哼唱式旋律之 MP3 資料庫檢索系統

Melody-Based Retrieval for MP3 Database by Humming

蔡易行
工業技術研究院電腦與通訊工業研究所
新竹縣竹東鎮中興路 4 段 195-11 號
03-5917948
alientsai@itri.org.tw

陳建發
工業技術研究院電腦與通訊工業研究所
新竹縣竹東鎮中興路 4 段 195-11 號
03-5915639
cfchen@itri.org.tw

摘要

在本論文中，將針對音訊內容之資料擷取(檢索，搜索，建立索引)建立一個 MP3 音樂資料庫多媒體管理系統(MMS: Multimedia Management System)，並使用音樂特性-音高、音程、旋律爲此 MP3 音樂資料庫的內涵式特徵值，經由 MPEG-7 的 DDL(Description Definition Language)將音高、音程、旋律等特徵值封裝成 metadata，同時將這些 metadata 儲存於 metadata 資料庫中，本系統即利用 metadata 來進行搜尋及過濾的動作，並提供下列三種方式讓使用者查詢:Query by example、Query by (default) index、Query by humming。在資料壓縮率爲 30~35 倍時，其查詢的命中率約爲 80%，資料濾除率約爲 56%。

關鍵字

MP3 音樂，MMS，多媒體音樂資庫， MP3 音樂的內涵式分析，MPEG-7，多媒體無線通訊。

1. 簡介

伴隨著 MP3 音樂物件的大量出現，以及網際網路的蓬勃發展，娛樂多媒體音樂已經成爲家庭必備的影音享受之一了例如，從早期只能在個人電腦(PC)上使用 WINAMP 應用軟體來播放 MP3 音樂，到 MP3 音響的硬體播放器，進而日本的公司 APPLE 推出 iPOD 的大容量(約可容納 1000 首 MP3 的流行歌曲，其儲存設備爲 5~10GB)MP3 隨身聽，在在都說明了在 MP3 音樂物件風行後，造就了一個數位多媒體 IT(Information Technology)時代。再輔以無線通訊的快速發展，未來多媒體資料與無線通訊的結合應用，是一股不可抵擋的風潮。

在多媒體音訊的研究中，許多的文獻都是針對非壓縮的音訊格式(如 WAVE，MIDI，…等)來深入探討[1][2][3][4]。而針對這些非壓縮的多媒體音訊的管理似乎都是以 MIDI 資料庫爲主[9][10][11][12][13][14]，其中以清華大學資工系張智星教授的超級點歌王最爲著名[15]。針對壓縮的音訊格式(MP3)文獻探討[5][6][7][8]中，尚未有內涵式的多媒體管理系統，而 MP3 的音樂檔案又如雨後春筍般的在網路上氾濫，所以爲此龐大的 MP3 資料設計一套使用者便利的查詢介面，是一件刻不容緩的事。爲了讓如此大量的多媒體數位影音資訊，包括了影音視訊（CD、VCD、DVD）、圖像資料（images）等，能夠讓使用者在網路上（有線的網際網路及無線的行動上網）快速、精確且使用最自然的語法來獲得使用者想要的多媒體資訊，而最重要的一點是要能夠具有跨平台的特性， MPEG-7 [16][17]就是依此目標爲出發點所制定的國際標準。Quackenbush[18]簡介 MPEG-7 的音訊標準，並說明其音訊標準結構以及音訊描述符號，同時以這些音訊描述符號來舉例說明 MPEG-7 音訊標準在網際網路上的應用。

Michael [19]提出以多媒體描述架構（MDF：Multimedia Description Framework）來設計一個不論是否以 MPEG-7 組成的多媒體檔案皆可以 MDF 來組成的多重描述整合系統架構，也

就是說，MDF 不僅能夠以物件資料內涵特徵的詮釋資料(METADATA)來描述多媒體檔案，而且可以將傳統的文數字特徵描述以 MDF 來表示。Michael [20]以符合 MPEG-7 所規範的多媒體描述定義語言（MD^2 L：Multimedia Description Definition Language）來定義各種的媒體資料如音訊、視訊、…等的描述大綱定義（DSD：Description Scheme Definition），以這些多媒體物件的詮釋資料(Metadata)來當成資料庫索引以及資料查詢比對的依據。

本論文的組織結構說明如下：在第二節裡，我們將簡介 MPEG-7 的標準；在第三節裡，我們會說明 MP3 的音訊標準；在第四節裡，我們會介紹以 MPEG-7 封裝而成的 MP3 詮釋資料(Metadata)爲基礎的 MMS 多媒體 MP3 音樂資料庫；在第五節裡，介紹 MP3 描述定義語言：在第六節裡，我們會以實作的 MMS 來分析其效能：而在最後一節裡爲結論及未來工作的說明。

2. MPEG-7 簡介

MPEG 爲 Motion Picture Experts Group 的簡稱，MPEG 是 ISO 國際組織下的一個工作小組，MPEG-7 爲該組織於 2001 年所訂定的標準，其主要的用途是不同於 MPEG-1、MPEG-2、MPEG-4 的標準，如表 1 所示。

表 1. MPEG 標準比較

標準	主要用途
MPEG-1	VCD 影音壓縮規格
MPEG-2	DVD 影音壓縮規格
MPEG-4	影音串流資料
MPEG-7	多媒體資料內涵描述語言

一般而言，我們會以多媒體內涵的描述介面之觀點來研究 MPEG-7，也就是說是一種用來描述形容資料的資料（data about the data），我們稱之詮釋資料(METADATA)。這跟以往 MPEG-1、MPEG-2 致力於多媒體資料壓縮的規範不同，更與 MPEG-4 以更高層次的內涵式技術來代表多媒體本身的意涵，所以我們可以這樣說 MPEG-1、MPEG-2、MPEG-4 針對多媒體資料提供使用者以內涵式的壓縮技術來獲得高品質而低容量的檔案，而 MPEG-7 就是制定關於多媒體內涵描述的標準（請參閱圖 1 所示），進而讓那些以 MPEG-1、MPEG-2、MPEG-4 標準所產生的多媒體物件能夠更快速、更便利且更準確的讓使用者在網路上取存這些影音娛樂資料，我們圖 2 來說明 MPEG-7 的應用及服務範圍。

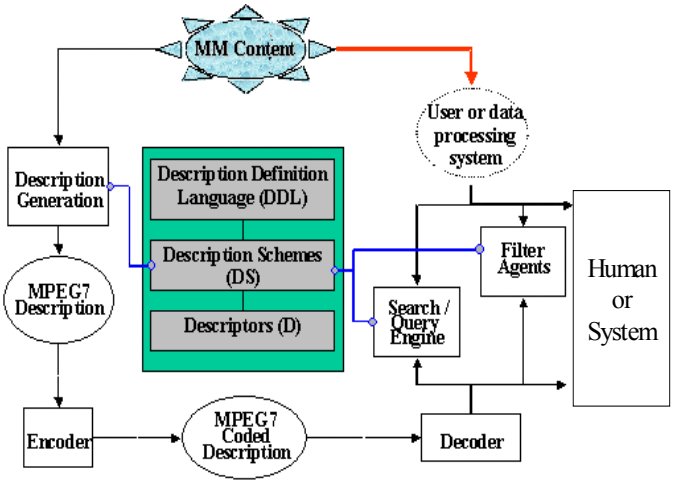

圖 1. 以 MPEG-7 描述多媒體物件內涵的方法 [21]

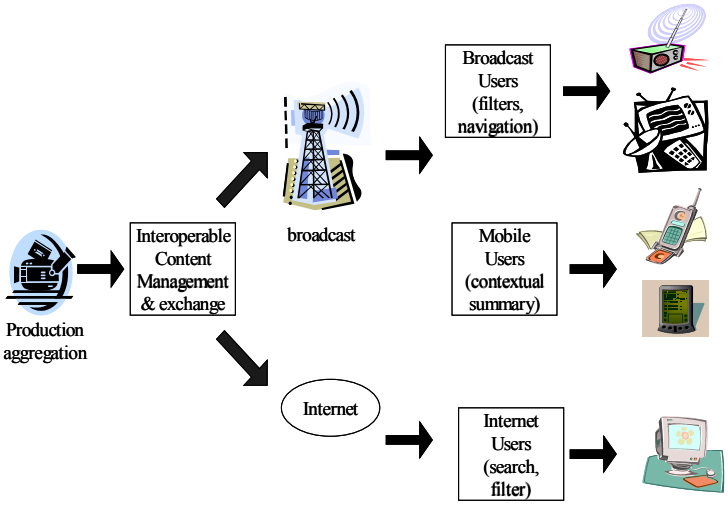

圖 2. MPEG-7 的作用示意圖 [18]

如圖 2 所示，MPEG-7 期望能以內涵式的多媒體物件描述讓各種不同的通訊器材如行動無線設備、視訊轉換器（set-top box）、個人電腦等皆能透過 MPEG-7 成爲一種彼此溝通的橋樑，進而成爲一個可處處都能找到有意義資料的資訊空間（Data everywhere）。

3. MP3 簡介

MP3 是 MPEG-1 Layer 3 的縮寫，其主要用途就是提供多媒體音訊的壓縮標準，而主要特徵就是能夠提供使用者高品質、高壓縮率的音訊資料，表 2 爲 MPEG-1 Layer 1、2、3 的比較表。

表 2. MPEG-1 Layer 1、2、3 的比較表

	壓縮比	音質
Layer 1	1:4	差
Layer 2	1:6~8	尚可
Layer 3	1:10~12	CD 音質

3.1 MP3 編碼原理

MPEG-1 音訊編碼主要是由四個方塊所組成的，其基本原理是將聲音訊號由時間域（time domain）轉換成頻率域（frequency domain），而且僅保留人類聽覺範圍 20~20kHz 的聲音訊號，因此我們亦可以將 MPEG-1 音訊編碼稱之爲聽覺心理學的演算法（psychoacoustic algorithm）。如下圖 3 所示。

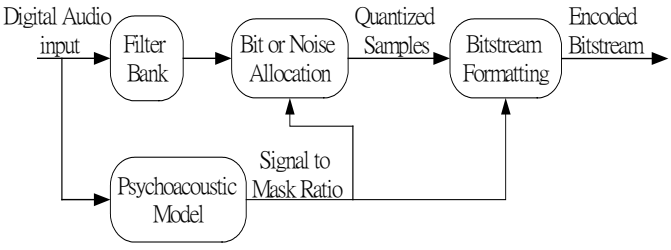

圖 3. MPEG-1 音訊編碼流程方塊圖 [22]

(1) Filterbank：以 MPEG-Audio 的演算法提供間域與頻率域的對應轉換。

(2) Psychoacoustic Model：以聽覺心理學的原理爲依據，針對 Filterbank 的每個子頻帶進行聲音訊號的遮罩。

(3) Bit or Noise Allocation：爲了能夠符合編碼器的位元率（bitrate）及遮罩（masking）需求，配置器會以 Filterbank 的取樣及 Psychoacoustic Model 所產生的 SMR 來調整位元或是雜訊配置。（SMR = Signal to Mask Ratio）

(4) Bitstream Formatting：將 Filterbank 輸出量化及位元或雜訊配置以及其它相關資料如 side information 等封裝成 MP3 格式。

3.2 MP3 解碼原理

圖 4. MPEG-1 音訊解碼示視圖 [22]

我們可以將 MP3 的解碼程序看成是反方向 MP3 編碼程序，如圖 4 所示。而 MP3 的解碼程序方塊圖如下圖 5 所示，詳細說明請參閱 [22]。

圖 5. MPEG-1 音訊解碼流程方塊圖

4. MMS 系統簡介

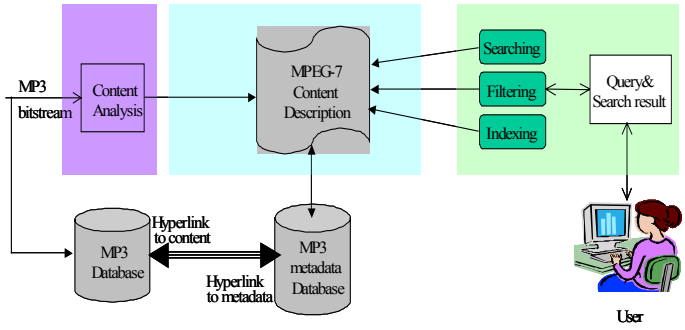

圖 6. MMS architecture

如上圖 6 所示，MMS 系統主要是由三個模組所構成的，MP3 objects content analysis，MPEG-7 content description，Matching/filtering framework。

(1)MP3 objects content analysis：此模組爲 MP3 音訊的內涵式分析，亦即萃取出 MP3 音樂的旋律特性。圖 7 爲 MP3 的特徵值係數擷取點。

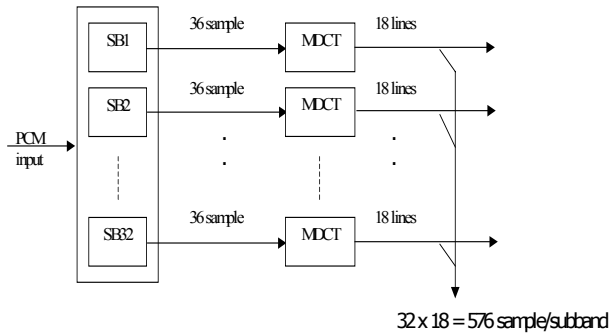

圖 7. 特徵值擷取點

我們使用 MDCT 的 576 個聲音訊號能量強度來計算出此一音框的最大能量是位於那一條頻率線，以及屬於那個子頻帶區間，進而將此音框推算出其音符，我們稱之爲 MP3 音高。因此我們定義 MP3 旋律線即爲一連串的 MP3 音高所組成的。而音程即爲彼此相鄰的 MP3 音高之差距。

(2)MPEG-7 content description：以 MPEG-7 的多媒體內涵描述介面(Multimedia Content Description Interface)來完成不同作業平台的交換機制。其 MP3 特徵值描述如下圖所示。

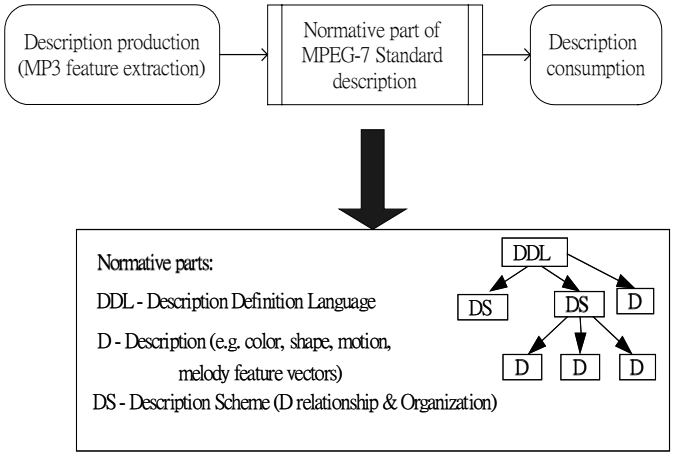

圖 8. MP3 的特徵值描述

(3)Matching/filtering framework：爲了評估使用者輸入的 MP3 音樂與資料庫中 MP3 音樂的相似程度，我們依據二者之間旋律的差異，以完全搜尋（Full Search）的方式計算出彼此之間的歐基里德距離（Euclidean distance），並以此來比對使用者輸入 MP3 音樂與資料庫中 MP3 的相似程度，如圖 9 所示。

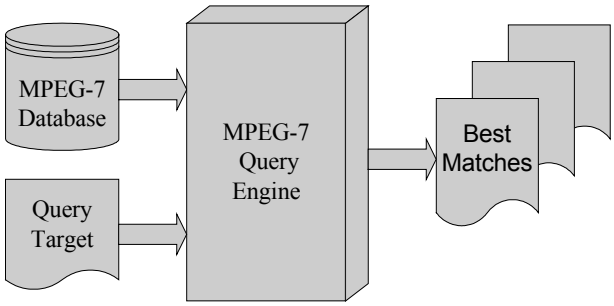

圖 9. 比對機制

5. MP3 Description Definition Language

5.1 MP3MelodyLine

5.1.1 Introduction

The MP3MelodyLineDS is a absolution for MP3 melodic information at the specification – Stereo，16 Bits，44.1 KHz. The MP3MelodyLineDS uses the real note position at the staff，i.e. the Melody contour is built by a serial notes.

5.1.2 MP3MelodyLineType

5.1.2.1 Syntax

```
<!-- ############################ -->
<!-- Definition of MP3MelodyLineType -->
<!-- ############################ -->
<complexType name=" MP3MelodyLineType ">
  <complexContent>
    <sequence>
      <element name="MusicalNote" type="mpeg7:
        MusicalNoteType"/>
      </sequence>
  </complexContent>
</complexType>
```

5.1.2.2 Semantics

Name	Definition
MP3MelodyLineType	A structure containing a absolute representation of the melody with interval MusicalNote and Track.
MusicalNote	The pitch contour of the melody of MusicalNoteType.

5.1.3 MusicalNoteType

The MusicalNoteType descriptor contains the real tones position representation of a melody.

5.1.3.1 Syntax

```
<!-- ############################ -->
<!-- Definition of MusicalNoteType -->
<!-- ############################ -->
<!-- Definition of " MusicalNoteType" -->
  <complexType name=" MusicalNoteType">
    <complexContent>
      <sequence>
        <element name="ToneData" type="PitchDataType"/>
      </sequence>
    </complexContent>
  </complexType>

<!-- ############################ -->
<!-- Definition of PitchDataType -->
<!-- ############################ -->
<simpleType name="PitchDataType">
<!-- 14 pitch values: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 -->
  <list>
    <simpleType>
      <restriction base="integer">
        <minInclusive value="1"/>
        <maxInclusive value="14"/>
      </restriction>
    </simpleType>
  </list>
</simpleType>
```

5.1.3.2 Semantics

Name	Definition
MusicalNoteType	A structure containing the musical note information.
ToneData	A series of integers ranging from 1 to 14, of ToneDataType. See Table 1.

The ToneData values equal each musical note, restricted from 1 to 14.

Table 1 — Musical note mapping table

Pitch value	Musical note
1	C4
2	D4
3	E4
4	F4
5	G4
6	A4

7	B4
8	C5
9	D5
10	E5
11	F5
12	G5
13	A5
14	B5

6. 實驗結果

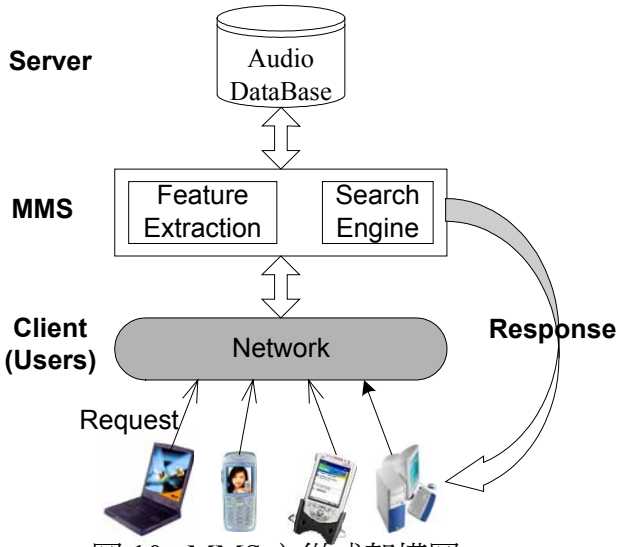

圖 10. MMS 主從式架構圖

MMS 使用主從式的系統架構，如圖 10 所示。我們的初步實驗爲使用者以筆記型電腦配合 802.11b 無線網卡，利用麥克風哼唱式輸入，MMS 會將資料庫中相似的 MP3 音樂索引回傳給使用者。表 2 爲哼唱式查詢的統計表。其中歌唱模式：男男(男生唱男生的歌)，男女(男生唱女生的歌)，女女(女生唱女生的歌)或是女男(女生唱男生的歌)。

表 2. 哼唱式查詢的統計表

歌唱模式	哼唱秒數(單位:秒)	哼唱方式	相似個數(單位:首)	歌唱模式	Hit	搜尋時間(單位:秒)
M_{-1}	3	唱	43	男男	√	17
M_{-2}	8	哼	1	男男		58
M_{-3}	3	唱	39	男男	√	19
M_{-4}	6	唱	3	男男		46
M_{-5}	6	唱	26	男女		35
M_{-6}	4	唱	37	男男	√	19
M_{-7}	8	哼	6	男男	√	52
M_{-8}	12	唱	5	男男		72
M_{-9}	2	唱	50	男女	√	12
M_{-10}	1	唱	51	男男	√	10
M_{-11}	2	哼	41	男男	√	15
M_{-12}	3	唱	13	男男	√	21
M_{-13}	3	唱	51	男男	√	12
M_{-14}	2	哼	5	男男	√	21
M_{-15}	3	哼	49	男男	√	13
M_{-16}	4	哼	22	男男	√	26
M_{-17}	5	唱	18	男男	√	32
M_{-18}	3	哼	20	男女	√	22
W_{-1}	5	唱	24	女女	√	29
W_{-2}	3	唱	27	女女	√	21
W_{-3}	6	哼	16	女女	√	37
W_{-4}	6	唱	17	女男		36
W_{-5}	6	唱	16	女男	√	37
W_{-6}	6	唱	30	女男	√	27
W_{-7}	7	唱	17	女男	√	37
W_{-8}	6	唱	35	女女	√	25
W_{-9}	7	哼	26	女女	√	33
W_{-10}	4	唱	46	女女	√	16
W_{-11}	3	唱	4	女女		26
W_{-12}	6	唱+哼	26	女女	√	29
W_{-13}	4	唱	48	女女	√	15
W_{-14}	7	哼	35	女女	√	32

6.1 實驗分析

經由哼唱式查詢的實驗結果觀察得之，國語流行歌曲（資料量爲 4~5M Bytes）經由 MPEG-7 descriptor 所建立的 MP3 description（STEREO format）其資料量約爲 135~145K Bytes， 所以其壓縮率約爲 30~35 倍，而其效能與歌曲數量之關係如圖 11 所示。其中

◇：代表歌唱模式 M_{-1} 資料庫歌曲數量與花費時間關係曲線

○：代表歌唱模式 M_{-2} 資料庫歌曲數量與花費時間關係曲線

□：代表歌唱模式 M_{-3} 資料庫歌曲數量與花費時間關係曲線

＋：代表歌唱模式 M_{-4} 資料庫歌曲數量與花費時間關係曲線

×：代表歌唱模式 M_{-5} 資料庫歌曲數量與花費時間關係曲線

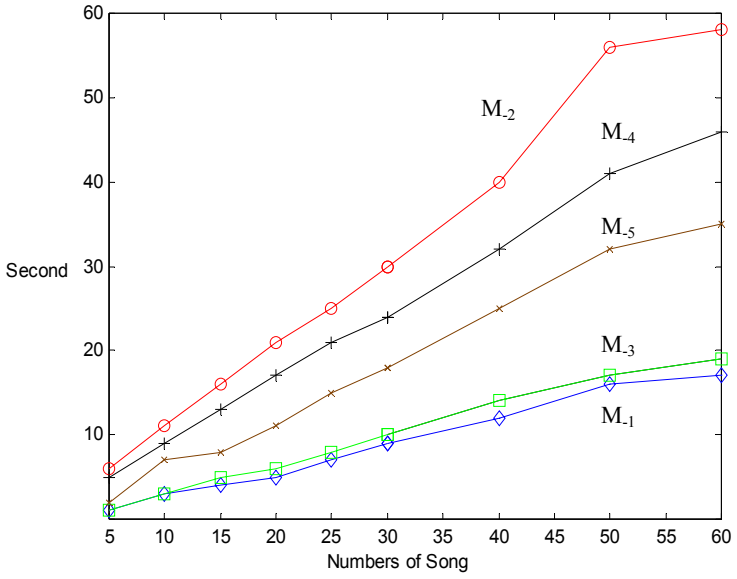

圖 11. 資料庫歌曲數量與搜尋花費時間曲線圖

7. 結論及未來工作

在本論文中，我們提供了使用者以最便利的方式來找尋 MP3 音樂，由實驗結果得知其查詢命中率約爲 80%，我們將以更精確的音樂特性當成我們的 MP3 音樂特徵值來提高查詢命中率。未來我們預計要將此一機制內建於 MP3 隨身聽，MP3 汽車音響，手機，PDA，STB(Set-Top Box)，…等設備，以及將來在數位廣播系統中對多媒體資料之搜尋及過濾提供一解決方案。

8. REFERENCES

[1] T. Zhang and C.-C. Jay Kuo, ”Hierarchical Classification of Audio Data for Archiving and Retrieving,” Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on Volume:6, 1999.

[2] J. Foote, ”A Similarity Measure for Automatic Audio classification,” Proc. AAAI 1997 Spring Symposium on Intelligent Integration and Use of Text, Image, Video, and Audio Corpora. Stanford, March 1997.

[3] G. Lu, and T. Hankinsion, ”A technique towards automatic audio classification and retrieval,” IEEE International Conference on Signal Processing. 1998.

[4] C.-C. Liu, J.-L. Hsu, and A.L.P. Chen, ”An Approximate String Matching Algorithm for Content-Based Music Data Retrieval,” IEEE Multimedia Computing and Systems, Vol.1, July 7-11, 1999.

[5] Y. Nakajima et al, ”A fast audio classification from MPEG coded data,” Acoustics, Speech, and Signal Processing, 1999 Proceeding., 1999 IEEE International Conference.

[6] P. Noll, ”MPEG Digit Audio Coding, ” IEEE Signal Processing Magazine Volume : 145, Sept. 1997.

[7] L. Yapp and G. Zick, ”Speech Recognition on MPEG/Audio Encoded Files , “ IEEE International Conference on Multimedia Computing and Systems, 1997.

[8] G. Tzanetakis and P. Cook, "Sound Analysis Using MPEG Compressed Audio," Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on , Volume: 2 , 2000

[9] J.-S. Roger Jang and Ming-Yang Gao, "A Query-by-Singing System based on Dynamic Programming," International Workshop on Intelligent Systems Resolutions (the 8th Bellman Continuum), 2000.

[10] 許文豪、高名揚、張智星, "直覺式歌唱輸入音樂搜尋引擎," Proceedings of the Fifth Conference on Artificial Intelligence and Applications (第五屆人工智慧與應用研討會), Nov 2000.

[11] A. Ghias, Logan, H., Chamberlain, D., Smith, B. C., "Query by humming-musical information retrieval in an audio database," ACM Multimedia , 1995.

[12] N. Kosugi et al, "A Practical Query-By-Humming System for a Large Music Database," ACM Multimedia , 2000.

[13] N. Kosugi et al, "Music Retrieval by Humming," IEEE PACRIM , 1999.

[14] N. Kosugi et al, "Let's Search for Songs by Humming!" ACM Multimedia , 1999.

[15] http://www.supermbox.com.tw/

[16] S.-F. Chang, A. Puri, T. Sikora, and H. Zhang, "Introduction to the Special Issue on MPEG-7," IEEE Trans. On Circuits and Systems for Video Technology, 2001.

[17] S.-F. Chang, T. Sikora, A. Puri, "Overview of the MPEG-7 Standard," IEEE Trans. On Circuits and Systems for Video Technology, 2001.

[18] S. Quackenbush and A. Lindsay, "Overview of MPEG-7 Audio," IEEE Trans. On Circuits and Systems for Video Technology, 2001.

[19] Michael J. Hu and Ye Jian, "Multimedia Description Framework (MDF) for Content Description of Audio/Video Documents," International Conference on Digital Libraries, ACM Conference on Digital Libraries, 1999.

[20] Michael J. Hu and Ye Jian, "MD^2 L：Content Description of Multimedia Documents for Efficient Process and Search/Retrieval," Research and Technology Advances in Digital Libraries, 1999.

[21] ISO/IEC JTC1/SC29/WG11 N4031, "CODING OF MOVING PICTURES AND AUDIO ," SIngapore, March 2001.

[22] ISO/IEC 11172-3:1993, "Information Technology — Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 Mbit/s — Part 3: Audio."