Multiple F0 Estimation and Source Clustering of Polyphonic Music Audio Using PLCA and HMRFs

Vipul Arora and Laxmidhar Behera

***Abstract*—Source transcription of pitched polyphonic music entails providing the pitch (F0) values corresponding to each source in a separate channel. This problem is an important step towards many important problems in music and speech processing. It involves 1) estimating the multiple F0 values in each short time frame, and 2) clustering the F0 values into streams corresponding to different sources. We address the problem in an unsupervised way, with only the total number of sources given beforehand. The framework of probabilistic latent component analysis (PLCA) is used to decompose the polyphonic short-time magnitude spectra for multiple F0 estimation and source-specific feature extraction. It is further embedded into the structure of hidden Markov random fields (HMRF) for clustering the F0s into different sources. This clustering is constrained by the cognitive grouping of continuous F0 contours as well as segregation of simultaneous F0s into different source streams. Such constraints are effectively and elegantly modeled by the HMRF's. Simulated annealing varies the degree of constraints for better clustering. The paper also proposes a novel strategy using the trade-off between precision and recall of multiple F0 estimation for better clustering. Evaluations over a variety of datasets show the efficacy of the proposed algorithm and its robustness to the presence of spurious F0s while clustering. It also outperforms a state-of-the-art unsupervised source streaming algorithm in a set of comparative experiments.**

***Index Terms*—Acoustic scene analysis, automatic music transcription, hidden Markov random fields, multiple F0 estimation, polyphonic instrument identification.**

I. Introduction

AUTOMATIC music transcription [1] converts the audio into some form of notation so as to tell which source (instrument or vocal) is playing what. This paper aims at source transcription of pitched polyphonic music, which provides the source-labeled pitch contours of different sources playing simultaneously. Source transcription can easily be extended for applications like source separation [2], instrument/singer identification [3], melody extraction [4], music search [5], music education [6] etc. The concepts used for source transcription are also closely related to those used in speech research, like speech separation [7]. This paper assumes each source to be playing single pitch contour (stream); chords, i.e., one instrument playing multiple notes simultaneously, are not considered here.

Source transcription can be seen as consisting of two subtasks, i.e., (i) estimating the pitch values in each time frame, and (ii) clustering all the pitch values into streams originating from different sources. There have been many works on these two tasks individually. The first task uses the periodicity of audio signal to find its time varying fundamental frequency (F0), and is known as multiple F0 estimation [8]–[10]. It also decomposes the signal into many parts based on the F0 values. The second task utilizes the source specific properties of these signal parts to identify or cluster the F0s originating from the same source. This task is known as musical source identification or clustering.

For multiple F0 estimation, a popular framework is Non-negative Matrix Factorization (NMF), which linearly decomposes the polyphonic magnitude spectra using a dictionary of basis spectra representing individual F0s [9], [10]. For monophonic source identification, a variety of source-specific timbre features can be extracted from the complete signal directly [11], [12]. However, for identifying all the sources playing simultaneously, as in polyphonic music, one has to decompose the signal into several parts, each of which corresponds to a different source, so as to obtain the source-specific features. This problem can be approached in supervised or unsupervised ways. There have been many works on the supervised approach to source identification [13]–[18], where the source specific features are pre-learned using source-labeled audio spectra and used later to recognize the sources in the test audio spectra. Many works model the harmonic envelope smoothness as well as temporal evolution for the decomposition and recognition of sources [14], [17]. Many others adopt NMF or its extension, Probabilistic Latent Component Analysis (PLCA). NMF/PLCA identify the sources by decomposing the polyphonic spectra using source-specific dictionaries of basis spectra [15], [16], [18].

On the contrary, unsupervised source clustering involves generating the source characterizing features from the test audio itself and subsequently using them for source clustering. This approach eliminates the need for labeled monophonic training data, and is also not limited by the sources used and the acoustic conditions during training. There is comparatively much less work done in unsupervised musical source clustering [19], [20]. However, it finds good parallels with the task of speaker diarization [21], [22], which is an active area of research. Moreover, the task of unsupervised source transcription is complementary to the task of unsupervised source separation, which aims at separation of sources into different audio channels. Many such source separation algorithms [23], [24] use NMF to decompose

Manuscript received June 08, 2014; revised September 19, 2014; accepted December 22, 2014. Date of publication January 05, 2015; date of current version January 15, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Nobutaka Ono.

The authors are with the Department of Electrical Engineering, Indian Institute of Technology, Kanpur 208016, India (e-mail: vipular.iitk@gmail.com; lbehera@iitk.ac.in).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2014.2387388

the magnitude spectra into spectral parts and cluster these parts into various sources. Some works [25], [26] also use various constraints like psycho-acoustic cues, temporal continuity, sparseness, etc., to enhance the source clustering. Besides these approaches, there are also semi-supervised approaches that make use of some user-input for source clustering [27]–[29].

For unsupervised source transcription, Duan *et al.* [20] use the output of multi-F0 estimator as the input to source transcription algorithm. It uses a constrained clustering algorithm for timbre based clustering of all the input F0 values. Our previous work [19] also employed constrained graph clustering over all the input F0s. These approaches, however, generally perform the two sub-tasks one after the other. In such a strategy, the errors caused in the first task propagate and adversely affect the accuracy of the second task too. The spurious F0s from the multi-F0 estimator can easily degrade the clustering efficiency, as there are no source specific cues other than the ones generated from the test audio using the F0 values. In this paper, we propose to alleviate this problem with the help of a novel strategy that utilizes the trade-off between precision and recall of the multi-F0 estimator.

This paper presents a novel scheme for the source transcription of pitched polyphonic music. Firstly, the multi-F0 values are extracted using source-filter model based Probabilistic Latent Component Analysis (PLCA) with harmonic dictionaries. Mostly NMF/PLCA based multi-F0 estimation schemes use spectral dictionaries for a fixed number of discrete F0 values [9], [10]. However, the proposed scheme uses source-filter model based harmonic dictionaries, which can estimate F0 values with better resolution and less computations. PLCA decomposition is also used to find source characterizing features by decomposing the polyphonic spectra according to the F0 values. The proposed multi-F0 estimator has a tunable parameter for varying the recall and precision of results. This is useful for the proposed precision-recall trade-off based strategy for unsupervised source clustering.

Secondly, the unsupervised source clustering scheme clusters the F0s into source specific clusters. This clustering is enhanced by various constraints. Though such constraints have also been considered by Duan *et al.* [20] and our previous work [19], the present work proposes a novel Hidden Markov Random Field (HMRF) model to cluster the F0s into source clusters, while taking into account these constraints. HMRF's have been a popular and elegant framework for unsupervised constrained clustering. An HMRF can be seen as an extension of Hidden Markov Models (HMM). An HMM is generally used to model a stochastic process in one dimension while an HMRF can efficiently model a 2-dimensional (2D) graph. HMRF's are popularly used for constrained clustering in image processing [30], [31]. The present work models the F0s as a 2D time-frequency graph and applies HMRF's for source clustering. The proposed HMRF's make use of simulated annealing to change the relative importance of timbral features and the degree of constraints so as to improve the clustering. Also, the proposed method uses under-clustering followed by agglomerative clustering in order to enhance the probability of forming the clusters more accurately. Section II describes the multiple-F0 estimation using PLCA. The framework for source clustering is detailed in Section III. Section IV gives a sketch of the complete implementation along with the precision-recall based strategy. Evaluations and discussions are presented in Section V, followed by conclusions.

II. Multiple F0 Estimation

The audio waveform is downsampled to 10 kHz and thence transformed to magnitude spectrum, $\bar{V}(f,t)$ with multi-resolution 2048-point short-time Fourier transform, using hanning window at 10 ms hop size. Here, f and t index the frequency bins and the time window, respectively. The total number of frequency bins is N_f and that of time frames is N_t. In order to have high frequency resolution at lower frequencies, i.e., up to 1000 Hz, a window of length 112 ms is used, and for frequencies above 1000 Hz, a window of length 56 ms is used. The obtained magnitude spectra $\bar{V}(f,t)$ are normalized with average sum per frame,

$$V(f,t) = \frac{\bar{V}(f,t)}{\sum_{f,t} \bar{V}(f,t)/N_t}.$$

This helps in reducing the effect of overall sound level on threshold values.

To extract the multiple F0 contours, first a larger number of F0 candidates are chosen at each time frame, using a harmonic structure based score function. From these F0 candidates, subsequently, multi-F0 contours are selected using PLCA decomposition, as described below.

A. F0 candidate Estimation

At each t, a harmonic score S is calculated at each spectral peak, F, lying between a range of frequencies, $F_{\min}$ to $F_{\max}$, as

$$S(F,t) = \sum_{f=1}^{N_f} \sum_{n=1}^{\lfloor N_f/F \rfloor} V(f,t) \frac{e^{(f-nF)^2}}{f}. \qquad (1)$$

Here, $\lfloor N_f/F \rfloor$ gives the number of harmonics and the $e^{(f-nF)^2}/f$ term models the harmonic spectrum with a roll-off of 6 dB/octave. A fixed number of peaks, F, with the largest scores are chosen as the F0 candidates. Further, the F0 candidates with $S(F,t)$ less than a threshold are rejected to avoid noisy peaks and also to save further processing time. The F0 candidates at time t are indexed with p and are denoted as $F0_{pt}$.

B. PLCA decomposition for Multi-F0 Estimation

A large number of F0 candidates obtained above have to be pruned so as to reject the spurious candidates. Further, they are used to decompose the polyphonic spectra, in order to obtain source-characterizing features for source clustering.

The magnitude spectrum of polyphonic audio, $V(f,t)$, is assumed to be a linear combination of the magnitude spectra of individual sources, and is decomposed using PLCA [32]. PLCA assumes the magnitude spectrum $V(f,t)$ to be generated by sampling from an underlying probability distribution function (pdf) $P_t(f)$. We use source-filter based PLCA [13], [16] with latent variables - p, s, a- which can take N_p, N_s, N_a values, respectively. The variable

- p_t indexes the F0 value at t;
- s_{pt} indexes the source underlying pitch p at t;
- a_s indexes the band pass filters, characterizing source s.

The subscripts are dropped for brevity, as the conditional pdf's below are defined in a way that takes care of these dependencies. The spectral pdf $P_t(f)$ is decomposed into a weighted mixture of the basis spectra $P(f|p,a)$ as

$$P_t(f) = \sum_{p,s,a} P_t(f|p,a)P_t(p)P_t(s|p)P(a|s) \tag{2}$$

Here, $P_t(p)$ represents the weight of $F0_{pt}$; $P_t(s|p)$ is the probability that $F0_{pt}$ belongs to the source s; and $P(a|s)$ is the weight of the ath band pass filter for the source s and is independent of t. This band pass filter bank has N_a triangular magnitude filters with centres uniformly distributed on a mel-frequency scale [16]. The basis spectrum $P_t(f|p,a)$ consists of Gaussian harmonic peaks at integer multiples of $F0_{pt}$ and passed through the ath band pass filter. Hence, it has a pre-defined shape of partials as well as spectral envelope, while the positions of partials depend upon $F0_{pt}$. The unknown PLCA parameters can be estimated using Expectation Maximization algorithm to minimize the Kullback-Leibler (KL) divergence between $P_t(f)$ and suitably normalized $V(f,t)$ [32]:

E-step:

$$P_t(p,s,a|f) = \frac{P_t(f|p,a)P_t(p)P_t(s|p)P(a|s)}{P_t(f)} \tag{3}$$

M-step:

$$P_t(p) = \frac{\sum\limits_{f,s,a} V(f,t)P_t(p,s,a|f)}{\sum\limits_{p}\sum\limits_{f,s,a} V(f,t)P_t(p,s,a|f)} \tag{4}$$

$$P_t(s|p) = \frac{\sum\limits_{f,a} V(f,t)P_t(p,s,a|f)}{\sum\limits_{s}\sum\limits_{f,a} V(f,t)P_t(p,s,a|f)} \tag{5}$$

$$P(a|s) = \frac{\sum\limits_{f,t,p} V(f,t)P_t(p,s,a|f)}{\sum\limits_{a}\sum\limits_{f,t,p} V(f,t)P_t(p,s,a|f)}. \tag{6}$$

F0 values are selected by putting a threshold on the energy allocated to $\langle p,t\rangle$, i.e. $A_{pt} = P_t(p)\sum_f V(f,t)$. The $F0_{pt}$ with A_{pt} less than the threshold are made null.

III. Source Clustering

For source clustering, the framework for three levels of cognitive clustering, as proposed in our previous work [19], is used:

1) *Pitched event decomposition:* The decomposition of spectra based on multi-F0 values
2) *Group object formation:* The clustering of temporally connected F0s across successive frames
3) *Source streaming:* The clustering of all F0s over all the frames using source-characterizing timbre features and constraints like group objects etc.

While pitched event decomposition is performed using PLCA, as explained in the previous section. A pitched event $\langle p,t\rangle$ corresponds to the estimated $F0_{pt}$. The other stages are detailed in this section as follows.

A. Group Objects formation

Multi-F0 estimation gives a set of F0s corresponding to the active pitched events at each frame. A source label can be assigned to each pitched event individually, based on its timbral features. However, the clustering process can be enhanced by grouping certain pitched events together, based on the temporal continuity of the F0 values. Apart from cognitive reasons, this grouping is justified by physical and musicological constraints as well.

Using F0 as a property of the source signal which varies smoothly over time, we track it in the increasing direction of time in order to form group objects. For tracking, a second order linear prediction filter is used, which gives the predicted F0 value for group g as

$$F0_t^g = F0_{p'(t-1)} + (F0_{p'(t-1)} - F0_{p''(t-2)}) \tag{7}$$

where $\{\langle p',t-1\rangle, \langle p'',t-2\rangle\} \subset G_g$; G_g being the set of pitched events clustered into the same group g. At time t, only one pitched event $\langle p,t\rangle$ is grouped into G_g based on the proximity between $F0_{pt}$ and $F0_t^g$ on a log scale. The implementation details are elaborated in our earlier work [19].

B. Source-Characterizing Features

The magnitude spectrum corresponding to each $F0_{p,t}$ can be estimated from the PLCA components with the help of Wiener filtering,

$$V_p(f,t) = \frac{P_t(f,p)}{P_t(f)}V(f,t) \tag{8}$$
$$\text{where,}\quad P_t(f,p) = \sum_{s,a} P_t(f|p,a)P_t(p)P_t(s|p)P(a|s).$$

Although, the magnitude spectra and dynamic properties of a group object can be used to find a number of acoustic features to characterize the source, the present method simply uses Mel Frequency Cepstral Coefficients (MFCC's). MFCC's have been used successfully for source characterization in both music and speech. They have also been used prominently in unsupervised speaker clustering for speaker diarization [21]. Contrastingly, the MFCC's used here have been derived from only the harmonic part of the spectrum.

The estimated spectrum $V_p(f,t)$ is passed through a filter bank of 50 triangular band pass filters with centres linearly spaced on mel-frequency scale. The discrete cosine transform coefficients of the logarithm of the output are called as the MFCC's. The first N_m coefficients, leaving the dc term, form the source characterizing feature vector $\mathbf{m}_{pt}$ of the pitched event $\langle p,t\rangle$.

C. Hidden Markov Random Fields

To carry out the source clustering, we have two kinds of information. Firstly, we have the source characterizing features of each pitched event $\langle p,t\rangle$. Secondly, we have some physical constraints, i.e., (i) *Grouping constraint*: all the pitched events in a group object should belong to the same source stream, and (ii) *Simultaneity constraint*: there is only one F0 value per source stream at a time. The simultaneity constraint is based on the assumption that a source plays only single pitch stream.

Such structurally constrained clustering of data can be efficiently done using the framework of Hidden Markov Random Fields (HMRF) [33], [34]. An HMRF has two components:

1) The observed field $\mathcal{M}$, which consists of the source characterizing features of all the pitched events $\langle p,t\rangle$.

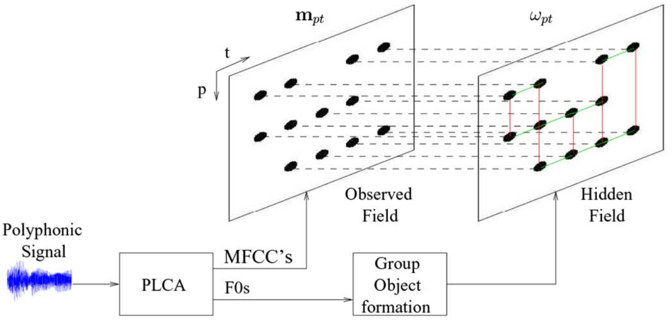

Fig. 1. Overview of proposed source transcription architecture: in the hidden field, green and red links show the grouping and simultaneity constraints, respectively.

2) The hidden field of labels Ω, which is the set of underlying cluster labels $\Omega_{pt} \in \{1, 2, \ldots, N_\omega\}$ of all the pitched events.

The hidden field is constrained by the grouping and simultaneity constraints. An overview of PLCA based HMRF architecture is illustrated in Fig. 1.

Let $\Omega = \{\Omega_{pt} | \forall \{p, t\}\}$, where Ω_{pt} symbolizes the cluster label of each pitched event $\langle p, t\rangle$, and let ω denote one the N_ω values that each Ω_{pt} can take. The clustering task can be seen as *maximum a posteriori* (MAP) estimation, i.e., maximizing the conditional probability of cluster labels Ω, given the observed data $\mathcal{M}$. The final label configuration $\hat{\Omega}$ is given by

$$\hat{\Omega} = \arg\max_{\Omega}\{P(\Omega|\mathcal{M})\} \tag{9}$$
$$= \arg\max_{\Omega}\{P(\Omega)P(\mathcal{M}|\Omega)\} \tag{10}$$

The probability of a label configuration Ω is given by the Gibbs-Boltzmann distribution,

$$P(\Omega) = \frac{1}{Z} e^{-U(\Omega)/T} \tag{11}$$

where Z is a normalization constant, T is temperature parameter and U is the energy function.

$$Z = \sum_{\Omega} e^{-U(\Omega)/T} \tag{12}$$

is called the partition function. T can be interpreted as controlling the sharpness of the peaks of $P(\Omega)$. If T is small, the effect of the difference in energies of different configurations on their respective probabilities is exaggerated. The energy function is given by,

$$U(\Omega) = \sum_{g} V_g(\Omega) + \sum_{t} V_t(\Omega). \tag{13}$$

Here, g, t are the group and time indices, respectively. It can be noted from Eq. (11) that the configuration with smaller energy is more probable. Also, we model these energy functions as pair interactions.

All the pitched events belonging to the same group g form a subgraph, whose energy is given by V_g. All such $\langle p, t\rangle \in G_g$ must be clustered together into the same cluster. The configurations which follow this grouping constraint should get low energy. Hence, we choose the pairwise energy function, which is the sum of interactions between all the pairs in G_g, as

$$V_g(\Omega) = - \sum_{\substack{\{\langle p,t\rangle, \langle p',t'\rangle\} \subset G_g \\ \langle p,t\rangle \neq \langle p',t'\rangle}} \delta(\Omega_{pt} - \Omega_{p't'}). \tag{14}$$

Here, $\delta(\cdot)$ is the Dirac delta function. The above equation ensures that V_g is low for the configuration in which all the pitched events in G_g have the same label.

All the pitched events at the same t form a subgraph, whose energy is given by V_t and they all must be segregated into different clusters. The configurations which obey this simultaneity constraint should get low energy. Hence, the pairwise energy function, which is the sum of interactions between all the pairs of active pitched events at time t, is chosen as

$$V_t(\Omega) = \zeta \sum_{\substack{\{\langle p,t\rangle, \langle p',t\rangle\} \\ p \neq p'}} \delta(\Omega_{pt} - \Omega_{p't}). \tag{15}$$

Here, ζ is the weighing constant. For the simultaneity constraint to be enforced strictly, ζ should have a large value. The aforementioned equation ensures that V_t is high for Ω in which the two pitched events at the same time t belong to the same cluster.

In equation (10), using the conditional independence property of MRF, $P(\mathcal{M}|\Omega)$ can be expressed as

$$P(\mathcal{M}|\Omega) = \prod_{p,t} P(\mathbf{m}_{pt}|\Omega) \tag{16}$$
$$= \prod_{p,t} P(\mathbf{m}_{pt}|\Omega_{pt}). \tag{17}$$

Given the cluster label, the emission probability of observed data can be modeled in an exponential form,

$$P(\mathbf{m}_{pt}|\Omega_{pt} = \omega) = \frac{1}{Z_{pt}} e^{-\|\mathbf{m}_{pt} - \boldsymbol{\mu}_\omega\|^2}. \tag{18}$$

Here, $\|\cdot\|$ denotes the Euclidean distance, Z_{pt} is a normalization constant and $\boldsymbol{\mu}_\omega$ is the cluster representative for cluster ω. This model assumes that each cluster can be characterized by a single point in the feature space and all the feature points belonging to a cluster ω lie closer to $\boldsymbol{\mu}_\omega$ than any $\boldsymbol{\mu}_{\omega' \neq \omega}$.

Using the above derived equations, the term to be maximized in Eq. (10) can be written as

$$P(\Omega)P(\mathcal{M}|\Omega) = \frac{1}{Z} e^{-U(\Omega)/T} \prod_{p,t} \frac{1}{Z_{pt}} e^{-\|\mathbf{m}_{pt} - \boldsymbol{\mu}_{\Omega_{pt}}\|^2}.$$

Maximizing this posterior distribution is equivalent to maximizing its log,

$$\hat{\Omega} = \arg\max_{\Omega}\{\mathcal{L}_{\mathrm{HMRF}}(\mathcal{M}, \Omega)\}$$
$$\mathcal{L}_{\mathrm{HMRF}}(\mathcal{M}, \Omega) = -\sum_{p,t} \|\mathbf{m}_{pt} - \boldsymbol{\mu}_{\Omega_{pt}}\|^2 - \frac{1}{T}\sum_{g} V_g(\Omega) - \frac{1}{T}\sum_{t} V_t(\Omega). \tag{19}$$

D. HMRF Parameter Estimation

1) Updating Ω *:* We propose the following two schemes for updating Ω.

Scheme 1: For maximizing the objective function $\mathcal{L}_{\text{HMRF}}(\mathcal{M},\Omega)$ with respect to (w.r.t.) the cluster labels Ω, we use an iterative method known as *iterated conditional modes* (ICM), which is a greedy coordinate ascent technique. After suitably initializing Ω, the pitched events are selected one by one. Given the cluster labels of other pitched events $\Omega_{\backslash pt} = \{\Omega_{p't'} : p' \neq p, t' \neq t\}$ and the feature data $\mathcal{M}$, the objective function is calculated at the selected $\langle p,t\rangle$ for different values of the label Ω_{pt},

$$\begin{aligned}\mathcal{L}_{\text{HMRF}}(\Omega_{pt}|\mathcal{M},\Omega_{\backslash pt}) = &-\|\mathbf{m}_{pt} - \boldsymbol{\mu}_{\Omega_{pt}}\|^2 \\ &+ \sum_{\substack{\{\langle p,t\rangle,\langle p',t'\rangle\}\subset G_g \\ \langle p,t\rangle\neq\langle p',t'\rangle}} \frac{\delta(\Omega_{pt}-\Omega_{p't'})}{T} \\ &- \zeta \sum_{\substack{\{\langle p,t\rangle,\langle p',t\rangle\} \\ p\neq p'}} \frac{\delta(\Omega_{pt}-\Omega_{p't})}{T}.\end{aligned}\tag{20}$$

The label with maximum value of objective function is used to estimate the new value of Ω_{pt},

$$\hat{\Omega}_{pt} = \arg\max_{\Omega_{pt}}\{\mathcal{L}_{\text{HMRF}}(\Omega_{pt}|\mathcal{M},\Omega_{\backslash pt})\}.\tag{21}$$

Generally, ICM involves randomly choosing the order of pitched events. However, the proposed schemes involve traversing successively through all the pitched events once and finally updating $\Omega \leftarrow \hat{\Omega}$. This forms one iteration of ICM.

During successive iterations of ICM, the temperature T is gradually lowered. The objective function in Eq. (21) has three terms - closeness to a cluster, grouping constraint and simultaneity constraint. Initially, when T is high, the grouping and simultaneity constraint terms are suppressed so that the clustering can take place freely. However, at later iterations, more than the clustering term, the constraint terms are desirable for smoothing out the clusters as per the constraints. Hence, at low T, the constraint terms contribute largely to the objective function.

Scheme 2: The scheme 1 is modified by explicitly imposing the simultaneity constraint. This makes the simultaneity constraint play an important role in all the iterations of ICM. The objective function of Eq. (19) is modified by removing the term corresponding to simultaneity constraint. In each ICM iteration, the objective function is computed for all pitched events at each t,

$$\begin{aligned}\mathcal{L}_{\text{HMRF}}(\Omega_{pt}|\mathcal{M},\Omega_{\backslash pt}) = &-\|\mathbf{m}_{pt}-\boldsymbol{\mu}_{\Omega_{pt}}\|^2 \\ &+ \sum_{\substack{\{\langle p,t\rangle,\langle p',t'\rangle\}\subset G_g \\ \langle p,t\rangle\neq\langle p',t'\rangle}} \frac{\delta(\Omega_{pt}-\Omega_{p't'})}{T}\end{aligned}\tag{22}$$

Based on the maximum value of $\mathcal{L}_{\text{HMRF}}(\Omega_{pt}|\mathcal{M},\Omega_{\backslash pt})$, the cluster labels and the pitched events at t are linked injectively in a greedy fashion to estimate $\hat{\Omega}_{pt}, \forall p$ at the given t. This scheme explicitly ensures that no two pitched events at the same t are given the same label.

2) Updating $\boldsymbol{\mu}_\omega$: The cluster representatives for each source can easily be updated as

$$\hat{\boldsymbol{\mu}}_\omega = \arg\max_{\boldsymbol{\mu}_\omega}\{\mathcal{L}_{\text{HMRF}}(\mathcal{M},\Omega)\}\tag{23}$$

$$= \frac{\sum_{p,t}\mathbf{m}_{pt}\delta(\Omega_{pt}-\omega)}{\sum_{p,t}\delta(\Omega_{pt}-\omega)}.\tag{24}$$

3) Agglomerative clustering: The clustering using HMRFs is quite dependent on initialization. Hence, if we cluster the group objects into the actual number of sources at the HMRF clustering stage itself, then there is a higher probability of getting stuck in a local maximum. To circumvent this problem, we use HMRF to give *more* number of clusters N_ω than the true number of sources N_s, and thence use hierarchical agglomerative clustering to obtain the final source streams.

Each cluster ω obtained from the HMRF based clustering is characterized by its centroid $\boldsymbol{\mu}_\omega$. At each iteration, the pair of clusters to be merged is chosen based on a cost function $\mathcal{C}$ which consists of (i) Euclidean distance between the cluster centroids, and (ii) simultaneity constraint, as follows

$$\begin{aligned}\mathcal{C}(\omega_1,\omega_2) = &\|\boldsymbol{\mu}_{\omega_1}-\boldsymbol{\mu}_{\omega_2}\|^2 \\ &+\sum_t\left(\sum_p\delta(\Omega_{pt}-\omega_1)\right)\left(\sum_p\delta(\Omega_{pt}-\omega_2)\right)\end{aligned}\tag{25}$$

The simultaneity constraint prevents a cluster ω_1, which has a pitched event at time t (i.e., $\delta(\Omega_{pt}-\omega_1)=1$ for some $\langle p,t\rangle$), to be merged with another cluster ω_2, containing any pitched event at the same t (i.e., $\delta(\Omega_{pt}-\omega_2)=1$ for some $\langle p,t\rangle$).

The agglomerative clustering can be carried out as follows:

1) Compute the cost function $\mathcal{C}$ between each pair of clusters;
2) Merge the two clusters with minimum cost to form an agglomerated cluster, by updating Ω and their centroid;
3) Repeat the above two steps till the total number of clusters equals the actual number of sources N_s.

IV. Implementation Strategy

A. Utilizing the Precision-Recall Trade-off

Unsupervised source transcription involves estimating the F0s and then clustering them into source streams. Hence, the problem is addressed in two stages - detection and classification. Errors in the first stage cause severe deterioration at the next stage, thereby degrading the overall performance. These errors at multi-F0 estimation stage are due to (i) spurious F0s, which form bad clusters later, and (ii) missed true F0s, which cannot participate in clustering.

In this paper, we devise a novel two-step strategy to address this issue. For the multi-F0 estimation task, there is a trade-off between precision and recall. Precision and recall can be defined as

$$\text{Prec} = 1 - \frac{\text{no. of spurious non-zero F0s estimated}}{\text{no. of all estimated non-zero F0s}}\tag{26}$$

$$\text{Rec} = \frac{\text{no. of correctly estimated non-zero F0s}}{\text{no. of all non-zero F0s in ground truth}}.\tag{27}$$

When there is high precision, the number of spurious F0s is low, even if all the correct F0s are not detected. On the other hand, when there is high recall, the number of correctly detected F0s

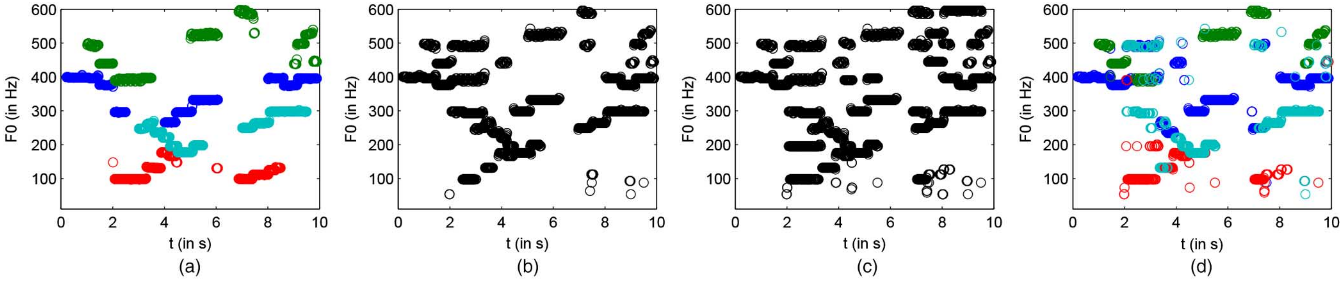

Fig. 2. Results for a polyphony from BACH dataset with $N_s = 4$: (a) Ground truth pitch contours, with different colors for different sources; (b) Estimated F0s in $\mathbf{F0}^{\mathrm{P}}$; (c) Estimated F0s in $\mathbf{F0}^{\mathrm{R}}$; (d) Estimated source streams, with different colors for different sources. For this example, Recall $= 73.7\%$, Precision $= 71.1\%$, Accuracy $= 56.7\%$ for source transcription.

is large, but the number of spurious F0s may also be large. The proposed strategy is based on the following observations:

•) The parameters (or thresholds) of the multi-F0 estimator can generally be adjusted so as to give either high precision or high recall.

•) For partitioning the feature space into source-dominant regions, precision is more desirable than recall, as spurious F0s can misrepresent the regions.

•) F0s from the high recall set can be added to the clusters based on their position in the partitioned feature space.

In the first stage, the set $\mathbf{F0}^{\mathrm{P}}$, with high precision, is used for unsupervised source clustering. The obtained clusters specify the region corresponding to each source in the space of source characterizing features. Subsequently, in the second stage, we expand the source clusters with the help of extra F0s which are there in the set $\mathbf{F0}^{\mathrm{R}}$, with high recall, but not in $\mathbf{F0}^{\mathrm{P}}$. Based on the position of such F0s in the feature space, they are linked to the source clusters. The spurious F0s can thus be spurned, thereby improving the overall transcription accuracy.

B. Complete Implementation

1) Multi-F0 Extraction: A fixed number (10 here) of spectral peaks, lying within the range $F_{\min}$ (50 Hz) to $F_{\max}$ (1400 Hz) and with harmonic score greater than a threshold ($= 0.01$), are chosen as F0 candidates at each t. Initializing all the PLCA parameters from uniform priors, except $P(a|s)$ that is randomly initialized, PLCA decomposition is performed using the EM algorithm for a fixed number (10 here) of iterations. By adjusting the threshold on A_{pt}, we obtain two sets of multi-F0 values. A larger value of threshold, ϵ_{P} ($= 0.3$ here), gives $\mathbf{F0}^{\mathrm{P}}$, a set of F0s with high precision but less recall. Correspondingly, a smaller value, ϵ_{R} ($= 0.2$ here), gives $\mathbf{F0}^{\mathrm{R}}$, a set with high recall but low precision. While estimating the high precision F0s set, it should be made sure that the recall is not too low as it may miss or under-represent some source in the given test audio.

2) Group Objects Formation: The group objects are formed for each set of F0s. For characterizing each $\langle p, t\rangle$, the number of MFCC's, N_m, is set to 5.

3) Source Clustering: The high precision set $\mathbf{F0}^{\mathrm{P}}$ is clustered into $N_\omega (> N_s)$ clusters using HMRF based clustering. For this, Ω is initialized randomly, with Ω_{pt} being sampled uniformly from $\{1, 2, \ldots, N_\omega\}, \forall \langle p, t\rangle$. For this work, we have taken $N_\omega = 2N_s$. Thence, $\langle p, t\rangle$'s are clustered into source clusters iteratively, using ICM. To strictly enforce the simultaneity constraint, ζ is set as 500. The temperature parameter T is slowly decreased along the exponential curve $T = e^{\tau}$, where τ decreases linearly in steps of 0.3, starting from 5. Slower cooling rates give better results but with increased computations. The maximum number of ICM iterations is set to 40. Based on the two schemes of updating Ω, we get two variations of the proposed algorithm, named as **Proposed-1** and **Proposed-2**. The obtained clusters are further agglomerated to reduce their number from N_ω to N_s.

4) Using the High Recall F0s: The obtained clusters are expanded with the help of extra F0s, which are present in the set $\mathbf{F0}^{\mathrm{R}}$ but not in $\mathbf{F0}^{\mathrm{P}}$. At each t, the unused source labels and the unused pitched events are linked injectively by maximizing the cosine similarity $\mathcal{D}_{\cos}(\boldsymbol{\mu}_s, \mathbf{m}_{pt})$ between the pair s and $\langle p, t\rangle$. The pairs with cosine similarity greater than a threshold $\epsilon_{\cos}$ (=0.8 here) are linked greedily, starting from the pair with maximum similarity. Here, cosine similarity measure has been used instead of Euclidean distance for convenience in defining a threshold on closeness.

5) Final Source Streaming: In the final source streams, there is a single F0 value at each t per stream. However, it is possible that in Ω, some source s is assigned to multiple $\langle p, t\rangle$'s at same t. In such a case, s is assigned to the one whose $\mathbf{m}_{pt}$ lies closer to $\boldsymbol{\mu}_s$ in terms of Euclidean distance. Since, annealing ends up in strict enforcement of the grouping constraint and agglomeration cannot disturb the grouping constraint, this stream assignment is done using only the timbre information and simultaneity constraint. The other $\langle p, t\rangle$'s are given any stream s', which has still not got any $\langle p, t\rangle$, based on their closeness to $\boldsymbol{\mu}_{s'}$. However, if no stream is available, then such $\langle p, t\rangle$'s are discarded.

An example visualization of the different stages of proposed-2 source transcription algorithm has been illustrated in Fig. 2. Please note the absence of many true F0s from $\mathbf{F0}^{\mathrm{P}}$ in Fig. 2(b) and presence of many spurious F0s in $\mathbf{F0}^{\mathrm{R}}$ in Fig. 2(c).

V. Experiments

Evaluation Results for Multi-F0 Estimation

A. Datasets

The proposed unsupervised source transcription algorithm has been evaluated over three datasets.

The first dataset BACH contains 10 pieces of J. S. Bach chorales performed using violin, clarinet, saxophone and bassoon. It has been derived from publicly available Bach10

TABLE I
EVALUATION RESULTS FOR MULTI-F0 ESTIMATION

Dataset	N_s	DuanMPS			Proposed Algorithm, $\mathbf{F0}^{\mathrm{P}}$			Proposed Algorithm, $\mathbf{F0}^{\mathrm{R}}$		
		Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
BACH	3	90.8 (6.1)	69.0 (2.5)	78.4 (3.9)	85.7 (8.1)	73.0 (5.7)	78.8 (6.8)	92.6 (4.2)	63.1 (7.2)	75.1 (6.5)
	4	90.1 (4.8)	58.9 (2.0)	71.2 (3.0)	79.5 (7.6)	80.4 (4.6)	80.0 (6.1)	89.2 (3.9)	68.1 (6.9)	77.2 (5.9)
MIREX	3	71.0 (3.4)	57.1 (2.3)	63.3 (2.8)	74.9 (3.5)	48.9 (0.6)	59.2 (1.5)	81.3 (3.0)	39.9 (2.3)	53.5 (2.7)
	4	71.1 (1.6)	54.8 (1.3)	61.9 (1.4)	62.5 (1.2)	53.7 (2.2)	57.8 (1.8)	77.6 (0.4)	48.5 (2.4)	59.7 (1.9)
IMIDI	3	73.5 (8.1)	49.4 (8.0)	59.1 (8.3)	75.7 (11.5)	55.7 (9.0)	64.2 (10.1)	84.9 (10.5)	49.2 (11.4)	62.3 (12.0)
	4	78.1 (6.3)	46.2 (6.3)	58.1 (6.7)	66.0 (8.5)	60.6 (9.5)	63.2 (9.1)	80.0 (9.4)	55.0 (8.8)	65.2 (9.3)

dataset [8]. Bach10 dataset has each piece 30s long, of which only first 10s audio has been used in BACH.

The second dataset is derived from the MIREX Multi-F0 development set. It consists of 2 audio files of a Beethoven's composition performed using bassoon, flute, oboe and horn. Each file is 27s long.

The third dataset IMIDI comprises of 10 songs derived from the Strings database [35]. The original midi files are converted into wav files using instruments like flute, violin, trumpet and cello. Each file is 10s in duration.

The sampling rate for all the audio files is 44.1 kHz. The corresponding pitch contours for the individual instruments in each audio file are available at 10 ms hop size.

These datasets generally contain continuous source streams with all the sources playing simultaneously almost all the time. This makes the task somewhat easier because the simultaneity constraint can help in distinguishing between different sources. However, a more practical situation is when the number of simultaneously active sources keeps changing. The task becomes even more difficult when there is no such time when *all* the source are active simultaneously. In order to incorporate these challenges into our evaluation, we suppress each source stream at certain times. We apply temporal mask on the waveform of each source using point-wise multiplication. The mask is set as 1 for first 2.5s and 0 for next 1.5s. This pattern of 4s is repeated throughout the length of the waveform. Moreover, the masks of different source waveforms are staggered by shifting them by integer multiples of 1s w.r.t. one another. Such masks have been applied over all the datasets for evaluations.

B. Existing State-of-the-Art Methods for Comparison

The proposed algorithm is compared with another state-of-the-art algorithm called as Multi-Pitch Streaming algorithm proposed by Duan *et al.* [20]. Their system, referred to as DuanMPS henceforth, consists of two parts. The first part performs multi-F0 estimation [8]. The second part clusters the obtained F0s in a constrained way into different source streams. Their MATLAB implementation is available as open source (from http://www.ece.rochester.edu/~zduan/). We use their code as it is. The maximum number of F0 streams, however, is set equal to N_s for evaluation purpose.

C. Evaluation Metrics

The overall transcription system finally provides N_s source streams, each of which consists of a single channel temporal F0 contour corresponding to a particular instrument. For evaluation, each estimated source stream is linked one-to-one with a ground truth source stream. The linking is done so as to maximize the accuracy of transcription. The recall, precision and accuracy of transcription are defined as [20]:

$$\mathrm{Rec} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \mathrm{Prec} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \mathrm{Acc} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

Here, TP is the number of correct non-zero F0 values, FP is the number of incorrect non-zero F0 values, and FN is the number of incorrect (zero or non-zero) F0 values which are actually non-zero in the ground truth source stream. A streamed F0 value is compared with the corresponding F0 value in the linked ground truth stream at the same time frame, and is considered as correct if the two lie less than half a semitone apart. Thus, to be considered as TP, an estimated F0 should have correct value as well as correct source label.

In addition to this, the sub task of multi-F0 estimation has also been evaluated using the measures of precision and recall as defined in Eqs. (26), (27). For this task, however, an estimated non-zero F0 is considered as correct if it is less than half a semitone away from any ground truth F0 at that time frame. Thus, to be considered correct, an estimated F0 should have correct value only. F1-measure is defined as the harmonic mean of precision and recall. All the metrics are reported in %ge values throughout this paper.

D. Evaluation Results and Discussion

Multi-F0 Estimation: The results of only multiple F0 estimation are presented in Table I. While DuanMPS system uses only one set of F0s, the proposed system uses two sets - $\mathbf{F0}^{\mathrm{P}}$ and $\mathbf{F0}^{\mathrm{R}}$. $\mathbf{F0}^{\mathrm{P}}$ shows higher precision than $\mathbf{F0}^{\mathrm{R}}$ while the later has higher recall than the former. As compared to DuanMPS, the proposed system shows similar performance in terms of the F1 values. For IMIDI and BACH ($N_s = 4$), the proposed system has higher F1 value than that of DuanMPS, but in other cases, DuanMPS performs better.

Overall Transcription: The overall transcription results obtained from different algorithms over different datasets are illustrated in Fig. 3. Both the proposed systems perform similarly, but proposed-2 performs better than the other for MIREX dataset. However, both the proposed algorithms perform better than DuanMPS over all the datasets. It should be noted that even if there is not substantial difference in the performance of multi-F0 estimators of DuanMPS and the proposed systems, the source transcription performance of the proposed ones outperforms that of the former. This shows that the proposed systems are more robust to the presence of spurious F0s.

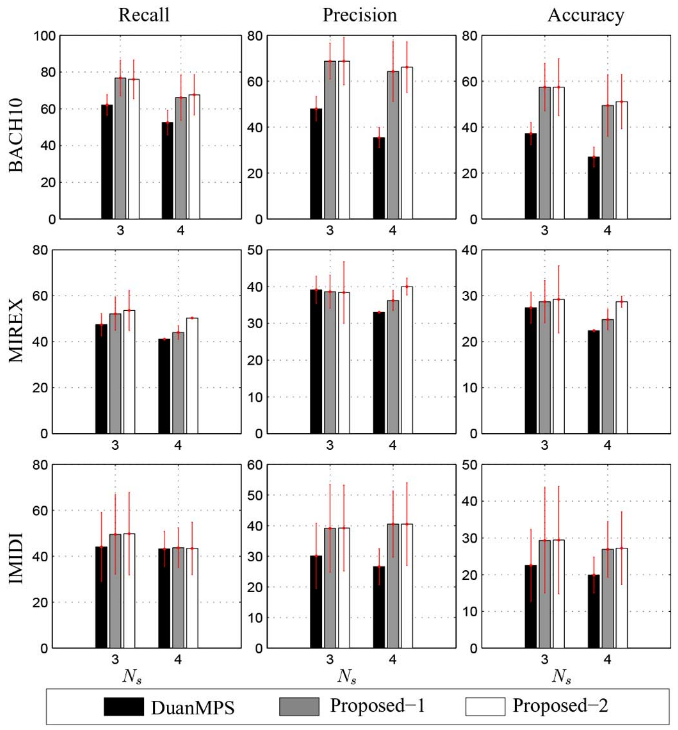

Fig. 3. Evaluation Results for overall transcription obtained from DuanMPS and the proposed schemes. Figure columns correspond to Rec, Prec and Acc, respectively, all in %ge. Rows correspond to the different datasets. The error bars show the standard deviation. In each figure, x-axis represents the number of sources N_s.

TABLE II
EVALUATION RESULTS FOR SOURCE TRANSCRIPTION WITH DIFFERENT F0 ESTIMATORS (F1 IN BRACKETS) OVER BACH DATASET, $N_s = 4$

F0 estimator	Algoritm	Recall	Precision	Accuracy
True (100.0)	DuanMPS	85.0 (6.9)	85.3 (7.0)	74.7 (10.6)
	Proposed-1	84.7 (6.6)	86.1 (6.3)	75.1 (9.9)
	Proposed-2	90.8 (5.0)	91.5 (5.0)	83.1 (8.3)
Duan (71.2)	DuanMPS	50.4 (7.6)	34.2 (5.6)	25.8 (5.0)
	Proposed-1	69.8 (9.8)	82.3 (7.7)	61.3 (10.6)
	Proposed-2	65.4 (12.1)	79.0 (10.5)	56.7 (13.5)
Vincent (69.5)	DuanMPS	46.9 (9.9)	34.8 (9.9)	25.4 (7.9)
	Proposed-1	51.2 (6.6)	62.2 (7.2)	39.2 (6.2)
	Proposed-2	51.7 (6.3)	59.4 (8.9)	38.2 (5.9)
Proposed (42.5)	DuanMPS	47.4 (5.2)	29.1 (3.4)	22.1 (3.1)
	Proposed-1	55.3 (15.3)	64.9 (13.4)	43.9 (14.1)
	Proposed-2	57.3 (14.5)	68.4 (11.8)	46.5 (13.8)

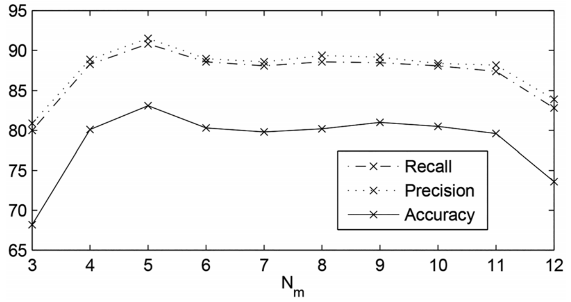

Fig. 4. Effect of varying N_m on the source transcription performance of proposed-2 algorithm over the BACH database for $N_s = 4$. The y-axis represents values in %ge.

The time complexity of each ICM iteration is of the order of $O(N_{pt}N_\omega)$ for scheme-1 and $O(N_{pt}N_\omega^2)$ for scheme-2. Here, N_{pt} is the total number of active F0 values. Since the search space is finite and HMRF objective function (for fixed T) always increases in each greedy update, the convergence of ICM method is guaranteed with random traversal of pitched events when T changes slowly [33]. However, successive traversal may lead to small oscillations because the centers and labels are updated only after completing each traversal. In practice, the convergence for scheme-1 is achieved in around 30 iterations, while scheme-2 mostly ends up in small oscillations.

Source Clustering with Different F0 Estimators: Experiments have been conducted to further assess the source clustering performance of the proposed systems, over the same F0 values given as input. Source clustering has been carried out with the given true F0 values as well as with the F0 values extracted using different multi-F0 estimators, viz., Duan [8], Vincent [9] and the proposed one. DuanMPS directly accepts the given F0 values, but the proposed systems have been provided the given F0s as F0 candidates so as to create the sets $\mathbf{F0}^{\mathrm{P}}$ and $\mathbf{F0}^{\mathrm{R}}$ using thresholds.

The evaluation results for these experiments for BACH dataset with $N_s = 4$ are tabulated in Table II. All the systems perform best over the true F0s as compared to the estimated F0s, which have estimation errors that propagate to adversely affect the clustering performance. Over true F0s, the proposed systems give better accuracy than DuanMPS, showing the efficacy of the proposed source clustering algorithms. All the F0 candidates of the proposed system (Section II-A) have been taken as estimated F0s here, and hence, the proposed F0 estimator has a low F1-measure. Nevertheless, with all the F0 estimators, the proposed clustering systems perform much better than DuanMPS algorithm. This shows the enhanced ability of the proposed systems to deal with estimation errors.

Effect of Agglomeration and Final Source Streaming: Underclustering followed by agglomeration helps in preventing local maxima. The advantage of this strategy is substantiated by the reduction in accuracy when it is not used. If the HMRF clustering directly results in N_s clusters, the overall accuracy for MIREX dataset, with $N_s = 4$, declines to 22.8%, while it is 28.7% for the complete system.

The final source streaming reinforces the simultaneity constraint using timbre information. If the streaming is done by invalidating all the different F0 values that have been assigned to the same source at the same t, the overall accuracy for MIREX dataset, with $N_s = 4$, drops to 25.5%. The drop is small because of the strict enforcement of simultaneity constraint at various preceding stages in source clustering.

Effect of varying N_m: An important factor for the source transcription performance is the vector of source characterizing features. The proposed systems use $N_m = 5$ for the above evaluations. To analyze the effect of changing the number of MFCC's, experiments have been performed over the BACH dataset. To keep the set of F0s same, the true F0 values are supplied as input to the source transcription system. The results of such experiments for $N_s = 4$ using the proposed-2 algorithm have been depicted in Fig. 4. These results appear surprising as only 5 MFCC values have been used here as source characterizing features, while commonly used values of N_m vary around 13 [36] to 21 [20].

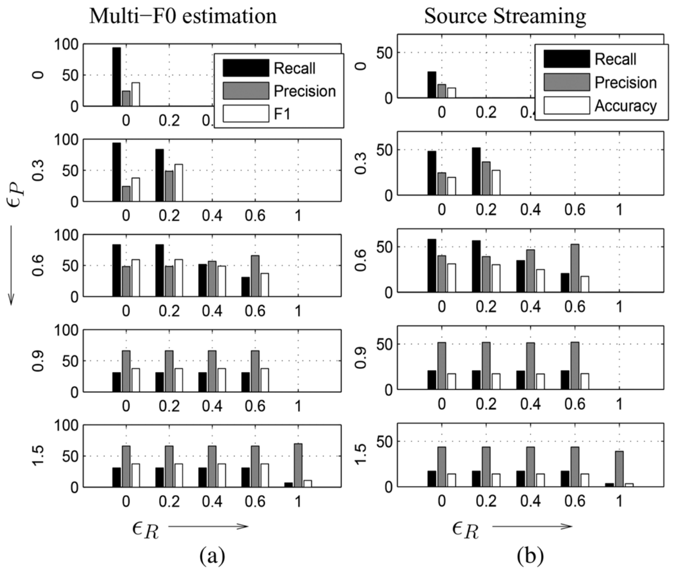

Fig. 5. Effect of varying ϵ_{P} and ϵ_{R} on the performance of (a) Multi-F0 estimation (b) overall Source transcription.

Effect of varying ϵ_{P} *and* ϵ_{R} *:* There are certain applications that require more recall in transcription or streaming, while others require more precision. One of the advantages of the proposed system is that these measures can be varied by tuning the parameters. The key parameters which directly affect the precision and recall of multi-F0s estimation are ϵ_{P} and ϵ_{R}. Another parameter, which decides that how many extra F0s from the set $\mathbf{F0}^{\mathrm{R}}\backslash\mathbf{F0}^{\mathrm{P}}$ are to be assimilated while expanding the source clusters, is $\epsilon_{\cos}$. For the above implementation, these values have been optimized over the BACH dataset. In the following experiments, their influence on performance is analyzed, using the MIREX dataset.

The effect of varying ϵ_{P} and ϵ_{R} on the performance of multi-F0 estimation (with $\mathbf{F0}^{\mathrm{R}}$) and overall source transcription is shown in Fig. 5. These experiments have been performed over MIREX dataset with $N_s = 4$. The role of $\epsilon_{\cos}$ has been obviated by setting it equal to -1. Since, $\epsilon_{\mathrm{R}} \leq \epsilon_{\mathrm{P}}$, some of the bar-plots are missing. In the multi-F0 estimation task, for a given ϵ_{P}, decrease in ϵ_{R} increases the number of F0s in $\mathbf{F0}^{\mathrm{R}}$ and hence the recall increases; but as a compromise, the precision decreases as a more number of spurious F0s also gets selected. Similar effects are seen in the corresponding source transcription performance. On the other axis, increase in ϵ_{P} decreases the number of F0s in $\mathbf{F0}^{\mathrm{P}}$ thereby increasing the precision for both the tasks. However, it seems that the clusters obtained from less data are not able to represent the sources well enough to include all the true pitched events into their respective true source clusters, and hence, the recall decreases. The best accuracy for source transcription over MIREX dataset is obtained around $\epsilon_{\mathrm{R}} \in [0, 0.2], \epsilon_{\mathrm{P}} \in [0.3, 0.6]$. However, it has been observed that $\epsilon_{\mathrm{R}} = 0.2, \epsilon_{\mathrm{P}} = 0.3$ gives good performance in general over all the datasets, while other values seem to overfit the MIREX dataset.

These experiments also illustrate the effect of using only single set of pitch estimates. For the cases with $\epsilon_{\mathrm{P}} = \epsilon_{\mathrm{R}}$, only one set of F0s is available. The effect of varying the common threshold value can be seen by interpolating/extrapolating the graph of Fig. 5. For example, at $\epsilon_{\mathrm{P}} = 0.6$, setting $\epsilon_{\mathrm{R}} = 0.6$ gives poorer transcription accuracy as compared to that by setting $\epsilon_{\mathrm{R}} = 0.2$. This substantiates the benefit of using two sets of F0 estimates with the proposed precision-recall based strategy.

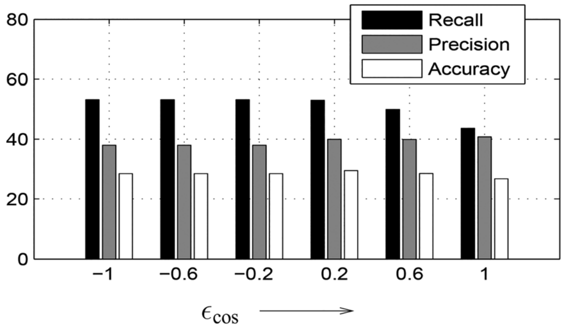

Fig. 6. Effect of varying $\epsilon_{\cos}$ on the performance of overall source transcription over MIREX dataset with $N_s = 4$.

Effect of varying $\epsilon_{\cos}$ *:* Keeping $\epsilon_{\mathrm{R}} = 0.2, \epsilon_{\mathrm{P}} = 0.3$, now the effect of $\epsilon_{\cos}$ on the source transcription performance is examined over the same data in Fig. 6. Increasing $\epsilon_{\cos}$ allows less number of extra F0s $\in \mathbf{F0}^{\mathrm{R}}\backslash\mathbf{F0}^{\mathrm{P}}$ to be assimilated into the clusters and hence the recall decreases. On the other hand, it also prevents the spurious F0s from getting into the clusters, thereby, increasing the precision. However, the accuracy level is very little affected by change in $\epsilon_{\cos}$.

VI. Conclusion

This paper presents a novel unsupervised procedure for source transcription of polyphonic music. The multiple F0 values are extracted using source-filter model based PLCA. These F0s are classified into streams corresponding to different sources using acoustic features along with cognitive grouping and simultaneity constraints. A novel framework based on HMRF's is proposed for this constrained clustering that uses annealing to vary the degree of constraints. It also includes the tactic of under-clustering followed by agglomerative clustering to avoid the problem of local maxima. Further, a novel strategy is proposed for unsupervised clustering based on the trade-off between precision and recall of F0 values to be clustered. It enhances the robustness of the system to estimation errors. Experimental evaluations show that the proposed algorithm outperforms a state-of-the-art algorithm in source clustering, while performs comparable to the state-of-the-art algorithms in terms of multi-F0 estimation.

The proposed system can be developed into a number of applications like source separation, music transcription etc. It would also be interesting to extend the system for speech applications.

References

[1] A. Klapuri and M. Davy, *Signal Processing Methods for Music Transcription*. New York, NY, USA: Springer-Verlag, 2006.

[2] S. Ewert, B. Pardo, M. Mueller, and M. Plumbley, "Score-informed source separation for musical audio recordings: An overview," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 116–124, May 2014.

[3] P. Leveau, E. Vincent, G. Richard, and L. Daudet, "Instrument-specific harmonic atoms for mid-level music representation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 116–128, Jan. 2008.

[4] V. Arora and L. Behera, "On-line melody extraction from polyphonic audio using harmonic cluster tracking," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 3, pp. 520–530, Mar. 2013.

[5] M. Marolt, "A mid-level representation for melody-based retrieval in audio collections," *IEEE Trans. Multimedia*, vol. 10, no. 8, pp. 1617–1625, Dec. 2008.

[6] C. Dittmar, E. Cano, J. Abeßer, and S. Grollmisch, "Music information retrieval meets music education," in *Multimodal Music Process.*, M. Müller, M. Goto, and M. Schedl, Eds. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2012, pp. 95–120, volume 3 of Dagstuhl Follow-Ups.

[7] P. Mowlaee, R. Saeidi, M. G. Christensen, Z.-H. Tan, T. Kinnunen, P. Franti, and S. H. Jensen, "A joint approach for single-channel speaker identification and speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 9, pp. 2586–2601, Nov. 2012.

[8] 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.

[9] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 528–537, Mar. 2010.

[10] E. Benetos and S. Dixon, "Multiple-instrument polyphonic music transcription using a temporally constrained shift-invariant model," *J. Acoust. Soc. Amer.*, vol. 133, no. 3, pp. 1727–1741, 2013.

[11] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Commun.*, vol. 52, no. 1, pp. 12–40, 2010.

[12] T. Zhang, "Automatic singer identification," in *Proc. Int. Conf. Multimedia Expo (ICME)*, Jul. 2003, vol. 1, p. I–633–6, vol.1.

[13] 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.

[14] J. Wu, E. Vincent, S. A. Raczynski, T. Nishimoto, N. Ono, and S. Sagayama, "Polyphonic pitch estimation and instrument identification by joint modeling of sustained and attack sounds," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1124–1132, Dec. 2011.

[15] 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, Dec. 2011.

[16] V. Arora and L. Behera, "Discriminative PLCA based polyphonic source identification," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2013.

[17] D. Giannoulis and A. Klapuri, "Musical instrument recognition in polyphonic audio using missing feature approach," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 9, pp. 1805–1817, Sep. 2013.

[18] V. Arora and L. Behera, "Instrument identification using PLCA over stretched manifolds," in *Proc. Nat. Conf. Commun. (NCC)*, Feb. 2014, pp. 1–5.

[19] 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.

[20] Z. Duan, J. Han, and B. Pardo, "Multi-pitch streaming of harmonic sound mixtures," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 1, pp. 138–150, Jan. 2014.

[21] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 356–370, Feb. 2012.

[22] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1557–1565, Sep. 2006.

[23] M. Spiertz and V. Gnann, "Source-filter based clustering for monaural blind source separation," in *Proc. Int. Conf. Digital Audio Effects*, 2009.

[24] 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.

[25] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[26] Y. Hu and G. Liu, "Instrument identification and pitch estimation in multi-timbre polyphonic musical signals based on probabilistic mixture model decomposition," *J. Intell. Inf. Syst.*, vol. 40, no. 1, pp. 141–158, 2013.

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

[28] N. J. Bryan and G. J. Mysore, "Interactive refinement of supervised and semi-supervised sound source separation estimates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2013, pp. 883–887.

[29] A. Diment, T. Heittola, and T. Virtanen, "Semi-supervised learning for musical instrument recognition," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2013, pp. 1–5.

[30] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm," *IEEE Trans. Med. Imag.*, vol. 20, no. 1, pp. 45–57, Jan. 2001.

[31] M. Ranzato, V. Mnih, J. M. Susskind, and G. E. Hinton, "Modeling natural images using gated MRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 9, pp. 2206–2222, Sep. 2013.

[32] M. Shashanka, "Latent variable framework for modeling and separating single channel acoustic sources," Ph.D. dissertation, Boston Univ., Boston, MA, USA, 2007.

[33] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 6, pp. 721–741, Nov. 1984.

[34] S. Basu, M. Bilenko, and R. J. Mooney, "A probabilistic framework for semi-supervised clustering," in *Proc. ACM SIGKDD Int. Conf. Knowl. Disc. Data Mining*, 2004, pp. 59–68, ACM.

[35] 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.

[36] H. Terasawa, M. Slaney, and J. Berger, "The thirteen colors of timbre," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, Oct. 2005, pp. 323–326.

Vipul Arora received the B.Tech. degree in electrical engineering from the Indian Institute of Technology (IIT), Kanpur, India. Currently, he is working towards the Ph.D. degree at IIT Kanpur. His research interests include music information retrieval, acoustic space modeling and semantic signal processing.

Laxmidhar Behera (S'92–M'03–SM'03) received the B.Sc. (engineering) and M.Sc. (engineering) degrees from NIT Rourkela in 1988 and 1990, respectively. He received the Ph.D. degree from IIT Delhi. He has worked as an Assistant Professor at BITS Pilani during 1995–1999 and pursued the postdoctoral studies in the German National Research Center for Information Technology, GMD, Sank Augustin, Germany, during 2000–2001. He is currently working as a Professor in the Department of Electrical Engineering, IIT Kanpur and as a Guest Professor at Hangzhou Dianzi University, China. He was a Reader in the Intelligent Systems Research Center (ISRC), University of Ulster, United Kingdom, on sabbatical from IIT Kanpur during 2007–2009. He has also worked as a Visiting Researcher/Professor at FHG, Germany, and ETH, Zurich, Switzerland. He has more than 180 papers to his credit published in refereed journals and presented in conference proceedings. His research interests include intelligent control, robotics, information retrieval, neural networks, cyber physical systems and cognitive modeling.