DISCRIMINATIVE PLCA BASED POLYPHONIC SOURCE IDENTIFICATION

Vipul Arora and Laxmidhar Behera

Department of Electrical Engineering,
Indian Institute of Technology, Kanpur
{vipular,lbehera}@iitk.ac.in

ABSTRACT

This work aims at searching for discriminatively learned features that characterize an audio source and make it identifiable even in polyphonic audio. Probabilistic latent component analysis (PLCA) is an effective method for decomposing a polyphonic signal into individual sources using source-specific dictionaries. This work proposes a novel discriminative approach to find better PLCA dictionaries which discriminate between the pitched sources more efficiently, by learning the differences in their harmonic spectra. The experimental results for source identification show promising advantages of the proposed approach over the generative PLCA approach for source classification.

***Index Terms*—** Probabilistic latent component analysis (PLCA), music information retrieval, source identification.

1. INTRODUCTION

Humans can carry out the task of pitch perception and source identification very easily, even in the case of polyphonic music, where several musical sources play together. But these tasks are extremely difficult to be mimicked by a machine. Monophonic instrument identification is carried out with quite good accuracy, using the features derived from the spectral envelope [1]. For polyphonic music, source (voice or instrument) identification becomes difficult due to the overlaps in the spectra of different sources playing simultaneously.

A convenient method to analyze overlapping spectra is non-negative matrix factorization (NMF). It models the observed spectrum as a weighted sum of spectral elements (dictionary) learnt from various sound sources. NMF is extended to probabilistic latent component analysis (PLCA) for better semantic interpretation, leading to enhanced modeling [2]. To reduce the number of spectral elements due to large number of pitches, the source filter model is used [3], such that the NMF model stores only the source specific spectral envelopes in the dictionary.

The problem of identifying the source associated with each given pitch value, in a mixture of pitched sounds from two or more sources, is addressed in this paper. The pitch values are assumed to be given as they can be extracted using a multi-pitch extractor [4]. The problem of source identification has been considered by several researchers. Some works [5] first separate the polyphonic music into separate sources and then perform monophonic instrument identification, using features derived from the separated channels. While some [6] attempt to identify the instrument from the mixed audio itself. Some others [7], [8] use NMF for modelling the audio signal and directly use the instrument specific gains to determine the instrument piano-roll.

The general NMF/PLCA approaches maximize the generative ability of the model, by minimizing the distance between the observed spectra and the modeled spectra. For classification problems, however, the regressional flexibility of the model can be better utilized by using discriminative training approaches [9]. A PLCA dictionary component does not represent the whole spectrum, but a part thereof. The discriminative approach can help in learning the parts in a way such that they represent the desired source well, but not so well the other sources. The empirical success of the discriminative approaches over the generative approaches has been observed in many frameworks. One of the most successful approaches, support vector machine [9], maximizes the margin between the classes. Pernkopf *et al.* [10] have used a maximum margin objective function to train Bayesian networks classifiers, but they do not consider latent variables. A discriminant NMF approach has been proposed for image classification applications in [11]. For audio signals, [12] uses discriminative Gaussian mixture model for single-channel audio source separation. A discriminative NMF approach has been proposed in [13], in the context of multi-pitch estimation. It appends the generative objective function with criteria to minimize the distance between the dictionary weights matrix and the ground-truth piano-roll representation. This approach has to be extended to the situations requiring decomposition into more than one latent variables, as in PLCA.

In this work, we aim at building a novel discriminative approach for the PLCA framework, in the context of musical source identification from polyphonic music, assuming that all the true pitch values at each time frame are given, as they can be extracted using a multi-pitch extractor. We model the observed polyphonic spectra using source-filter based PLCA approach. We classify the sources in the PLCA framework itself without the help of any post-PLCA classification scheme. We propose a novel discriminative learning approach for the PLCA dictionaries so as to maximize an objective function, which is highly representative of the classification error between the different sources. Our goal is to enhance the discriminative ability of the dictionaries by learning the differences between the source spectra. The proposed discriminative PLCA approach is compared with the generatively trained source-filter based PLCA. The main novel contribution of our work is to develop a discriminative framework for training the PLCA dictionaries.

Our method of representing the audio signal in source-filter based PLCA framework is explained in Section 2. Section 3 describes the generative and the proposed discriminative approaches to train PLCA dictionaries, as well as the source identification scheme. Experimental settings and the results are discussed in Section 4.

2. SIGNAL REPRESENTATION

The polyphonic audio signal is transformed to the spectrogram, $V(f,t)$ taking a 2048-point short time Fourier transform magnitude, computed using a Hanning window of length 56ms and hop size 10ms. Here, f,t are frequency and time indices, respectively. Each magnitude spectrum is boosted by 6dB per octave by multiplying the amplitude of each frequency bin by the corresponding frequency f, so as to enhance the spectral envelope while maintaining the linear separability of the sources.

Basic PLCA involves factorizing the (normalized) spectrogram as,

$$P_t(f) = \sum_z P(f|z)P_t(z). \tag{1}$$

Here, $P(f|z)$ represents the zth dictionary element and $P_t(z)$ represents the weight of the zth element. However, in order to obtain a more meaningful representation (similar to [7]) and to incorporate the source-filter model (as in [5]), we factorize the spectrogram into discrete latent variables p, s, z, a, f, which can take N_p, N_s, N_z, N_a, N_f values, respectively. We are given the pitch values at each time t, which are indexed by p. The goal is to identify the source underlying each pitch value. The trained source models are indexed by s, each of which has spectral dictionaries indexed by z. Each dictionary component contains the weights of N_a (= 20 here) band pass filters, indexed by a. This band-pass filter bank consists of triangular magnitude response filters, with centres uniformly distributed on Mel-frequency scale as in [5]. Time index t is a deterministic parameter. Hence, we can modify (1) as,

$$P_t(f) = \sum_{p,s,z,a} \int_{F_0} P(f|F_0, a)P_t(F_0|p) \\ P_t(p)P_t(s|p)P_t(z|p,s)P(a|z,s)\, dF_0. \tag{2}$$

F_0 represents the pitch value at time t, which is deterministic if the source p is known, i.e., $P_t(F_0|p) = \delta(F_0 - F_0^{p,t})$, $\delta(\cdot)$ being the Dirac delta function. Hence, we can write

$$\int P(f|F_0, a)P_t(F_0|p)dF_0 = P(f|p, F_0^{p,t}, a) = P_t(f|p,a)$$

which simplifies (2) to:

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

Here, we have omitted $F_0^{p,t}$ for the ease of representation. $P_t(f|p,a)$ represents the spectrum of the output of ath filter, at time t and pitch p. We model it using the source-filter model as,

$$P_t(f|p,a) = \frac{e(f|F_0^{p,t}) \cdot h(f|a)}{\sum_f e(f|F_0^{p,t}) \cdot h(f|a)}. \tag{4}$$

Here $e(f|F_0)$ is the excitation spectrum generated at fundamental frequency F_0 and $h(f|a)$ is the magnitude transfer function of the ath triangular band pass filter. $P_t(p)$ can be interpreted as the contribution fraction of the source with pitch p to the whole spectrum. $P_t(s|p)$ is the probability that it is the sth source which corresponds to the pitch p at the time instant t. $P_t(z|p,s)$ is the probability of the zth dictionary component of source s, whose spectral envelope is formed by $P(a|s,z)$, in the form of coefficient of ath band-pass filter.

Hence, instead of modeling the whole spectrum as a linear sum of basis spectra, we model the spectral envelopes, in the form of filter-bank coefficients, as a linear sum of basis envelopes. These basis elements are stored in a dictionary while training and are subsequently used for the separation task. In general, the goal of PLCA is to estimate the model probabilities (in (3)) from the observed spectra $V(f,t)$.

3. THE METHOD

The source-filter model has been incorporated into NMF by [5]. This section incorporates the source-filter model into the PLCA framework and then proposes a novel discriminative training method for the same.

3.1. PLCA Generative Model

The PLCA decomposes the signal into various components according to (3). The decomposition takes place with a goal that the model closely represents the observed spectra. This is accomplished using Expectation Maximization (EM) algorithm which minimizes the Kullback-Leibler divergence between $P_t(f)$ and $V(f,t)$. The likelihood

$$G = \sum_{f,t} V(f,t) \log P_t(f) \tag{5}$$

is maximized by sequentially iterating the E and M steps, subject to the constraints that the probabilities sum to unity. In the E-step, the current model probabilities are used to find the posterior distribution of the latent variables.

$$P_t(p, s, z, a|f) = \frac{P_t(f|p, a)P_t(p)P_t(s|p)P_t(z|p, s)P(a|s, z)}{P_t(f)} \tag{6}$$

where, $P_t(f)$ is given by (3).

In the M-step, the model probabilities are updated, so as to maximize the likelihood function, as

$$P_t(p) = \frac{\sum_{f,s,z,a} V(f, t)P_t(p, s, z, a|f)}{\sum_p \sum_{f,s,z,a} V(f, t)P_t(p, s, z, a|f)} \tag{7}$$

$$P_t(s|p) = \frac{\sum_{f,z,a} V(f, t)P_t(p, s, z, a|f)}{\sum_s \sum_{f,z,a} V(f, t)P_t(p, s, z, a|f)} \tag{8}$$

$$P_t(z|p, s) = \frac{\sum_{f,a} V(f, t)P_t(p, s, z, a|f)}{\sum_z \sum_{f,a} V(f, t)P_t(p, s, z, a|f)} \tag{9}$$

$$P(a|s, z) = \frac{\sum_{f,t,p} V(f, t)P_t(p, s, z, a|f)}{\sum_a \sum_{f,t,p} V(f, t)P_t(p, s, z, a|f)}. \tag{10}$$

For training the dictionary for sth source, we use its monophonic recordings for which the pitch contour is given. Hence, in (3) and the EM equations, $N_p = N_s = 1$, and $P_t(p)$ and $P_t(s|p)$ become fixed to unity. M-step computes $P_t(z|p, s)$ and $P(a|s, z)$, out of which $P(a|s, z)$ are stored as the characterising dictionary of source s.

3.2. Discriminative PLCA

The proposed discriminative training of PLCA aims to maximize the soft multi-class margin based objective function,

$$M = \sum_t \log \frac{P_t(s^t|p)}{\max_{s \neq s^t} P_t(s|p)} \tag{11}$$

$$\approx \sum_t \log \frac{P_t(s^t|p)}{[\sum_{s \neq s^t} P_t(s|p)^\eta]^{1/\eta}}, \text{for } \eta \gg 1 \tag{12}$$

with s^t being the ground truth source at time t. We train from monophonic audio, i.e., $p = 1$ for all t and hence, we omit p further in this section for the ease of representation.

The objective function is maximized using adaptive gradient ascent algorithm. At each iteration, the probabilities $P_t(s), P_t(z|s)$ are updated through the generative approach (Eqs. (8), (9)). Then, the dictionaries of spectral envelopes - $P(a|s, z)$ are updated by moving in the direction which best enhances the objective function M.

$$w^{i+1} = w^i + \epsilon_1 \frac{\partial M}{\partial w^i} + \epsilon_2(w^i - w^{i-1}) \tag{13}$$

Here, w^i represents the parameter values after the ith iteration, and ϵ_1, ϵ_2 are the fixed learning rates. The derivative of the objective function with respect to the parameters is given as:

$$\frac{\partial M}{\partial P(a|s, z)} = \sum_{t,f} \mathcal{V}_{fts} \frac{P_t(f|a)P_t(s)P_t(z|s)}{P_t(f)}. \tag{14}$$

Here,

$$\mathcal{V}_{fts} = V(f, t) \times \left(\frac{V_s^t}{P_t(s)} - \frac{\sum_{s',z',a'} V_{s'}^t P_t(f|a')P_t(z'|s')P(a'|s', z')}{P_t(f)} \right)$$

and, $V_s^t = \begin{cases} 1, & \text{for } s = s^t, \\ \frac{-(P_t(s|p))^\eta}{\sum_{s \neq s^t}(P_t(s|p)^\eta)}, & \text{otherwise.} \end{cases}$

To keep the non-negativity and sum-to-unity constraints for probability values, we re-parametrize the parameters as

$$P(a|s, z) = \frac{\exp(\beta(a|s, z))}{\sum_{a'} \exp(\beta(a'|s, z))}. \tag{15}$$

Hence,

$$\frac{\partial M}{\partial \beta(a|s, z)} = P(a|s, z) \left[\frac{\partial M}{\partial P(a|s, z)} - \sum_{a'} P(a'|s, z) \frac{\partial M}{\partial P(a'|s, z)} \right]$$

is used to learn the parameters $\beta(a|s, z)$ using (13), which are further used to estimate the dictionary components $P(a|s, z)$ using (15). The dictionary components learnt in this way tend to discriminate better between the sources by learning the differentiating features.

On comparing the numerator of the generative update of $P(a|s, z)$ (Eq. (10)) with the derivative of the discriminative objective function w.r.t the same (Eq. (14)), we find that the $V(f, t)$ in the former is replaced with $\mathcal{V}_{fts}$ in the latter. On analyzing the term $\mathcal{V}_{fts}$, we find that for $s = s^t$, it is mostly similar to positively scaled $V(f, t)$, but for $s \neq s^t$, it is similar to a negatively scaled version of the same. Fig. 1 pictorially represents the same. This negative contribution of $V(f, t)$ to the incorrect classes enhances the classification margin.

3.3. Source Identification

The goal of this stage is to identify the sources in the given audio mixture, corresponding to the given pitch values. This task is carried out by first decomposing the given spectrogram with the help of trained dictionary components $P(a|s, z)$ for each source. The decomposition is carried out using EM algorithm described in subsection 3.1, while keeping $P(a|s, z)$ fixed to the trained values.

The source identification is carried out frame-by-frame without considering the streaming rules, which are supposed to further enhance the accuracy. Temporal evolutions also handle the pitch overlaps more effectively [14]. Further, the identification is carried out in two different ways: unconstrained and constrained.

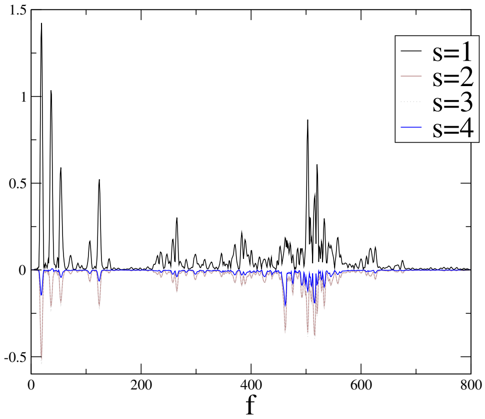

Fig. 1. $\mathcal{V}_{fts}$ (zoomed to $f \in [1, 800]$) while training for 4 sources, for frame t where only the source $s^t = 1$ is active.

For unconstrained identification, the source corresponding to pitch p is chosen using the MAP estimator,

$$\hat{s}_t(p) = \arg\max_s \{P_t(s|p)\}. \tag{16}$$

For constrained identification, we use the constraint that one source corresponds to only one pitch at a time, i.e., $\hat{s}_t(p) \neq \hat{s}_t(p')$, for $p \neq p'$. This obviously implies $N_s \geq N_p$. We represent the source-pitch mapping with a matrix $\mathbf{R}_t$,

$$R_t(s,p) = \begin{cases} 1, & \hat{s}_t(p) = s \\ 0, & \text{otherwise} \end{cases}$$

constrained to,

$$\sum_p R_t(s,p) \leq 1, \forall s.$$

R_t can be seen to be formed by row switching operations on an $N_s \times N_p$ matrix which has all diagonal elements as unity and others as zero. Hence,

$$\mathbf{R}_t = \arg\max_{\mathbf{R}_t} \{\sum_i \sum_j P_t(s_i|p_j) R_t(i,j)\}. \tag{17}$$

4. EXPERIMENTS

The proposed discriminative training algorithm (Subsection 3.2) is compared with the generative training algorithm (Subsection 3.1). Notably, the training algorithms are different, but same identification algorithm is used for both the approaches. The generative model used in this work is similar to the one proposed in [2], with the source-filter model as in [5], and source identification using the model weights directly, as in [7]. The dataset consists of singer voices from the publicly available MIR-1k dataset[1], along with their pitch values. We used the voices from 4 different singers (2 males and 2 females) and added their waveforms to produce the polyphonic mixtures. Although this random mixing of sounds produces musically non-meaningful polyphonic audio, the evaluations over such audio are acceptable indicators of performance [5]. We used 10 audio files, each around 8s in duration, from each singer.

Table 1. Classification Accuracies (standard deviations), in %, for $N_p = 4$

Algorithm	Unconstrained	Constrained
G-PLCA	43.0 (5)	48.4 (7)
D-PLCA	43.1 (3)	47.9 (6)
DF-PLCA	41.7 (3)	48.1 (7)
A-PLCA	43.7 (5)	46.5 (8)

Table 2. Classification Accuracies (standard deviations), in %, with G-PLCA and D-PLCA

	Unconstrained		Constrained	
N_p	G-PLCA	D-PLCA	G-PLCA	D-PLCA
2	65.8 (9)	71.5 (7)	69.0 (9)	74.4 (8)
3	48.7 (7)	50.3 (6)	52.1 (8)	53.6 (7)
4	43.0 (5)	43.6 (4)	48.4 (7)	49.0 (7)

Several versions of discriminative PLCA were implemented. The one with all dictionary components trained by maximizing the margin is denoted as D-PLCA. This version models the differences in the sources, but the spectral parts which are common among two or more sources may remain uncaptured. Hence, we tried another variant (DF-PLCA) of it, which first learns the discriminative components and then learns some filler components using generative learning. These filler components tend to capture the common spectral parts unmodeled by the discriminative components. Another augmented version (A-PLCA) trains the components so as to maximize an augmented objective function, which maximizes both the generative as well as discriminative functions in turn, like the coordinate ascent scheme.

The training of dictionaries was performed from monophonic files (unmixed), 2 for each source. The dictionary components were initialized randomly. The parameters for discriminative training are chosen as $N_z = 10, \eta = 10, \epsilon_1 = 1, \epsilon_2 = 0.2$. For filler dictionaries (in DF-PLCA and A-PLCA), out of N_z components, half were taken as discriminative and the other half were taken as filler components. These parameters were set heuristically and may be optimized to get better perfomance.

For evaluating the performances, the classification accuracies are compared. The classification accuracy is calculated

[1] http://mirlab.org/dataset/public/MIR-1K.rar

as the number of correctly estimated source labels to the total number of source labels, in each time frame, over the given pitch values. The accuracies obtained by using various algorithms are shown in Table 1. These are averaged over dictionaries trained from several random initializations.

The average performance of the discriminative PLCA is comparable to that of the generative one. Notably, the objective function associated with D-PLCA is non-convex, giving rise to local maxima. Hence, the initialization point is important in determining the classification performance. For this purpose we created a development set for training and tried several initializations of $P(a|s,z)$. One such instance of dictionaries giving good performance over the development set is used to perform classification over the test set for various orders of polyphony (N_p) and the obtained results are reported in Table 2. The improvement due to discriminative approach is more for lower order polyphonies and it reduces with increasing N_p. The better classification accuracies of D-PLCA as compared to those of G-PLCA shows that the discriminative approach has the capability to enhance the classification accuracies.

5. CONCLUSION & FUTURE WORK

In this paper, we have presented a novel discriminative learning approach for the probabilistic latent component analysis framework for source identification, maximizing the margin between the source classes. The simulation results show improvement in the classification ability of the discriminative PLCA over that of the generative PLCA for some good initialization points, and for a general initialization point, both the approaches show similar classification performances. This shows that the proposed scheme has the capability to improve the classification ability. We may take help of genetic algorithms for determining the good initialization points. The use of more sophisticated schemes for non-convex optimization may also improve the performance, as the gradient ascent method used in this work is highly susceptible to be stuck in local maxima. The experiments have been performed for vocal sources, which have a larger variation in timbre due to various vowels as compared to instruments, but the approach can be used for instrument identification as well.

6. REFERENCES

[1] B. L. Sturm, M. Morvidone, and L. Daudet, "Musical instrument identification using multiscale mel-frequency cepstral coefficients," in *Proc. European Signal Process. Conf.*, 2010.

[2] M. Shashanka, B. Raj, and P. Smaragdis, "Probabilistic latent variable models as nonnegative factorizations," *Computational Intelligence and Neuroscience*, 2008.

[3] J. L Durrieu, G. Richard, and B. David, "Singer melody extraction in polyphonic signals using source separation methods," in *Proc. ICASSP*, 2008, pp. 169–172.

[4] C. Yeh, A. Roebel, and X. Rodet, "Multiple fundamental frequency estimation and polyphony inference of polyphonic music signals," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 18, no. 6, pp. 1116 –1126, 2010.

[5] T. Heittola, A. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation," in *Proc. Int. Society for Music Information Retrieval Conf.*, 2009.

[6] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Instrogram: Probabilistic representation of instrument existence for polyphonic music," *IPSJ Digital Courier*, vol. 3, pp. 1–13, 2007.

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

[8] J.J. Carabias-Orti, T. Virtanen, P. Vera-Candeas, N. Ruiz-Reyes, and F.J. Canadas-Quesada, "Musical instrument sound multi-excitation model for non-negative spectrogram factorization," *IEEE J. Selected Topics in Signal Process.*, vol. 5, no. 6, pp. 1144–1158, 2011.

[9] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag New York, Inc., 2006.

[10] F. Pernkopf, M. Wohlmayr, and S. Tschiatschek, "Maximum margin Bayesian network classifiers," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, pp. 521–532, 2012.

[11] I. Kotsia, S. Zafeiriou, and I. Pitas, "A novel discriminant non-negative matrix factorization algorithm with applications to facial image characterization problems," *IEEE Trans. Info. Forensics and Security*, vol. 2, no. 3, pp. 588 –595, 2007.

[12] V. Emiya, E. Vincent, and R. Gribonval, "An investigation of discrete-state discriminant approaches to single-sensor source separation," in *Proc. Work. Appli. Sig. Proces. Audio and Acous. (WASPAA)*, 2009, pp. 97–100.

[13] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "Discriminative non-negative matrix factorization for multiple pitch estimation," in *Proc. Int. Society for Music Information Retrieval Conf.*, 2012.

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