The 2016 Conference on Computational Linguistics and Speech Processing
ROCLING 2016, pp. 131-141

多通道之多重音頻串流方法之研究

Multi-channel Source Clustering of Polyphonic Music

官志誼 Chih-Yi Kuan
國立中央大學資訊工程學系
Department of Computer Science and Information Engineering
National Central University
103522038@cc.ncu.edu.tw

蘇黎 Li Su
中央研究院資訊科技創新研究中心
Academia Sinica Research Center for Information Technology Innovation
lisu@citi.sinica.edu.tw

秦餘皞 Yu-Hao Chin
國立中央大學資訊工程學系
Department of Computer Science and Information Engineering
National Central University
kio19330@gmail.com

王家慶 Jia-Ching Wang
國立中央大學資訊工程學系
Department of Computer Science and Information Engineering
National Central University
jcw@csie.ncu.edu.tw

摘要

基礎頻率分析在數位訊號處理中是一項重要課題並可以延伸到許多相關的研究，無論是在音樂或者語音上皆有其中要性，本論文主要討論多個單音音源的音頻串流方法，本論文提出之系統需要三個輸入，分別為音源個數、基頻偵測結果、混合音檔。而整體系統可以分為兩個階段，第一階段為依據基頻偵測結果將每一個音高取得相對應特徵參數，第二階段則將上述所有資料進行的聚類，最後輸出各個音源的音頻串流，簡單來說即是每個時刻每個音源演奏哪些音高的資訊。

本論文在特徵參數方面我提出了新的多通道方位特徵參數，並與其他音色特徵參數融合成為更加強健的特徵參數，聚類方面我們基於粒子群最佳化演算法提出了新的架構，並融合領域知識於其中來提高準確率。另外本論文特別針對音源音域接近、音頻串

流纏繞頻繁的音檔來設計並能有更好的準確率。

關鍵詞：基礎頻率分析，音頻串流，粒子群最佳化演算法

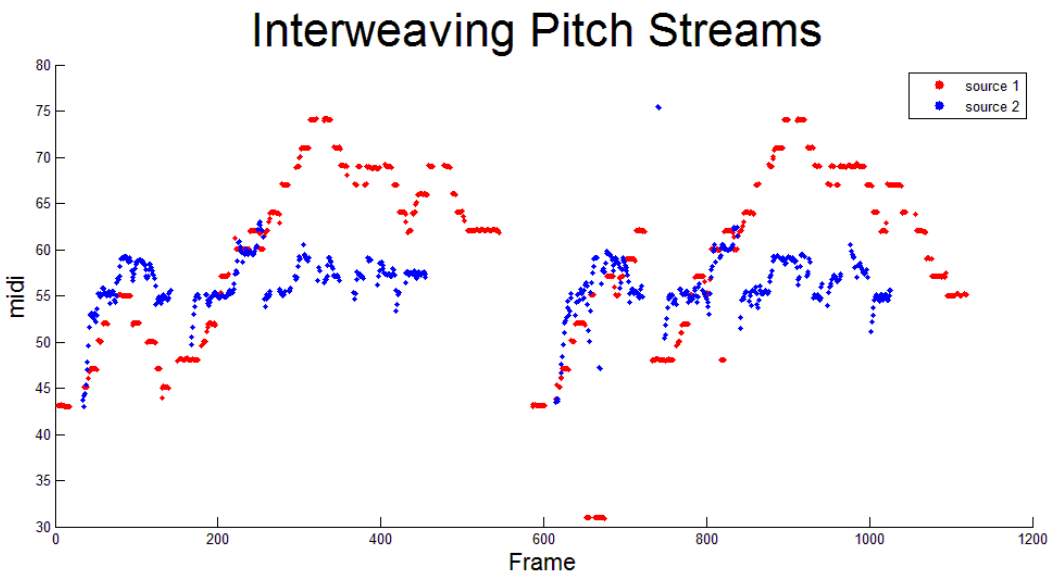
一、緒論

人類的聽覺非常奇妙，我們對音頻與音色的變化相當地敏銳，Robert W. Gutman 在莫札特傳記中寫道:14 歲的莫札特，曾在羅馬西斯廷教堂聽過一次 Gregorio Allegri: Miserere，之後便憑著記憶將曲譜寫下來，並且幾乎沒有存在錯誤。然而這是很有天賦的人或者經過專業訓練的音樂家才能辦到的，因此我們希望能依靠電腦的音訊分析替我們做到這項挑戰。

本論文探討的音頻串流分離(Multi-pitch-streaming)可以應用在樂譜依據的音訊分離(score inform source separation)[1]以及自動轉譜(Automatic music transcription)[2]等等相關議題上，其內容是在已知音源個數的情況下，透過雙麥克風的資訊，將各個音源的音高資訊分析出來，簡單來說即是在某個時刻各個音源發出的音高的資訊。在本研究中，我們需要混合訊號的音檔、音源數、基頻信號偵測的結果，在透過音頻及音色分析之後，最後將每一音源的音頻串流分別輸出出來，但是這項研究仍然是一項挑戰，尤其在音源數目眾多、音源的音高相近、音源音色相近…等，而當音源混合在一起時則頻譜會互相重疊，造成各別音頻的特徵參數擷取不易，為了解決這樣的問題，許多方法被提出像是:採用非負矩陣分解 NMF[3]-[6]，Probabilistic Latent Component Analysis-based 的方法[7][8][9]…等，這些方法都是將各別音頻相對應的頻譜從混合頻譜中分離出來之後再對各別頻譜做處理，而 Baseline 的特徵參數 Uniform Discrete Cepstrum (UDC)[10]，則是直接對混合頻譜做處理，計算某音源在混合頻譜中稀疏且非均勻點的倒頻譜。然而上述方法都是基於單通道下的方法，在本論文中我們則提出新的多通道方法來提升特徵參數的鑑別性。另外當音源音域接近時還會有音高串流之間彼此交纏(interweaving)的問題如下圖一。

特徵參數擷取完成之後下一階段就是將所有資料聚類，在本研究上有很多新的聚類架構被提出，在非監督式的方法有: Spectral Clustering[11][12] ，Baseline 的 Constrained Clustering[10]，而監督式方法有: Factorial Hidden Markov Model [13][14], PLCA spectral dictionary [15]，但監督式方法需要事先訓練，在應用上較為限制，因此本論文專注於非監督式的方法，基於粒子群最佳化演算法提出新的聚類架構，將在後面章節介紹。

圖一、音頻串流交纏的情況



二、特徵參數擷取

本研究在特徵參數擷取上比較特殊，是依據基頻信號偵測的結果(MPE)來做相對應的特徵參數擷取，亦即每一個基頻信號偵測出來的音高都有一個相對應的特徵參數，因此在同一個音框(Frame)中若有多個估測音高(Pitch)就會截取出相對應的多個特徵參數來，再將各個音高相對應的特徵參數作後續的聚類分析，最後把特徵參數相近的視為同一個音源所發出的音高，並將所有同一類別的音高依時間串成各自的音高串流。在特徵參數擷取階段，本論文分為音色特徵參數與方位特徵參數，最後在將兩者特徵參數融合在一起作為最後的特徵參數。

音色特徵部分我們採用 Uniform Discrete Cepstrum (UDC)[10]來做為我們的音色特徵參數，UDC 是一種稀疏、非均勻的倒頻譜表示方法。我們將混合音檔做離散傅立葉

轉換 discrete Fourier transform (DFT)得到混合頻譜，令 $\mathbf{f}=[f_1,...,f_N]^T$ 與 $\mathbf{a}=\left[a_1,...,a_N\right]^T$ 分別為混合頻譜的全頻帶頻率與振幅的對數函數(log-amplitudes)，令 $\hat{\mathbf{f}}=[\hat{f_1},...,f_n]^T$ 與 $\hat{\mathbf{a}}=[\hat{a_1},...,a_n]^T$ 為其子集合，代表我們欲觀察音源在頻譜上其對應音高的頻帶，我們稱為該音源的觀察點，在此方法中我們取輸入基頻的五十個諧波點做為該音源的觀察點，而 UDC 的計算方式如下:

$$\mathbf{F}^{\text{udc}}=\hat{\mathbf{Y}}^T\hat{\mathbf{a}} \tag{1}$$

$$\hat{\mathbf{Y}}=\begin{pmatrix}1 & \sqrt{2}\cos(2\pi 1\hat{f_1}) & \cdots & \sqrt{2}\cos(2\pi(d_{\text{udc}}-1)f_1)\\ \vdots & \vdots & \vdots & \vdots\\ 1 & \sqrt{2}\cos(2\pi 1\hat{f_n}) & \cdots & \sqrt{2}\cos(2\pi(d_{\text{udc}}-1)f_n)\end{pmatrix} \tag{2}$$

在本研究中我們需要處理混合的頻譜與許多盲訊號分離[16]的問題相似，但不同在於盲訊號分離多是處理全頻帶資訊，而我們則是處理特定頻帶的資訊。由於我們的輸入音檔為雙通道的資訊，我們可以藉此來求得方位的資訊。首先，我們先將 *J* 個麥克風所收到的資訊，分別採用短時距傅立葉轉換（short time Fourier transform, STFT）將時域上的混合訊號，轉換成全頻域的混合頻譜(mixture spectrum)，以 $\mathbf{f}^c=[f_1^c,...,f_N^c]^T$ 表示，其中 *N* 為總頻帶個數，f_n^c 為 *c* 麥克風第 *n* 個頻帶的 STFT 係數，之後我們依據輸入時給定位於第 *t* 音框時的第 *p* 個音高(*t*,*p*)，找到該音高對應的頻帶 *n*，並將該對應頻域附近的頻譜(上下 m 個頻帶，一共 2m+1 個)取出為 Ω^c 如下:

$$\Omega^c=\{f_i^c \mid n-m\le i\le n+m\},\ f_i^c \text{ for some } i\in\{1,2,3,...,N\}c\in\{1,2,3,...,J\} \tag{3}$$

其中 *c* 為麥克風編號， *J* 為麥克風數目，我們將 Ω^c 作為估測音高(*t*,*p*)的音源頻譜觀察點，並將這些頻譜分別取能量值(magnitude)後得到 $\mathbf{v}^c$ 如下:

$$\mathbf{v}^c=\left[\left|f_{n-m}^c\right|,...,\left|f_m^c\right|,...,\left|f_{n+m}^c\right|\right]^T \tag{4}$$

之後我們將所有麥克風中的 $\mathbf{v}^c$ 值相加作為該音高的總能量值 $\sum_{c=1}^{J}\mathbf{v}^c$ ，再分別將各自音高頻譜觀察點的能量值除以對應的總能量值，稱能量比值(level ratio) [16]:

$$\mathbf{l}^c = \frac{\mathbf{v}^c}{\sum_{c=1}^{J} \mathbf{v}^c} \tag{5}$$

其中除法為點除。最後我們將音高資料(*t*,*p*)每個麥克風中的$\mathbf{l}^c$值都計算出來，維度為2*m*+1，再將其所有麥克風的值做矩陣合併作為該音高(*t*,*p*)對應的方位特徵參數，維度為$(2m+1)\cdot J$，如下:

$$\mathbf{F}^{\text{level}} = [\mathbf{I}^1, \mathbf{I}^2, \mathbf{I}^3, ..., \mathbf{I}^J]^T \tag{6}$$

分別計算完音色特徵參數$\mathbf{F}^{\text{udc}}$與方位特徵參數$\mathbf{F}^{\text{level}}$之後，下一步我們將其合併成為一個完整的特徵參數，但由於兩個特徵參數的維度不同對於往後聚類方法計算特徵參數距離合時會造成不平等的影響，因此我們採用標準差總和正規畫[17]的方式來合併。首先，我們分別計算出各特徵參數的第 *i* 個維度的標準差(standard deviation)σ_i，再將各自特徵參數的各維度的標準差值相加，算出各別特徵參數的標準差值總和:

$$\sigma^{\text{udc}} = \sum_{i=1}^{d_{\text{udc}}} \sigma_i \qquad \sigma^{\text{level}} = \sum_{i=1}^{d_{\text{level}}} \sigma_i \tag{7}$$

其中d_{udc}，d_{level}分別為特徵參數$\mathbf{F}^{\text{udc}}$及$\mathbf{F}^{\text{level}}$的維度，之後再將兩個特徵參數分別除以各自標準差總和，最後將兩個特徵參數直接作矩陣合併:

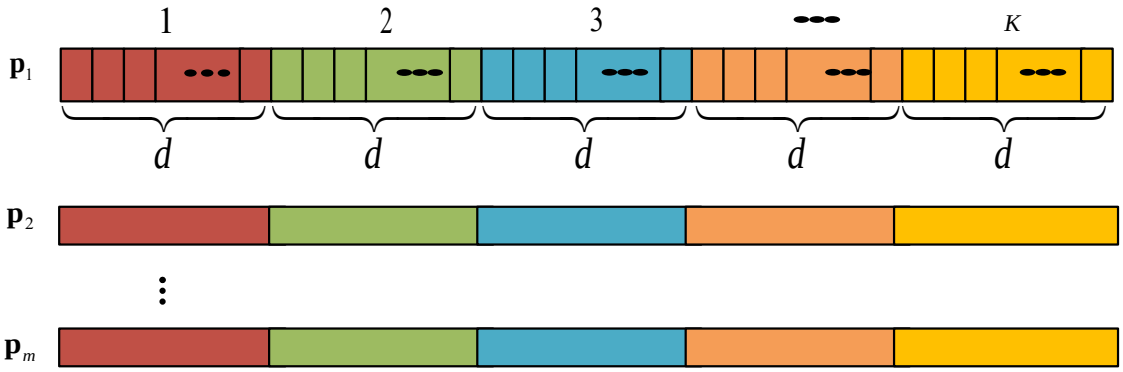
$$\mathbf{F}^{\text{fusion}} = \begin{bmatrix} \dfrac{\mathbf{F}^{\text{udc}}}{\sigma^{\text{udc}}} \\ \dfrac{\mathbf{F}^{\text{level}}}{\sigma^{\text{level}}} \end{bmatrix} \tag{8}$$

我們可以發現在經過正規劃後，維度較高的特徵參數也會有較大標準差總和值，借此來平衡算特徵參數距離時，維度不對等的問題。

三、限制型粒子群最佳聚類演算法

粒子群最佳化演算法運用在聚類問題上有許多研究[18]，而運算核心在於評估各粒子的適應函數，如何根據求解問題設計粒子形式與其適應函數則是關鍵，在本研究中，我們的目的是將各個音高資料作聚類，所以我們的求解問題的答案即是各個音高資料屬於

哪一群聚，而在粒子群最佳化演算法中一個粒子即代表著一組解答，我們使用下述方法來定義粒子形式與計算適應函數值。其概念類似 K-means，不同的地方在於 K-means 的群中心只有一組且是不斷更新後平均得到的，而在粒子群聚類演算法中，我們會生成多組群中心讓它們各自去搜尋最佳結果，每一組群中心代表一個粒子為 $d \times K$ 維的向量，我們定義為 $\mathbf{P}_n$，其中 n 為粒子編號，K 為類別個數(音源數)，d 是特徵參數維度，如下圖:



圖二、以群中心為主的粒子型式設計

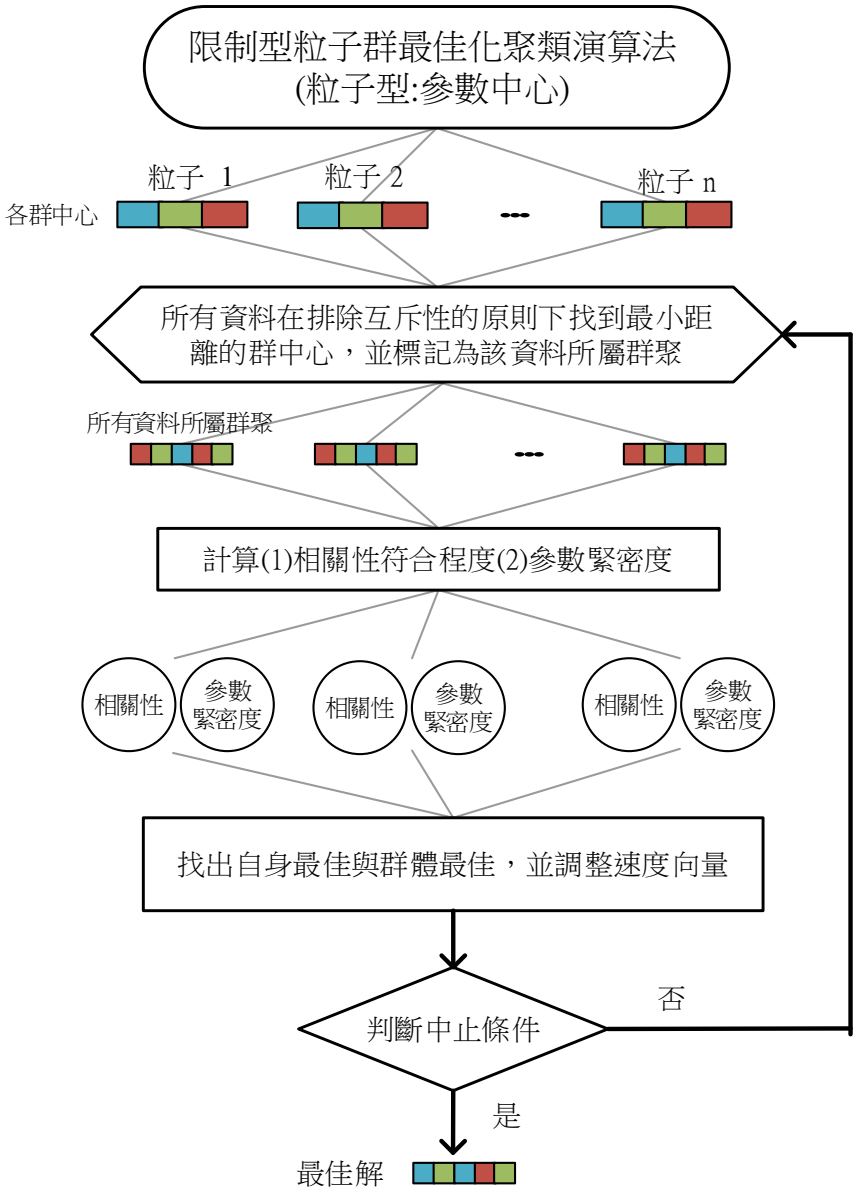
上述的一個粒子 $\mathbf{P}_n$ 即代表著一組群中心，如同 K-meas 一般我們將所有資料依據這樣的群中心來作分類，亦即我們將每筆資料的特徵參數 $\mathbf{F}_{t,p}$ 與每一個粒子的(不同顏色)群中心算距離，取其中最短者將該資料標示為該類別，有了所有資料的所屬群聚之後，我們先計算所有同群聚資料的平均群中心:

$$\mathbf{M}_k = \frac{\sum_{F_{t,p} \in S_k} \mathbf{F}_{t,p}}{\sum_{F_{t,p} \in S_k} 1} \tag{9}$$

其中 S_k 代表第 k 個群聚的特徵參數集合，(t,p) 為 t 音框下的第 p 筆音高資料，$\mathbf{F}_{t,p}$ 為其對應的特徵參數。記算好各群聚的平均群中心 $\mathbf{M}_k$ 後，我們將各音高資料的特徵參數減去相對應的群中心，再將同一群聚的差值總合來得到整體群聚的緊密程度，並定義為該粒子 $\mathbf{P}_n$ 的特徵參數緊密度適應函數如下式，K 為音源個數:

$$fitness(\mathbf{P}_n) = \sum_{k=1}^{K} \sum_{\mathbf{F}_{t,p} \in S_k} \left\| \mathbf{F}_{t,p} - \mathbf{M}_k \right\|^2, \; n = 1, \ldots, m \tag{10}$$

而在本研究中由於我們有兩點可用的領域知識: (1)互斥性:同一時間下的一個音源只會生產一個音高，因此若是兩筆音高資料屬於同一個音框則可以知道它們理論上不該被分至同一類別。(2)相關性:由於音高之間常有連續性，若兩個音框相近且音高值也相近的資料則很有可能是同一類別，因此 我們可以將這樣條件加入在我們評估粒子的適應函數中，其方式為上述每筆資料再依據各粒子聚類時，預先排除掉所有互斥性的可能再去計算適應函數值，另外當所有資料皆依據粒子中心聚類完之後，我們再依據這樣的聚類結果計算相關性的符合程度，並視為該粒子的另一個適應值。最後考量兩個是適應值來對各粒子進行最佳化運算，其整體流程如下圖。



圖三、限制型粒子群最佳化聚類演算法流程圖

四、實驗

本論文實驗輸入使用 Ground Truth 的 MPE 作為輸入，其生成方式為使用單軌音檔透過 YIN[21]製作，混合音檔則使用 Roomsim[20]混合各單軌音源而成，而評估方法為分別計算 Accuracy、Precision、Recall、Avg. Accuracy，其中計算方法如下:

$$Accuracy = TP / (TP + FP + FN) \tag{11}$$

$$precision = TP / (TP + FP) \tag{12}$$

$$Recall = TP / (TP + FN) \tag{13}$$

實驗部分分成兩階段，一是固定聚類方法(K-means)來評估各特徵參數，二是固定特徵參數(Fusion)來評估各聚類方法，實驗資料庫為 Bach10[19]資料庫，為 10 首四個樂器的重奏，以及一首交纏頻繁的範例音檔為 MedleyDB[22]中的曲子 Country 1，由吉他、人聲、貝斯演奏而成。

表一、各特徵參數在 Bach10 資料庫的評估

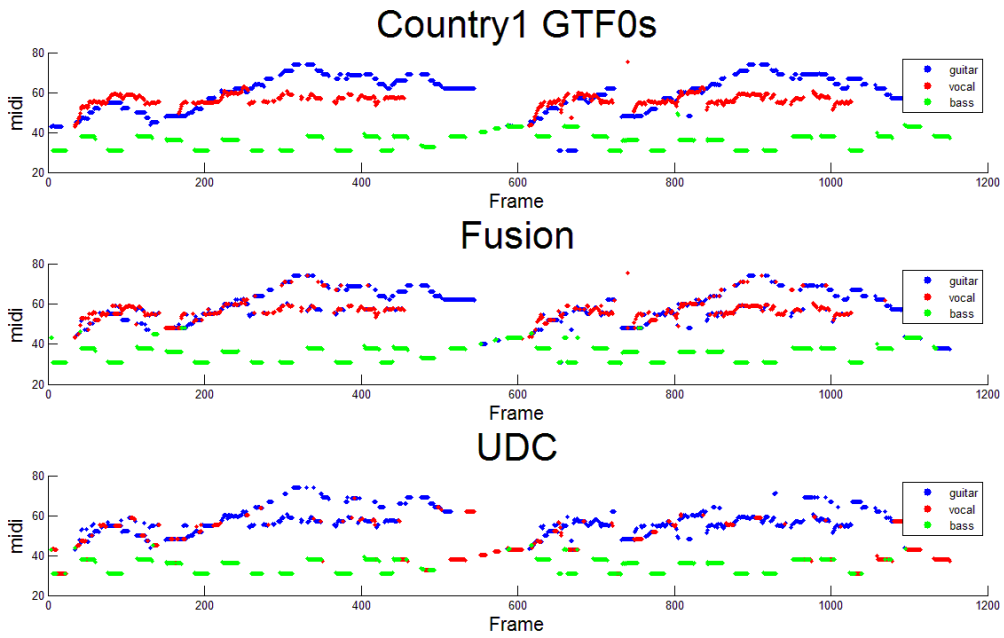
資料庫	特徵參數	Avg. Accuracy
Bach10 Overall	UDC	0.4322±0.0930
	Proposed Fusion	**0.4622±0.0854**

表二、各聚類方法在 Bach10 的評估

資料庫	聚類方法	Avg. Accuracy
Bach10 Overall	K-means	0.4622±0.0854
	Baseline[10]	0.8544±0.0659
	Proposed	**0.8798±0.0465**

表三、特徵參數在交纏音檔 Country 1 的表現

交纏音檔: Country 1				
GT MPE use	特徵參數	Accuracy	Precision	Recall
	UDC	0.3650	0.6278	0.4658
	Proposed Fusion	**0.5692**	**0.7997**	**0.6638**



圖四、特徵參數在交纏音檔 Country 1 的音頻串流圖比較

五、結論

本篇論文提出了新的方位特徵參數並與其他音色特徵參數融合成為更強健的特徵參數，聚類架構部分則基於粒子群最佳化演算法提出了限制型粒子群最佳化聚類演算法，並在準確率上有更好的表現。本論文另一個重點在於處理交纏頻繁的音檔，我們可以從實驗結果得知提出的方法對於這類混合音檔擁有更好的處理能力。

參考文獻

[1] R. Hennequin, B. David, and R. Badeau, “Score informed audio source separation using a parametric model of non-negative spectrogram,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2011, pp. 45–48.

[2]A. Klapuri and M. Davy, Eds.,“Signal Processing Methods for Music Transcription”. New York, NY, USA: Springer, 2006.

[3] T. Heittola, A. Klapuri, and T. Virtanen, “Musical instrument recognition in polyphonic audio using source-filter model for sound separation,” in Proc. Int. Symp. Music Inf. Retreival (ISMIR), 2009.

[4] E. Benetos, M. Kotti, and C. Kotropoulos, “Musical instrument classification using

non-negative matrix factorization algorithms and subset feature selection," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2006, pp. 221–224

[5]R. Jaiswal, D. FitzGerald, D. Barry, E. Coyle, and S. Rickard, "Clustering NMF basis functions using shifted NMF for monaural sound source separation," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2011, pp. 245–248.

[6]F. Rigaud, A. Falaize, B. David, and L. Daudet, "Does inharmonicity improve an NMF-based piano transcription model? " in Proc. IEEE Int.Conf. Acoust., Speech, Signal Process. (ICASSP), 2013, pp. 11–15.

[7] P. Smaragdis, B. Raj, and M. Shashanka, "A probabilistic latent variable model for acoustic modeling," Adv. Models for Acoust. Process.NIPS, vol. 148, 2006.

[8]G. Grindlay and D. P. W. Ellis, "Transcribing multi-instrument polyphonic music with hierarchical eigeninstruments," IEEE J. Sel. Topics Signal Process., vol. 5, no. 6, pp. 1159–1169, Oct. 2011.

[9]V. Arora and L. Behera, "Semi-supervised polyphonic source identification using PLCA based graph clustering," in Proc. Int. Symp. Music Inf. Retreival (ISMIR), 2013.

[10]Zhiyao Duan, Jinyu Han, and Bryan Pardo, "Multi-pitch streaming of harmonic sound mixtures," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 22, no. 1, Jan. 2014.

[11] V. Arora and L. Behera, "Musical source clustering and identification in polyphonic audio IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 22, no. 6, pp. 1003–1012, Jun. 2014.

[12] L. G. Martins, J. J. Burred, G. Tzanetakis, and M. Lagrange, "Polyphonic instrument recognition using spectral clustering.," in Proc. Int. Symp. Music Inf. Retreival (ISMIR), 2007.

[13]M. Wohlmayr, M. Stark, and F. Pernkopf, "A probabilistic interaction model for multipitch tracking with factorial hidden Markov models," IEEE Trans. Audio, Speech, Lang. Process., vol. 19, no. 4, pp. 799–810, May 2011

[14]F. Bach and M. Jordan, "Discriminative training of hidden Markov models for multiple pitch tracking," in Proc. IEEE Int. Conf. Acoust. Speech, Signal Process. (ICASSP), 2005, pp. 489–492.

[15]M. Bay, A. F. Ehmann, J. W. Beauchamp, P. Smaragdis, and J. S. Downie, “Second fiddle is important too: Pitch tracking individual voices in polyphonic music,” in Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR), 2012, pp. 319–324.

[16] Shoko Arakia,, Hiroshi Sawada , Ryo Mukai , Shoji Makino “Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors” Signal Processing 87 (2007) 1833–1847.

[17]Guodong Guo and Stan Z. Li ‘’Content-Based Audio Classification and Retrieval by Support Vector Machines,” IEEE Trans. Neural Neteorks, vol. 14, no. 1, Jan. 2003.

[18] Shuai Li; Xin-Jun Wang; Ying Zhang “X-SPA: Spatial characteristic PSO clustering algorithm with efficient estimation of the number of cluster” Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 2008.

[19] Z. Duan, B. Pardo, and C. Zhang, “Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions,” IEEE Trans. Audio, Speech, Lang. Process., vol. 18, no. 8, pp. 2121–2133, Nov. 2010.

[20] D. Campbell, K. Palomäki, and G. Brown, “A matlab simulation of shoebox room acoustics for use in research and iimm teaching,” Comput. Inf. Syst. J., vol. 9, no. 3, pp. 48–51, Oct. 2005.

[21]A. de Cheveigné and H. Kawahara, “Yin, a fundamental frequency estimator for speech and music,” J. Acoust. Soc. Amer., vol. 111, pp.1917–1930, 2002.

[22] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello. Medleydb: “A multitrack dataset for annotation-intensive mir research,” in Proc .Int. Soc. Music Info. Retrieval Conf., 2014.