SEMI-SUPERVISED POLYPHONIC SOURCE IDENTIFICATION USING PLCA BASED GRAPH CLUSTERING

Vipul Arora, Laxmidhar Behera
Department of Electrical Engineering, Indian Institute of Technology, Kanpur
vipular@iitk.ac.in, lbehera@iitk.ac.in

ABSTRACT

For identifying instruments or singers in the polyphonic audio, supervised probabilistic latent component analysis (PLCA) is a popular tool. But in many cases individual source audio is not available for training. To address this problem, this paper proposes a novel scheme using semi-supervised PLCA with probabilistic graph clustering, which does not require individual sources for training. The PLCA is based on source-filter approach which models the spectral envelope as a weighted sum of elementary band-pass filters. The novel graph based approach, embedded in the PLCA framework, takes into account various perceptual cues for characterizing a source. These cues include temporal cues like the evolution of F0 contours as well as the acoustic cues like mel-frequency cepstral coefficients. The proposed scheme shows better results in identifying vocal sources than a state of the art unsupervised scheme. In addition, the proposed framework can be used to incorporate perceptual cues so as to enhance the performance of supervised schemes too.

1. INTRODUCTION

We humans can selectively focus our attention on listening to a particular sound source even in the midst of many interfering sounds. This ability makes it possible for us to listen to the polyphonic music where we can intently hear a particular instrument (or singer) as if it were playing alone. In order to understand the nature of human cognition it is important to find out what features human ears identify in order to be familiar with and cognize a sound. In music, several perceptual attributes have been found to correspond to particular mathematical properties of the audio signal. For example, pitch approximately corresponds to the fundamental frequency (F0), loudness roughly corresponds to the amplitude/intensity etc. However, even if the pitch and loudness may vary during the song, we tend to recognize and cluster apart a particular musical source from the mixture of sounds, even if we have not listened to that source before. [9] discusses various acoustic attributes which help in perceptual grouping of various acoustic stimuli. This grouping is conceptualized as taking place at two levels. At the first level, the acoustic stimuli are grouped into group objects or the perceptual units, and at the second level, these group objects are linked to different source streams. Our present work mostly focusses on the second level of grouping, which relies upon the features which characterize a source. The first level of grouping in one time frame is characterised by the given F0 value and its harmonics.

There have been several works on finding the multiple pitches or fundamental frequencies (F0's) in a polyphonic audio, so much so that it is a popular task in the MIREX challenge [1]. This paper focuses on identifying the instruments associated with producing those F0's. One popular paradigm to quantify the quality of sound is the spectral envelope. There are several works in the literature that learn spectral envelopes for instrument identification in a supervised fashion. [7] uses sinusoidal modeling to represent the audio signal and groups the peaks over a few consecutive frames using heuristic auditory cues and spectral clustering. These clusters are then assigned to a source with the help of timbre models trained beforehand. [6] models the harmonic partials and temporal evolution using Gaussian mixture model (GMM). Instrument identification is performed using pre-trained support vector machine classifier over statistical features derived from the spectral and temporal parameters. A popular tool to analyze polyphonic sound is the non-negative matrix factorizarion (NMF). [12] uses a source-filter model based NMF for sound separation and subsequently, the instrument identification is carried out using pre-trained GMM's over Mel-Frequency cepstral coefficients.

These supervised methods are dependent on the availability of isolated sounds from each contributing source. But many times, the isolated source audio is not available, for example for commercial CD's. Hence, we have to resort to unsupervised techniques for source identification. This possibility is asserted by the human ability to cognize and cluster any singer from the polyphonic song even if the song or the singer is heard for the first time.

Several works aiming in this direction cluster the various sources in an unsupervised way by matching their timbral properties. [8] applies NMF to extract dictionary components, which are then grouped into sources by unsupervised clustering over the Mel-frequency cepstral co-

[1] http://www.music-ir.org/mirex/wiki/2012

efficients extracted from the dictionary components. [11] improves upon this method by using shifted NMF for the unsupervised clustering of the dictionary components. [5] uses agglomerative approach to cluster the matrices, obtained by the NMF decomposition of the input signal, into different sources. These works have their foundations rooted in the hypothesis that the sounds having similar spectral envelops originate from the same source.

The present work is an exploration of this hypothesis for identifying the sources present in a polyphonic audio in a semi-supervised way. A polyphonic audio, with several musical sources playing together, is given. Since this work focuses on identifying the underlying instrument, we assume the active fundamental frequencies to be given apriori, as they can be extracted using a multi-F0 extractor [3]. Our goal is to identify the instrument associated with each F0. The system is semi-supervised in the sense that a human user annotates a few segments of pitch contours with the corresponding source labels and feeds this information for initializing the system. The proposed algorithm then labels the entire audio with the corresponding instruments. A source-filter based Probabilistic Latent Component Analysis (PLCA) framework is used to decompose the given polyphonic audio spectrum into an appropriate representation in terms of various elementary spectra embedded in a Bayesian network of various lateral variables. The clustering is performed using probabilistic graph clustering framework. All the samples in the given F0 contours act as nodes of the graph and the edges connecting them are modelled by a similarity relation. The similarity between nodes is derived from various acoustic cues like pitch continuity, spectral characteristics and simultaneity constraints.

The major advantage of our approach is that it does not need individual source audios for training but requires only a few annotations per instrument. Hence this approach can be used even for unseen instruments and commercial recordings. Conceptually, the novelty of this work is that it applies graph clustering to the NMF framework. Graph based method helps to model the metrics (or distances) between the sound objects in a non-Euclidean way and hence gives a lot of flexibility to model the auditory space. The probabilistic graph based clustering can incorporate various perceptual cues which may help in grouping the sources, thereby enhancing the performance. In this work, we propose novel ways to model these constraints.

The polyphonic signal representation using PLCA framework is explained in Section 2. The graph based modeling of instrument similarity is proposed in Section 3 followed by the probabilistic graph clustering method in Section 4. Section 5 gives the complete overview of the proposed algorithm and the experimental evaluation of the same.

2. SIGNAL REPRESENTATION

The single channel polyphonic audio is considered to be a linear additive mixture of various source waveforms. The audio waveform is transformed to frequency domain by taking its 2048-point short-time Fourier transform with hanning window of length 56ms and hop size of 10ms. The magnitude spectrum is boosted by 6dB per octave. PLCA [10] assumes the magnitude spectrogram $V(f,t)$ of the polyphonic signal to be a histogram of *energy quanta* piled up in the time-frequency bins, indexed by t and f respectively, and generated by an underlying probability distribution function (pdf) $P_t(f)$. It further assumes that the magnitude spectrum of polyphonic audio is the linear sum of the magnitude spectra of individual source signals. Using the source-filter model [12], $P_t(f)$ can be factorized into the following latent variable representation:

$$P_t(f) = \sum_{p,s,z,a} P_t(f|p,a)P_t(p)P_t(s|p)P_t(z|p,s)P(a|s,z) \tag{1}$$

Here, p, s, z, a are latent variables underlying the generation of energy quanta at each time-frequency bin, which can take values from 1 to N_p, N_s, N_z, N_a, respectively. p indexes the pitch, s indexes the source instrument, z indexes the dictionary component and a indexes the frequency band associated with each t-f bin. $P_t(f|p,a)$ is the fixed spectrum consisting of the Gaussian harmonic peaks placed at the integer multiples of the given F0 associated with the pitch index p at time t and filtered through the ath triangular mel-frequency band pass filter. The other parameters can be learnt using the Expectation Maximization (EM) algorithm which minimizes the Kullback-Leibler divergence between the observed spectrum $V(f,t)$ and the underlying distribution model $P_t(f)$, as explained below.

E-step:

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

M-step:

$$P_t(p) = \frac{\sum_{f,s,z,a} V(f,t)P_t(p,s,z,a|f)}{\sum_p \sum_{f,s,z,a} V(f,t)P_t(p,s,z,a|f)} \tag{3}$$

$$P_t(s|p) = \frac{\sum_{f,z,a} V(f,t)P_t(p,s,z,a|f)}{\sum_s \sum_{f,z,a} V(f,t)P_t(p,s,z,a|f)} \tag{4}$$

$$P_t(z|p,s) = \frac{\sum_{f,a} V(f,t)P_t(p,s,z,a|f)}{\sum_z \sum_{f,a} V(f,t)P_t(p,s,z,a|f)} \tag{5}$$

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

Also, we assume the total number of instruments N_s to be known apriori. Our goal is to cluster all the (t,p) units based on the underlying sources s. This problem can be seen as that of estimating $P_t(s|p)$.

3. INSTRUMENT SIMILARITY MODELING

After suitably modeling the spectra, the next task is to cluster them based upon their originating sources. In order to cluster the sources, we need to know what features quantify the distinction among the spectra of different sources and the similarity between those of the same source.

All the objects to be clustered, i.e. the (t, p) units, are modeled as the nodes of a graph. The edges of the graph are modeled by the link-weights matrix A, such that $A_{ki;lj}$ represents the closeness between the pair of nodes (t_k, p_i) and (t_l, p_j). Here, t_k implies that the time index is k and p_i implies that the pitch index is i. The link-weights are constructed to lie within the range $[0, 1]$. A larger value of the link-weight denotes more closeness between the pair of nodes, i.e. it is more likely that both the nodes correspond to the same cluster.

To model the closeness between the nodes, we consider three factors which quantify the similarity between the spectra from same sources, namely, the spectral envelope, the temporal continuity of the pitch contours and the simultaneity constraint. Each one of these is explained as follows.

3.1 Spectral Envelope

This step uses Mel-frequency cepstral coefficients (MFCC's) in order to measure spectral-timbral closeness of sounds. Although the source-filter PLCA is also based on spectral envelope, but MFCC features provide different advantages, like de-correlating the filter bank weights. MFCC features have been very successful in the areas of speech research like speaker characterization [4]. The use of MFCC features for source characterization is based on the hypothesis that the spectra from the same source have similar spectral envelopes. The MFCC's quantify the position of the node (t, p) in the space of spectral envelopes. They are computed from the PLCA representation as follows.

The ath band-filtered spectrum associated with pitch index p at time t is estimated using the following Weiner filter,

$$V_{p,a}(f, t) = \frac{P_t(f, p, a)}{P_t(f)} V(f, t) \tag{7}$$

where, $P_t(f)$ is given by Equation (1) and

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

The energy values in the N_a sub-band spectra associated with a (t, p) unit are used to compute the 13 dimensional MFCC vectors $\mathbf{m}_{t,p}$. The closeness between two nodes, based on these features, is modeled using the cosine similarity measure which is defined as

$$\mathcal{D}_{\text{cos}}(\mathbf{m}_1, \mathbf{m}_2) = \frac{\mathbf{m}_1 \cdot \mathbf{m}_2}{|\mathbf{m}_1||\mathbf{m}_2|} \tag{8}$$

where, $\cdot$ represents the vector dot product and $|\mathbf{m}|$ stands for the Euclidean norm of vector $\mathbf{m}$. The link-weight matrix is estimated as

$$A^{\mathcal{S}}_{ki;lj} = \frac{\mathcal{D}_{\text{cos}}(\mathbf{m}_{t_k,p_i}, \mathbf{m}_{t_l,p_j}) + 1}{2} \tag{9}$$

with the superscript $\mathcal{S}$ denoting that it has been derived from the spectral features. The aforementioned equation ensures that although $\mathcal{D}_{\text{cos}} \in [-1, 1]$, $A^{\mathcal{S}}$ is constrained to lie in the interval $[0, 1]$.

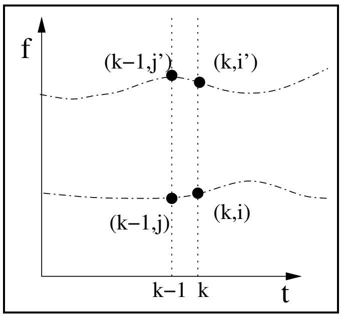


Figure 1. Temporal Continuity: dashed-dotted line shows two pitch contours.

3.2 Temporal Continuity

The pitch contour produced by a musical source changes slowly due to physical as well as musicological constraints. We can use this information as a cue [13] for clustering up the (t, p) units by forming the link-weight matrix $A^{\mathcal{T}}$, where the superscript $\mathcal{T}$ denotes the temporal connectedness measure. The closeness between F0s is modeled as

$$d^{lj}_{ki} = \exp\{-(F0_{ki} - F0_{lj})^2/\sigma^2\}. \tag{10}$$

For $A^{\mathcal{T}}_{ki;lj}$ to be large, we need to satisfy the following constraints:

1. $|t_k - t_l| = 1$

2. If d^{lj}_{ki} is large, $d^{lj'}_{ki}$ should be small, for $j' \neq j$, in order to avoid the ambiguity when the F0s at both the nodes are close to each other. For the same reason, $d^{ki'}_{ki}$ should also be small, for $i' \neq i$ (see Figure 1 with $l = k - 1$).

3. If (t, p) is well connected to (t', p') which is further well connected to (t'', p''), then (t, p) should also be well connected to (t'', p''). We call this as agglomeration.

Constraint 2 ensures that the temporal connectedness is strong only when the F0 trajectory under consideration is far apart in frequency from the other F0 trajectories.

In accordance with the above factors we devise the following novel scheme to construct $A^{\mathcal{T}}$.

```
for $k = 2$ to $N_t$ do
    $A^{\mathcal{T}}_{ki;(k-1)j} = d^{(k-1)j}_{ki} \left[ \frac{d^{(k-1)j}_{ki}}{\sum_j d^{(k-1)j}_{ki} + \sum_{j \neq i} d^{kj}_{ki}} \right]$    (11)
    {Agglomeration:}
    if $A^{\mathcal{T}}_{ki;(k-1)j} > \epsilon$ then
        $A^{\mathcal{T}}_{ki;k'i'} \leftarrow A^{\mathcal{T}}_{(k-1)j;k'i'}, k' = \{1, ..., (k-2)\}, \forall i'$
    end if
    {Making $A^{\mathcal{T}}$ symmetric:}
    $A^{\mathcal{T}}_{k'i';ki} \leftarrow A^{\mathcal{T}}_{ki;k'i'}, \forall k', \forall i'$
end for
```

3.3 Simultaneity Constraint

We assume that an instrument can play only a single note at a time. This assumption is undoubtedly valid for the vocal sources. For instruments, this assumption is not valid for harmony-based music systems where same instrument plays many notes simultaneously to form chords, but is valid for melody-based music systems where an instrument plays only one note at a time. This constraint implies that $N_s = N_p$.

The final link-weight matrix A is formed by combining $A^{\mathcal{S}}$ and $A^{\mathcal{T}}$ as,

$$A = (1-\alpha)A^{\mathcal{S}} + \alpha A^{\mathcal{T}} \tag{12}$$

where, $\alpha \in [0, 1]$. The simultaneity constraint is imposed as,

$$A_{ki;kj} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

Another implication of the simultaneity constraint in graph clustering is mentioned in the next section, viz., Section 4.

4. GRAPH CLUSTERING

For clustering, several techniques are used like agglomerative clustering, probabilistic clustering, spectral clustering etc. In this work, we use the probabilistic framework for graph clustering as proposed by [1]. The clustering problem is seen as an optimization problem which aims at maximizing the log-likelihood function,

$$\mathcal{L}(\Omega, A) = \sum_{s} \sum_{k,l,i,j} \ln P(\omega_{ki}^s, \omega_{lj}^s | A_{ki;lj}) \tag{14}$$

where, $\omega_{ki}^s \in \Omega$ is the cluster membership function which measures the degree of affinity of the unit (t_k, p_i) to the sth source cluster. We can see that ω_{ki}^s corresponds to $P_{t_k}(s|p_i)$. $P(\omega_{ki}^s, \omega_{lj}^s | A_{ki;lj})$ is modeled as a Bernoulli distribution,

$$P(\omega_{ki}^s, \omega_{lj}^s | A_{ki;lj}) = (A_{ki;lj})^{\omega_{ki}^s \omega_{lj}^s} (1 - A_{ki;lj})^{1-\omega_{ki}^s \omega_{lj}^s} \tag{15}$$

This equation ensures that if the link weight $A_{ki;lj}$ is large, then there is a large probability of ω_{ki}^s and ω_{lj}^s to be close to unity, meaning that the two nodes get clustered into the same source s. On putting this in Equation (14) and maximising with respect to ω_{ki}^s, we get,

$$\frac{\partial \mathcal{L}(\Omega, A)}{\partial \omega_{ki}^s} = \sum_{lj} \omega_{lj}^s \ln \frac{A_{ki;lj}}{1 - A_{ki;lj}} \tag{16}$$

The updated cluster membership function is estimated using soft-assign ansatz as

$$\hat{\omega}_{ki}^s = \frac{\exp[\partial \mathcal{L} / \partial \omega_{ki}^s]}{\sum_s \exp[\partial \mathcal{L} / \partial \omega_{ki}^s]} \tag{17}$$

This equation also takes care of the fact that ω_{ki}^s represents $P_{t_k}(s|p_i)$ and hence ought to be normalized with respect to s.

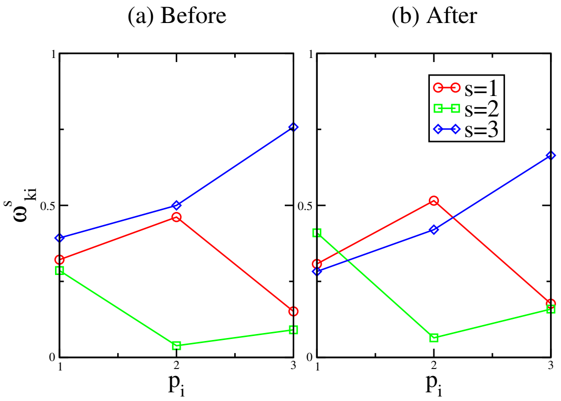


Figure 2. ω_{ki}^s vs p_i for different values of s and a fixed t_k. (a) All p_i are strongly linked to the cluster $s = 3$; (b) Equation (18) imposes the simultaneity constraint.

Under this formulation, it is quite possible that both ω_{ki}^s as well as ω_{kj}^s, for some $j \neq i$, have large values. However, the simultaneity constraint prevents this from happening. This is achieved by considering the point-cluster relation in a two-way sense, i.e. how simultaneous (t, p) units link to one cluster as well as how many clusters relate to one such unit. The simultaneity constraint is imposed in a non-rigid way by simply updating the cluster membership function as,

$$\omega_{ki}^s \leftarrow \frac{\omega_{ki}^s}{\sqrt{\sum_j \omega_{kj}^s}}, \tag{18}$$

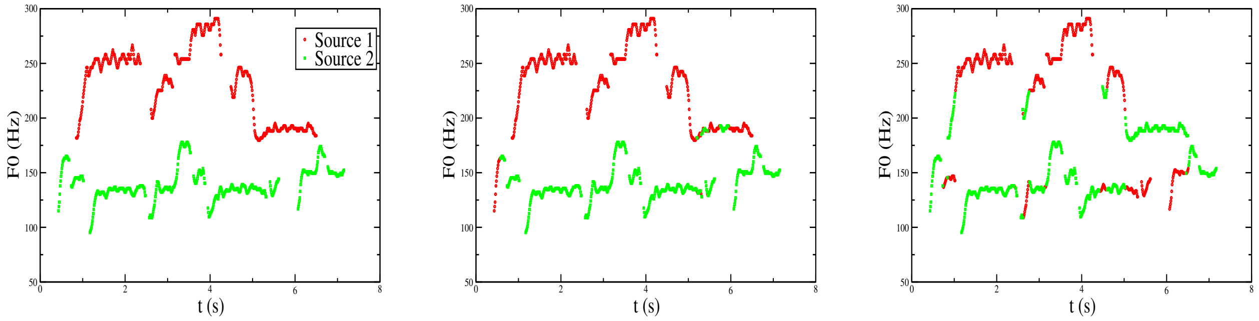
followed by normalization with respect to s, so as to maintain its probabilistic interpretation as $P_t(s|p)$. To understand this equation, let us consider the binary clustering case with $N_s = 2$. For a certain k, if $\omega_{kj}^{s_1} > \omega_{kj}^{s_2}, \forall j$, it implies that both the simultaneous units, $(t_k, p_j), j = \{1, 2\}$, are strongly associated with the source s_1. This situation implies that $\sum_j \omega_{kj}^{s_1} > 1$ and $\sum_j \omega_{kj}^{s_2} < 1$. The simultaneity constraint however imposes that only one p should be associated with one s. Hence, we attenuate $\omega_{kj}^{s_1}, \forall j$ and amplify $\omega_{kj}^{s_2}, \forall j$ by simply dividing with $\sqrt{\sum_j \omega_{kj}^s}$. This effect is also achieved for larger values of N_s, as graphically depicted in the Figure 2.

5. EXPERIMENTS

5.1 Implementation of Complete Algorithm

Given a polyphonic song along with all the active pitch values indexed by p at time frame index t, our task is to cluster the (t, p) units into the underlying sources. Instead of completely unsupervised clustering, we make the task semi-supervised by randomly choosing a few (3 here) time indices and labeling the (t, p) units at those times with their respective sources. This is the only little annotation work which has to be manually performed in case of absence of any other training data.

The source dictionaries $P(a|s, z)$ are randomly initialized and pre-trained using this little annotated data from



(a) Ground Truth

(b) Proposed Algorithm

(c) Algorithm [8]

Figure 3. Example of Source Clustering for a mixture of Ken-Gen singer pair. (a) Ground truth (b) the proposed algorithm and (c) the algorithm of [8]

the same song. These source dictionaries are then used as seeds for the complete algorithm as explained below.

1. PLCA decomposition of the entire song is performed, using the EM algorithm Equations (2)-(6), iterated 10 times.

2. The spectral and temporal link-weight matrices are computed as explained in Section 3.

3. The probabilistic source labels ω_{ki}^{s} are initialized as $P_{t_k}(s|p_i)$ and are updated using graph clustering as explained in Section 4. Then, $P_t(s|p)$ is updated as

$$P_{t_k}(s|p_i) \leftarrow \frac{P_{t_k}(s|p_i) + \omega_{ki}^{s}}{2}$$

   because generally ω_{ki}^{s} is found to have extreme values (0 or 1) due to its exponential update rule.

4. These steps 1 to 3 are repeated for a fixed number of times (3 here).

Finally, the source index is estimated using maximum aposteriori estimator along with the simultaneity constraint as

$$\hat{\mathbf{s}}_t = \arg\max_{\mathbf{s}_t}\{\sum_p P_t(\mathbf{s}_t(p)|p)\}. \quad (19)$$

Here, $\mathbf{s}_t$ is the vector formed by permutations of the vector $[1, ..., N_s]$, so that the final instrument labels are $\hat{s}_{t,p} = \hat{\mathbf{s}}_t(p)$.

Notably, the $P_t(s|p)$ for the given (t, p) units are fixed to the given values throughout the algorithm. The various parameters were chosen heuristically. Number of dictionary components for each source is set as $N_z = 2$ and the number of mel-frequency band filters $N_a = 20$. In subsection 3.2, we set $\sigma = 20$ and $\epsilon = 0.6$.

5.2 Evaluation

The clustering of vocal sources is more challenging than that of most of the musical instruments. The reason is that for a certain F0, the spectrum of the latter remains mostly the same but that of the former varies significantly due to the variety of voiced sounds (or vowels) which can be produced by a human singer.

The proposed algorithm was evaluated with the help of vocal songs from the MIR-1k database [2], which is publicly available. The evaluation dataset consisted of audio recordings of 4 different singers - 2 males (Abj, Gen) and 2 females (Ken, Hey). Total 60 single-channel audio mixtures of polyphony order $N_s = 2$ were formed by linearly adding the monophonic audio waveforms, with a total duration of about 350 seconds.

The proposed approach was compared with that proposed by Spiertz and Gnann [8], whose implementation is available as open source. Their method is oriented to produce separated signals instead of instrument clusters. To estimate their performance on instrument identification, we estimated $P(s|p, t)$ by comparing the spectra using the cosine similarity measure, defined in Equation (8).

$$\hat{P}^{SG}(s|p, t) = \mathcal{D}_{\cos}(\mathbf{Y}_{p,t}, \mathbf{S}_{s,t}) \quad (20)$$

Here, $\mathbf{Y}_{p,t}$ and $\mathbf{S}_{s,t}$ are the magnitude spectral vectors of the pth output channel and the sth monophonic source used to construct the input, respectively, at time frame t.

For evaluation metrics, all the (t, p) units are clustered into N_s number of groups. Each obtained cluster is linked to a ground truth source in a way which maximizes the classification accuracy. The classification accuracy is measured as the fraction of (t, p) units correctly labeled with the ground truth source.

The evaluation scores are presented in Table 1. We can see that the proposed semi-supervised algorithm performs better than the unsupervised algorithm. Although this comparison is not well balanced, but still it gives an idea of the performance. Also, we can see that for our algorithm, the male-female voice differentiability is larger than the other two cases, viz., male-male and female-female.

This algorithm works quite well with short audio excerpts, as we tested it for upto 10s long excerpts, i.e. the size of A was upto 2000×2000. But for long songs,

Singers pair	Classification Accuracy in %	
	Proposed algorithm	Algorithm [8]
Abj-Hey	79.5 (16.5)	62.7 (15.0)
Ken-Hey	72.9 (13.7)	68.3 (13.5)
Gen-Hey	86.2 (13.0)	65.4 (11.2)
Abj-Ken	79.5 (16.8)	74.3 (9.9)
Abj-Gen	68.1 (13.5)	66.0 (10.0)
Ken-Gen	87.2 (9.9)	71.4 (14.8)

Table 1. Evaluation Results with standard deviations in brackets

the size of the link-weight matrix becomes very large and hence the computations become extensive. Thus, it is wise to break the long song into shorter segments ($<$ 10s in duration). In such a case, one may pre-train the dictionaries for each audio segment using the few labeled (t, p) units, even if they fall in different audio segments, and perform the proposed algorithm over one segment at a time.

6. CONCLUSION

This paper proposed a novel algorithm for instrument identification in polyphonic music. The major advantage of the algorithm is that it is semi-supervised as it does not require any training data in the form of individual source audios. It can initialize from annotations only at a few time instants. The method uses source-filter based PLCA for decomposing the magnitude spectra of the polyphonic audio. This PLCA based representation is used to form a graph which is clustered into the underlying sources with the help of various perceptual cues, including the spectral and temporal cues. Novel ways of modeling these cues were proposed. The temporal constraints take care of not only the continuity of the pitch contour individually but also its interaction with other pitch contours. To model the simultaneity constraints, a new equation is introduced which takes care of the point-cluster relation in a two-way sense, so that the simultaneously occuring points are assigned to different clusters.

The experimental results show that the proposed algorithm performs better than a state of the art unsupervised source separation algorithm adapted for instrument identification. The proposed framework can also be used with supervised schemes so as to enhance the performance by incorporating various acoustic cues.

7. REFERENCES

[1] A. Robles-Kelly, and E. R. Hancock: “A Probabilistic Spectral Framework for Grouping and Segmentation,” *In Pattern Recognition*, Vol. 37, No. 7, pp. 1387–1405, 2004.

[2] C.-L. Hsu, and J.-S. R. Jang: “On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset,” *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 18, No. 2, pp. 310–319, 2010.

[3] C. Yeh, A. Roebel, and X. Rodet: “Multiple fundamental frequency estimation and polyphony inference of polyphonic music signals,” *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 18, No. 6, pp. 1116-1126, 2010.

[4] D. A. Reynolds: “An overview of automatic speaker recognition technology,” *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vol. 4, pp. 4072–4075, 2002.

[5] J. M. Becker, M. Spiertz, and V. Gnann: “A probability-based combination method for unsupervised clustering with application to blind source separation,” *Latent Variable Analysis and Signal Separation (LVA/ICA)*, pp. 99-106, 2012.

[6] 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 Journal of Selected Topics in Signal Processing*, Vol. 5, No. 6, pp. 1124–1132, 2011.

[7] L. G. Martins, J. J. Burred, G. Tzanetakis, and M. Lagrange: “Polyphonic Instrument Recognition Using Spectral Clustering,” *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, 2007.

[8] M. Spiertz, and V. Gnann: “Source-filter based clustering for monaural blind source separation,” *Proceedings of International Conference on Digital Audio Effects*, 2009.

[9] M. Weintraub: “A theory and computational model of monaural auditory sound separation,” *Ph. D. dissertation,* Stanford University, 1985.

[10] P. Smaragdis, B. Raj, and M. Shashanka: “A probabilistic latent variable model for acoustic modeling,” *Advances in models for acoustic processing, NIPS*, 2006.

[11] R. Jaiswal, D. FitzGerald, D. Barry, E. Coyle, and S. Rickard: “Clustering NMF basis functions using shifted NMF for monaural sound source separation,” *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 245–248, 2011.

[12] T. Heittola, A. Klapuri, and T. Virtanen: “Musical instrument recognition in polyphonic audio using source-filter model for sound separation,” *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, 2009.

[13] V. Arora, and L. Behera: “On-line melody extraction from polyphonic audio using harmonic cluster tracking,” *IEEE Transactions on Audio, Speech and Language Processing,* Vol. 21, No. 3, pp. 520–530, 2013.