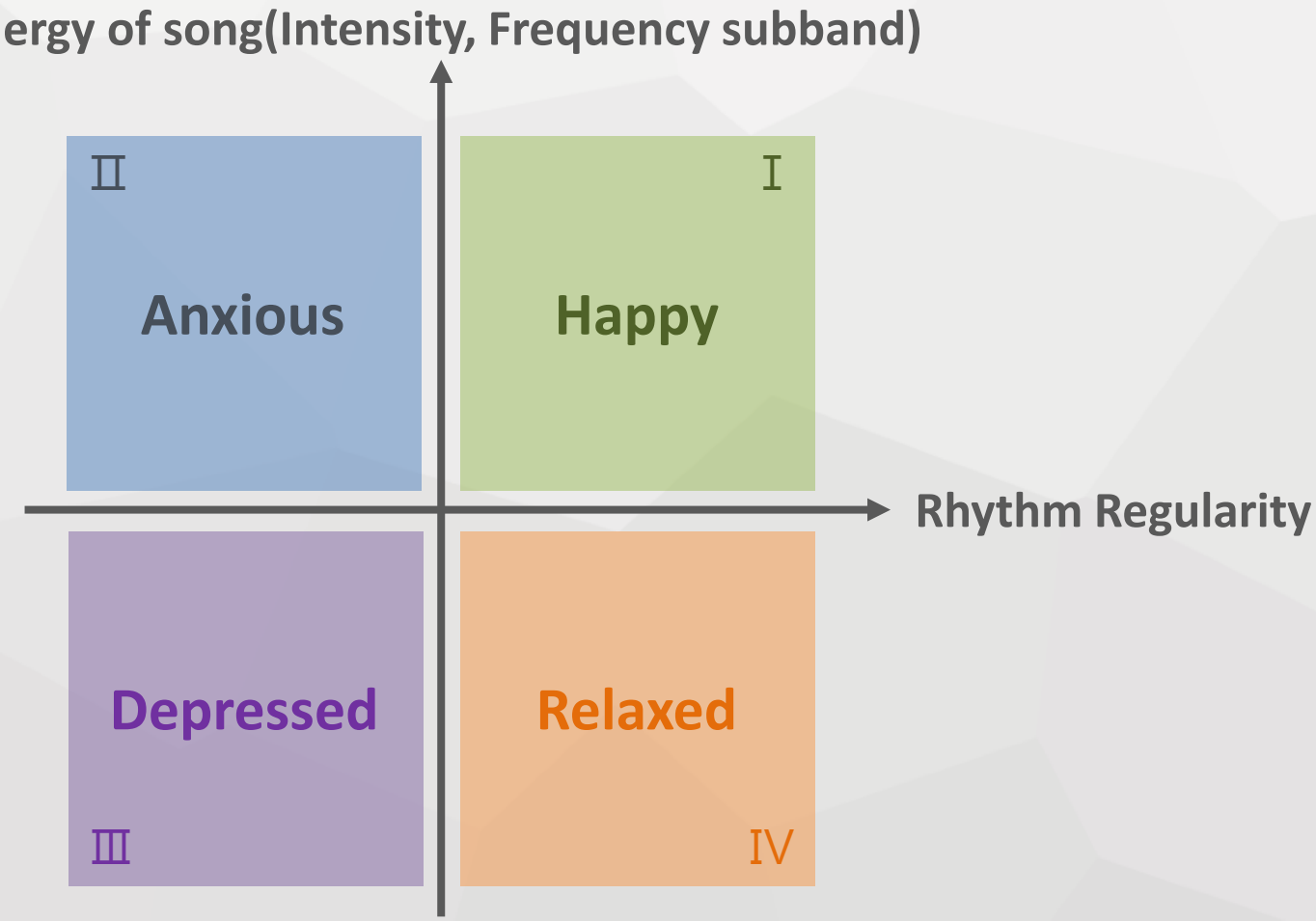
detect the emotion of a song

Fig. 8 Flowchart of emotion detection

Proposed Method

a) Emotion model establishment

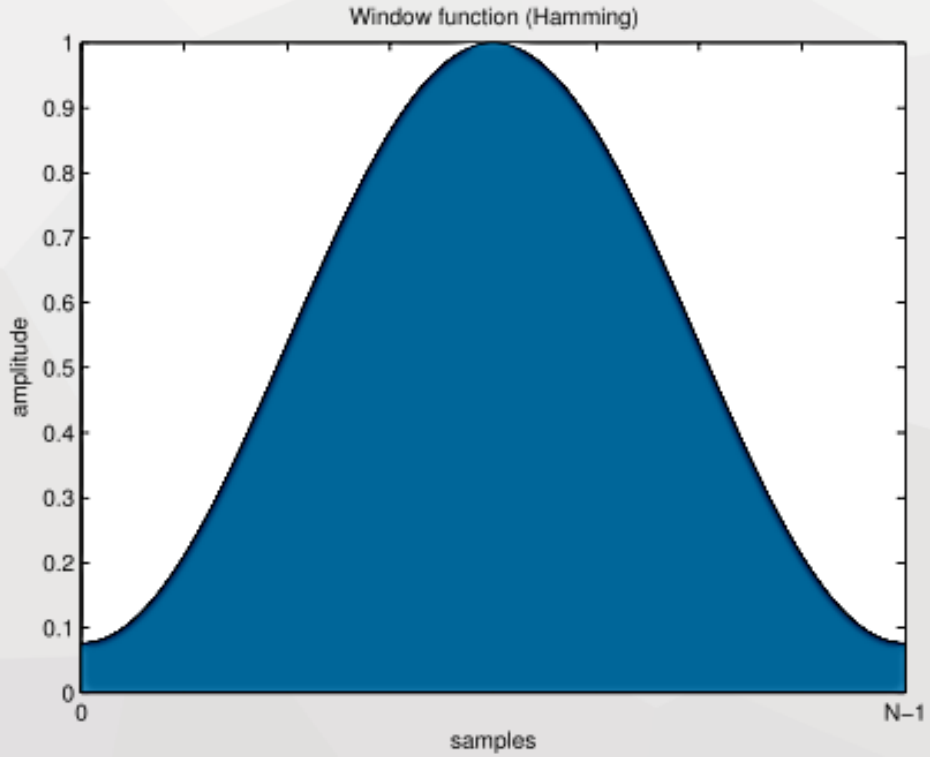
Thayer's 2-D emotion model is used in our emotion detection scheme.



Proposed Method

b) Feature extraction

Hamming window : minimize signal discontinuities



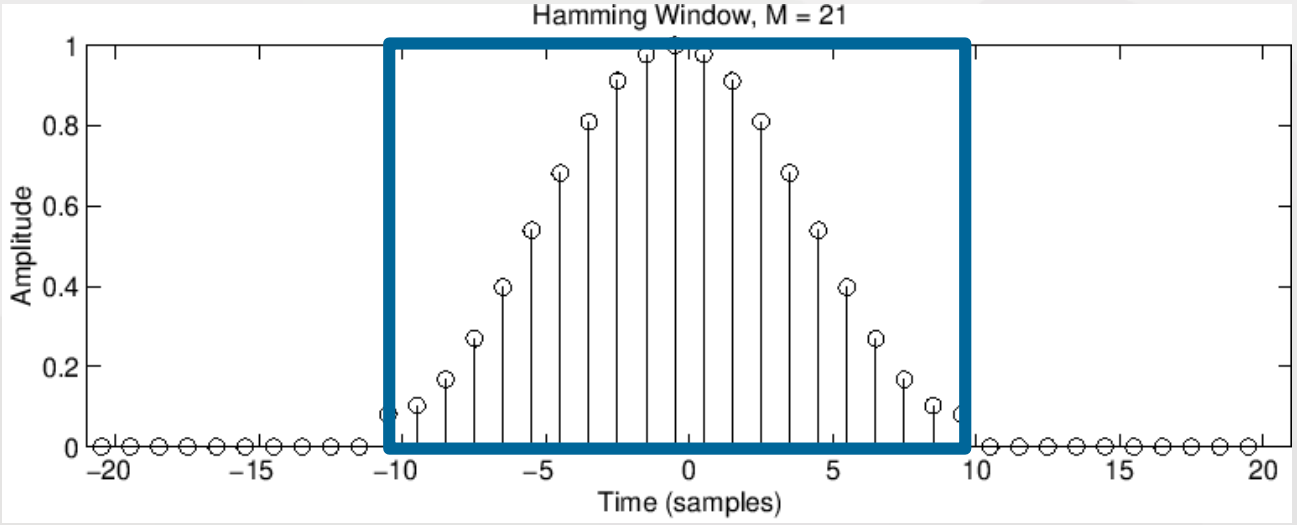


Fig. 9 Hamming windows

Proposed Method

b) Feature extraction

① **Intensity** : The energy summation of the frequency domain is represented below

$$E(k) = \sum_{m=1}^{N} |f_k(m)|^2$$

- $E(k)$ is the intensity of the k^{th} frame.
- $f_k(m)$ is a m^{th} Fourier coefficient of the k^{th} frame.
- N represents the frame size.

② **Frequency Subband**

Subband #	1	2	3	4
Frequency range (Hz)	1~128	128~256	256~512	512~1 k
Subband #	5	6	7	8
Frequency range (Hz)	1 k~2 k	2 k~4 k	4 k~8 k	8 k~22 k

→ low frequency bands

→ high frequency bands

Table. 1 Frequency range of 8 sub-bands

Proposed Method

b) Feature extraction

③ **Rhythm Regularity** : D_k stands for the change of amplitude in the frequency of adjacent frames

$$D_k = \begin{cases} |E(n,k) - E(n-1,k)|, & E(n,k) \geq T_1 \\ 0, & E(n,k) < T_1 \end{cases}$$

$$S_v(n) = \sum_{k=1}^{N} D_k$$

- T_1 is a predefined threshold.
- n is frame number.
- k is the index of FT coefficient.
- N is the frame size.

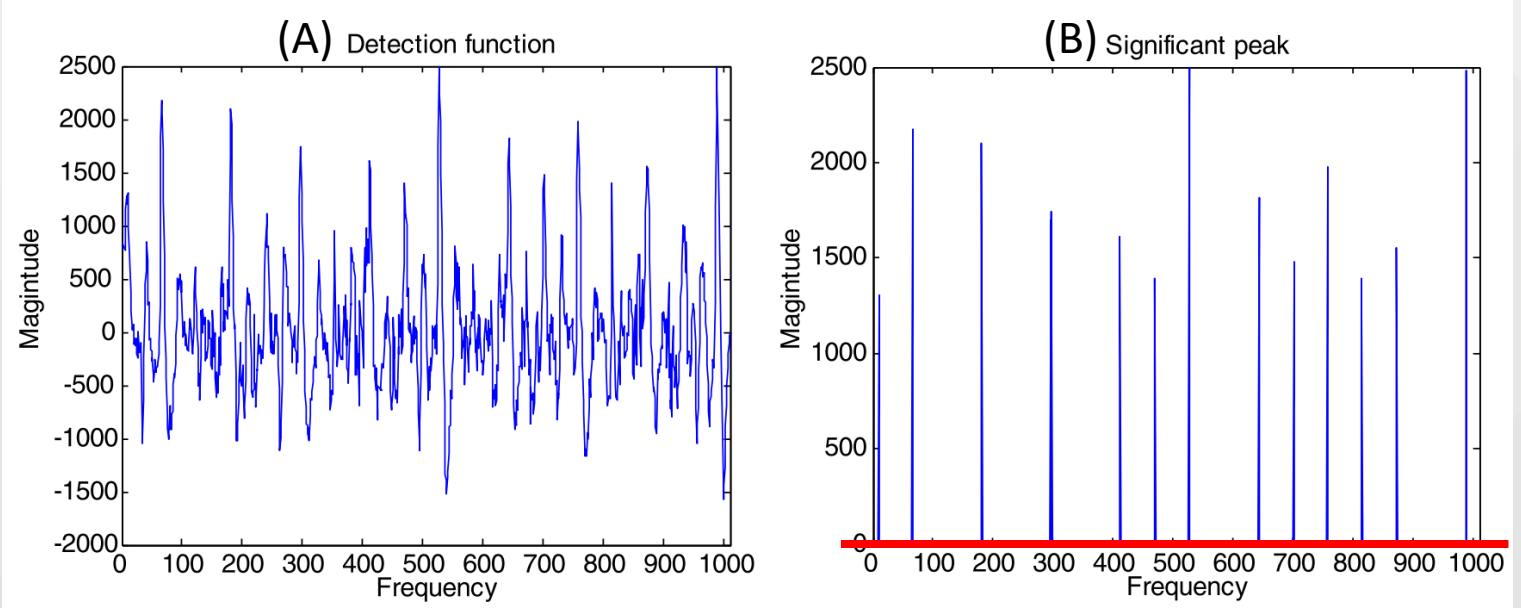


Fig. 10 An example of A detection function and B its significant peak extraction

Proposed Method

c) Hierarchical Adaboost Classifier

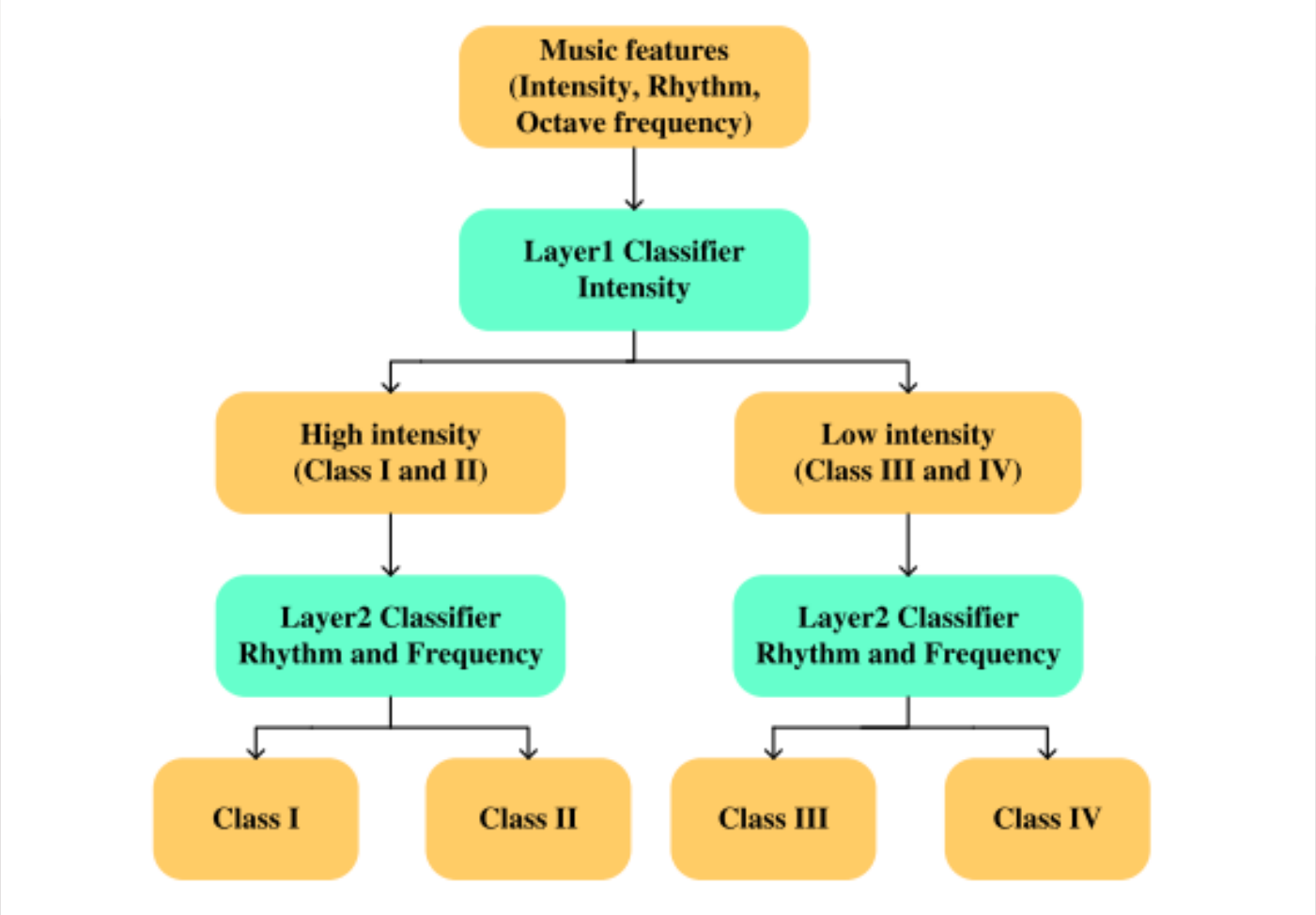


Fig. 11 Structure of hierarchical Adaboost classifier

Experimental Results

a) Database

	Own database	MIREX 2009 Collection(Beatles)
Training	200	40
Testing	150	39
Total	350	79

b) Precision & Recall

$$P = \frac{|\{P_m\} \cap \{P_h\}|}{|\{P_m\}|} \qquad R = \frac{|\{P_m\} \cap \{P_h\}|}{|\{P_h\}|}$$

P_m represents the set of segments detected by the proposed algorithm.
P_h represents the set of segments identified as chorus by human directly.

http://www.music-ir.org/mirex/wiki/2009:Structural_Segmentation

Experimental Results

c) Results of chorus detection

Table. 2 Results of chorus detection

	Total length (sec)				Recall	Precision
	Test songs	Correct	False	Miss		
Chorus section	35515	15454	2959	767	95.27 %	83.93 %

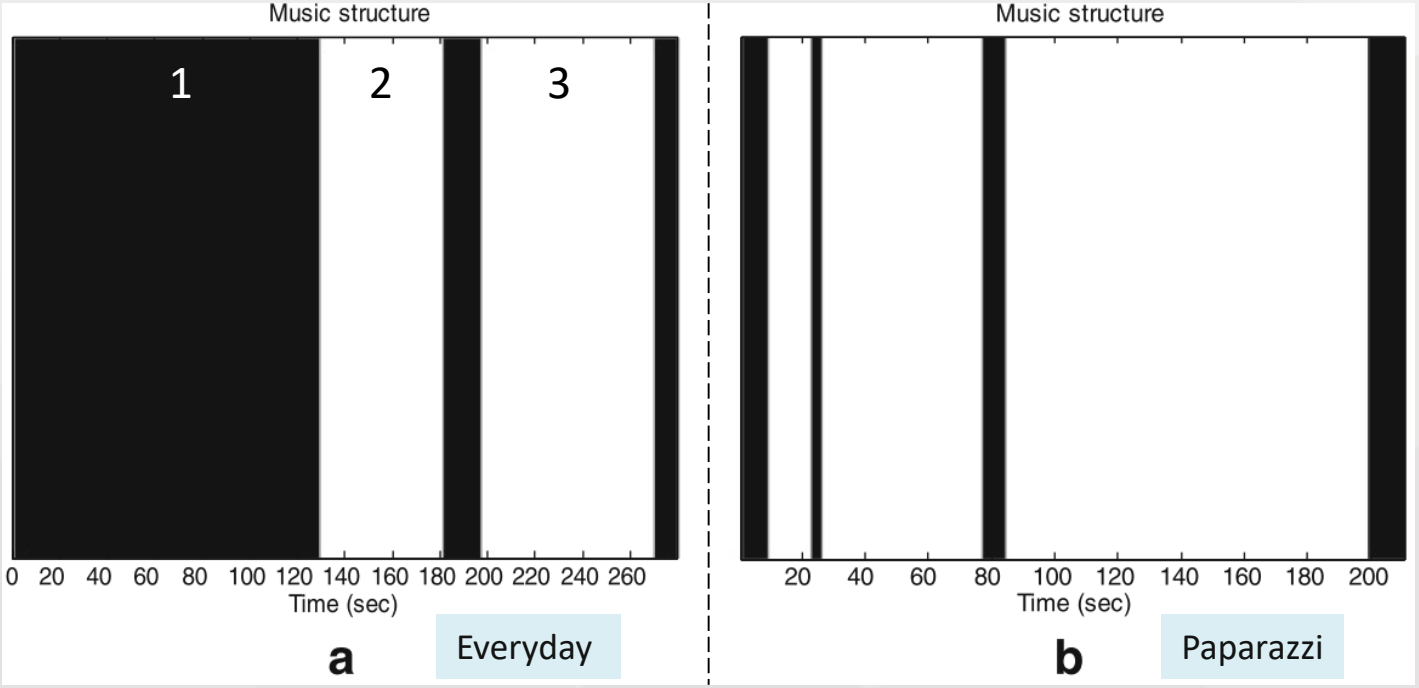


Fig. 12 Colormap, structure information and chorus section

Experimental Results

d) Results of emotion detection

$$Precision = \frac{Correct}{Detected}$$

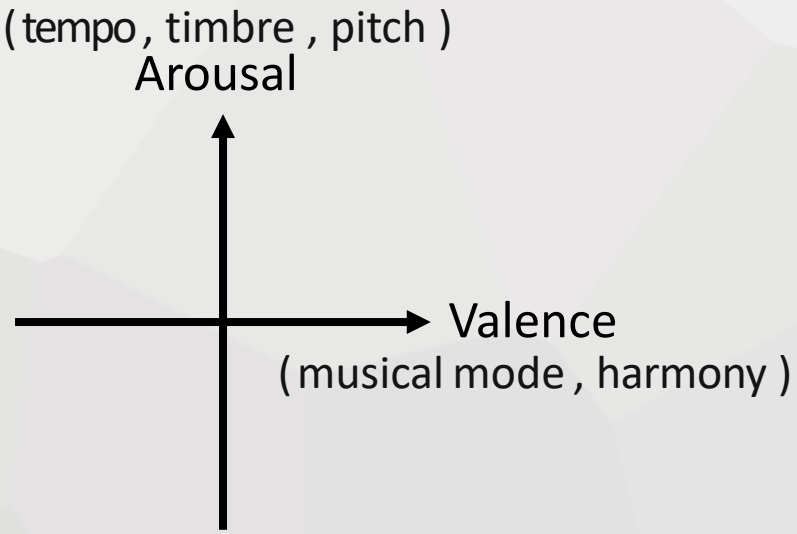
$$Recall = \frac{Correct}{Number\ of\ songs}$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

average precision rate = 92 %

Table. 3 Intermediate comparisons of the proposed emotion detection

	Number of songs	Detected	Correct	Precision	Recall	F-measure
Class I	53	53	50	94.34 %	94.34 %	0.9434
Class II	19	20	19	95.00 %	100 %	0.9744
Class III	34	37	31	83.78 %	91.18 %	0.8732
Class IV	44	40	38	95.00 %	86.36 %	0.9048
Class I & Class II	72	73	69	94.52 %	95.83 %	0.9517
Class II &Class III	53	57	50	87.72 %	94.34 %	0.9091
Class III& Class IV	78	77	69	89.61 %	88.46 %	0.8903
Class IV& Class I	97	93	88	94.62 %	90.72 %	0.9263



Experimental Results

d) Results of emotion detection

Table. 4 Results of emotion detection for our own database

	Class	Number of songs	Detected	Correct	Precision	Recall	F-measure
Chang et al. [7] (2011)	I	53	21	16	76.19 %	30.19 %	0.4324
	II	19	20	6	30.00 %	31.58 %	0.3077
	III	34	83	29	34.94 %	85.29 %	0.4957
	IV	44	26	14	53.85 %	31.82 %	0.4000
	Average				48.74 %	44.72 %	0.4090
Chin et al. [9] (2013)	I	53	69	41	59.42 %	77.36 %	0.6721
	II	19	15	4	26.67 %	21.05 %	0.2353
	III	34	37	18	48.65 %	52.94 %	0.5070
	IV	44	29	12	41.38 %	27.27 %	0.3288
	Average				44.03 %	44.66 %	0.4358
Proposed	I	53	53	50	94.34 %	94.34 %	0.9434
	II	19	20	19	95.00 %	100.00 %	0.9744
	III	34	37	31	83.78 %	91.18 %	0.8732
	IV	44	40	38	95.00 %	86.36 %	0.9048
	Average				92.03 %	92.97 %	0.9239

Table. 5 Results of emotion detection for MIREX 2009 Collection database

	Class	Number of songs	Detected	Correct	Precision	Recall	F-measure
Chang et al. [7] (2011)	I	10	13	6	46.15 %	60.00 %	0.5217
	II	4	4	1	25.00 %	25.00 %	0.2500
	III	13	8	2	25.00 %	15.38 %	0.1905
	IV	12	14	8	57.14 %	66.67 %	0.6154
	Average				38.32 %	41.76 %	0.3944
Chin et al. [9] (2013)	I	10	13	5	38.46 %	50.00 %	0.4348
	II	4	19	3	15.79 %	75.00 %	0.2609
	III	13	2	2	100.00 %	15.38 %	0.2667
	IV	12	5	2	40.00 %	16.67 %	0.2353
	Average				48.56 %	39.26 %	0.2994
Proposed	I	10	9	5	55.56 %	50.00 %	0.5263
	II	4	6	3	50.00 %	75.00 %	0.6000
	III	13	13	6	46.15 %	46.15 %	0.4615
	IV	12	11	6	54.55 %	50.00 %	0.5217
	Average				51.56 %	55.29 %	0.5274

Experimental Results

e) Results of emotion detection of cover songs

- collect 10 cover songs in different languages (same melody different lyrics)
- The precision of cover songs' emotion detection is approximately 90 %.

Table. 6 Results of emotion detection of cover song database

Test music	Emotion	Detected emotion
A_CH	1	1
A_KR	1	1
B_CH	3	3
B_US	3	3
C_CH	2	2
C_KR	2	3
D_CH	1	1
D_KR	1	1
E_CH	3	3
E_JP	3	3
E_TW	4	4
F_TW	3	3
F_JP	3	3
G_CH	4	4
G_US	4	4
H_UK	1	1
H_CH	1	1
I_JP	2	2
I_CH	2	2
J_JP	2	2
J_CH	3	3

Conclusion

- A new approach of mapping the audio signals into the red, green and blue color space.
- Emotion classification is based on 3 music features : intensity , rhythm regularity , frequency subband. In future, music search/retrieval by emotion is highly desired and most straightforward for users.
- Emotion detection algorithm provides consistent results for the same melody sang in different languages and lyrics (cover songs).

Comment

- Use the elements of music as features for classification, which is different from the previous use of text-based classification.
- I think there is a chance to improve the accuracy of chorus detection and emotion detection.
- Increase the size of the data set and modify the content of the data set
- The database defined by itself has no public content.
- The application of audio into a picture (the audio mood corresponds to the picture of the emotion)