14th International Conference on Latent Variable Analysis and Signal Separation, LVA/ICA, Guildford, UK, 3–5 July 2018

# LATENT MIXTURE MODELS FOR AUTOMATIC MUSIC TRANSCRIPTION

*Cian O'Brien, Mark D. Plumbley*

Centre for Vision, Speech and Signal Processing
University of Surrey
{cj.obrien,m.plumbley}@surrey.ac.uk

## ABSTRACT

Polyphonic music transcription is a challenging problem, requiring the identification of a collection of latent pitches which can explain an observed music signal. Many state-of-the-art methods are based on the Non-negative Matrix Factorization (NMF) framework, which itself can be cast as a latent variable model. However, the basic NMF algorithm fails to consider many important aspects of music signals such as low-rank or hierarchical structure and temporal continuity. Here we propose a probabilistic model to address some of the shortcomings of NMF. Based on the Probabilistic Latent Component Analysis framework, we propose an algorithm which represents signals using a *collection* of low-rank dictionaries built from a base pitch dictionary. Experiments on a standard music transcription data set show that our method can successfully decompose signals into a hierarchical and smooth structure, improving the quality of the transcription.

## 1. INTRODUCTION

Automatic Music Transcription (AMT) attempts to reproduce the pitch content of a music signal. That is, it seeks a representation which specifies what musical pitches are present at each time frame. Given a time-frequency matrix  of a recorded music signal, we aim to find a binary *transcription matrix* which specifies the presence/absence of each musical pitch at every time frame. The most commonly used approach for AMT is the Non-negative Matrix Factorization (NMF) algorithm. In terms of AMT, the signal matrix  is usually a magnitude-frequency representation of the signal and we seek a factorization of the form

$$\mathbf{X} \approx \mathbf{W}\mathbf{H} \tag{1}$$

where $\mathbf{W}$ is a matrix whose columns contain individual pitches and the matrix  represents the final transcription.

## 2. PROBABALISTIC LATENT MIXTURE MODELS FOR AMT

We propose a latent variable model for the automatic transcription of polyphonic music using a probabilistic framework which is closely related to the PLCA model of Smaragdis et al. [1]. We model the observed time-frequency signal as a joint distribution $P(f,t)$. We may factorize the joint distribution according to our generative model as follows

$$P(f,t) = P(t)\,P(f\,|\,t). \tag{2}$$

Further factorizing $P(f\,|\,t)$ gives the asymmetric PLCA model

$$P(f,t) = P(t)\sum_p P(f\,|\,p)P(p\,|\,t). \tag{3}$$

Under this model, $P(f,t)$ is given as a sum over $p$ latent pitches which explain this signal. As opposed to standard PLCA, in this work we suppose that each latent model corresponds to a *collection* of pitches from some base dictionary, so that higher level concepts such as intervals and chords can be introduced. Formally, suppose that the signal is composed of $m$ latent models, each of which is of rank-$r$. Each model should be constrained to lie in a so-called *pitched-subspace*, so that they correspond to linear combination of valid pitches.

We can introduce this into the model given in (3) by defining

$$P(p\,|\,t) = \sum_{m,r} P(p\,|\,m,r,t)P(r\,|\,m,t)P(m\,|\,t). \tag{4}$$

Combining (4) and (2) gives the proposed *Hierarchical Latent Mixture Model* (HLMM)

$$P(f,t) = P(t)\sum_{p,m,r} P(f\,|\,p)P(p\,|\,m,r,t)P(r\,|\,m,t)P(m\,|\,t). \tag{5}$$

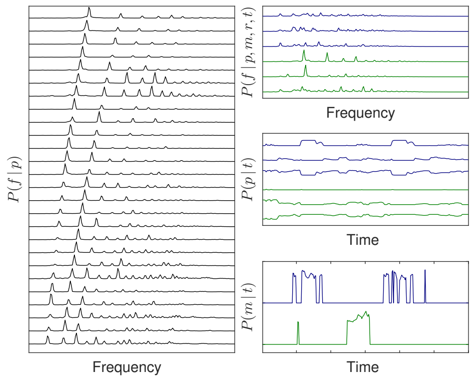


Figure 1: The proposed approach applied to piano music. In total, a collection of 30 models of rank-3 were learned.

## 3. INFERENCE

Given an observed time-frequency signal $\pi(f,t)$ the factors can be learned using the expectation-maximization (EM) algorithm to maximize the following log-likelihood

$$\sum_{f,t} \pi(f,t) \log \big(P(t)P(f \mid t)\big). \tag{6}$$

During the E-step, we compute the posterior distribution of the latent variables $p$, $m$ and $r$ which by Bayes theorem is given by

$$P(p,m,r \mid f,t) = \frac{P(f \mid p)P(p \mid m,r,t)P(r \mid m,t)P(m \mid t)}{P(f \mid t)}. \tag{7}$$

During the M-step, the remaining factors are updated using this posterior:

$$P(f \mid p) = \frac{\sum_{m,r} P(p,m,r \mid f,t)\, \pi(f,t)}{\sum_{f,m,r} P(p,m,r \mid f,t)\, \pi(f,t)} \tag{8}$$

$$P(p \mid m,r,t) = \frac{\sum_{f} P(p,m,r \mid f,t)\, \pi(f,t)}{\sum_{f,p} P(p,m,r \mid f,t)\, \pi(f,t)} \tag{9}$$

$$P(r \mid m,t) = \frac{\sum_{f,p} P(p,m,r \mid f,t)\, \pi(f,t)}{\sum_{f,p,r} P(p,m,r \mid f,t)\, \pi(f,t)}. \tag{10}$$

$$P(m \mid t) = \frac{\Big(\sum_{f,p,r} P(p,m,r \mid f,t)\, \pi(f,t)\Big)^{\alpha}}{\sum_m \Big(\sum_{f,p,r} P(p,r \mid f,t)\, \pi(f,t)\Big)^{\alpha}} \tag{11}$$

where $\alpha \geq 1$. After updating $P(m \mid t)$ we set values below a set threshold to zero before renormalizing. After solving for each latent factor in (3), the joint distribution $P(p,t)$ is given by

$$P(p,t) = \frac{\sum_{m,r} P(p \mid m,r,t)P(r \mid m,t)P(m \mid t)P(t)}{\sum_{p,m,r} P(p \mid m,r,t)P(r \mid m,t)P(m \mid t)P(t)}. \tag{12}$$

which is the desired transcription.

## 4. EVALUATION

The proposed system was evaluated on 30-second excerpts from the *EnStDkcl* subset of the Midi Aligned Piano Sounds (MAPS) dataset, which consists of recordings of classical piano music. The global dictionary $P(f \mid p)$ was initialized by training NMF models on recordings of isolated pitches from the test instrument. The proposed method significantly outperforms the standard NMF approach across all metrics.

| Reference | Model | $\mathcal{P}$ | $\mathcal{R}$ | $\mathcal{F}$ |
|---|---|---|---|---|
| | $\beta$-NMD | 73.13 | 70.90 | 71.70 |
| | PLCA | 72.26 | 72.41 | 72.07 |
| [2] | LR-$\beta$-NMD | 73.83 | 73.17 | 73.50 |
| [3] | W$\beta$-NMD | | | 73.70 |
| [4] | GS-KL-NMD | | | 74.10 |
| Proposed | HLMM | | | |

Table 1: Transcription results on the *EnStDkcl* data set.

## 5. REFERENCES


[1] P. Smaragdis, B. Raj, and M. Shashanka, "A probabilistic latent variable model for acoustic modeling," *Advances in models for acoustic processing, NIPS*, vol. 148, pp. 8–1, 2006.

[2] C. O'Brien and M. D. Plumbley, "Automatic music transcription using low rank non-negative matrix decomposition," in *Signal Processing Conference (EUSIPCO), 2017 25th European*, pp. 1848–1852, IEEE, 2017.

[3] K. O'Hanlon and M. D. Plumbley, "Automatic music transcription using row weighted decompositions," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013*, pp. 16–20, IEEE, 2013.

[4] K. O'Hanlon, H. Nagano, N. Keriven, and M. D. Plumbley, "Non-negative group sparsity with subspace note modelling for polyphonic transcription," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 3, pp. 530–542, 2016.