MIREX 2014 MULTIPLE FUNDAMENTAL FREQUENCY ESTIMATION AND TRACKING SUBMISSION: CONSTRAINED NMF

Daniel Recoskie
David R. Cheriton School of Computer Science
University of Waterloo
dprecosk@uwaterloo.ca

Richard Mann
David R. Cheriton School of Computer Science
University of Waterloo
mannr@uwaterloo.ca

ABSTRACT

In this submission we apply nonnegative matrix factorization (NMF) to the task of multiple fundamental frequency estimation and tracking. NMF is an unsupervised learning method which finds an additive model of data. Since each time point in a musical piece is composed of a sum of notes, NMF is a suitable analysis tool. We constrain the standard NMF model to be piecewise smooth and aligned in order to exploit the general structure of music.

1. INTRODUCTION TO NMF

Lee and Seung popularize the use of NMF by applying it to facial recognition (among other applications) [4]. Smaragdis and Brown [7] later introduced the idea of using NMF for music transcription. Since then, there has been much work in the area such as: realtime transcribing utilizing sparsity constraints [2], sparsity and temporal smoothness constraints [9, 10], harmonic constraints [8], a Bayesian approach to enforce harmonicity and smoothness constraints [1]. The methodology in this submission is fully described in [6].

In NMF we wish to factorize a data matrix $X_{d\times n}$ into the product of two smaller matrices $B_{d\times r}$ and $G_{r\times n}$ such that

$$X \approx BG \tag{1}$$

where we measure the closeness of X and BG using some error function. In this work we will consider the following metric which we will refer to as *divergence* as defined in [5]

$$D(X||BG) = \sum_{i,j}\left(X_{ij}\log\frac{X_{ij}}{(BG)_{ij}} - X_{ij} + (BG)_{ij}\right) \tag{2}$$

where M_{ij} refers to the element of M in the i^{th} row and j^{th} column.

We obtain our X matrix by taking the magnitude of the short-time Fourier transform (STFT) of our music sample. We can then use NMF to factorize the data into B and G matrices. The novelty of NMF is that the columns of B generally correspond to the magnitude of the Fourier transforms of each individual note in the piece (as well as a noisy column), and the rows of G correspond to each note's activation. See Figure 1 for an illustration of the process. The only parameter needed to be set in the standard model is r, the number of notes in the musical piece.

We impose sparsity and piecewise smoothness constraints on G by adding penalties to the error function.

Sparsity:

$$\lambda\sum_{i,j}(1 - e^{-G_{i,j}^2/2\sigma^2}) \tag{3}$$

Piecewise smoothness:

$$\lambda\sum_{i,j}\left(1 - e^{-(G_{ij}-G_{ij-1})^2/2\sigma^2}\right) \tag{4}$$

Piecewise smoothness with aligned notes:

$$\lambda\sum_{j}\left(1 - e^{-\sum_i(G_{ij}-G_{ij-1})^2/2\sigma^2}\right) \tag{5}$$

Similar piece-wise smoothness constraints have been used for the task of hyper spectral unmixing [3]. To obtain a factorization we use gradient descent in order to minimize the error function. We begin by randomly initializing our B and G matrices. We derive update rules based on our error function for each matrix separately. Once we have update rules for both B and G, we iteratively update each until convergence. Note that it is possible to train the model ahead of time by precomputing the B matrix using real or synthetic music samples. For details of the update rules see [6].

2. METHODOLOGY

Our submission is for the task of multiple fundamental frequency estimation and tracking. Our methods for each subtask (1. frame level evaluation and 2. note tracking) are very similar and differ only in preprocessing and postprocessing. In either subtask we start by finding the magnitude of the STFT of the music signal. For subtask 1 we use 20 ms triangular windows overlapping by 10 ms. For subtask 2 we use 100 ms triangular windows overlapping by 50 ms. Frequency bins are evenly spaced up to a maximum of 5 kHz.

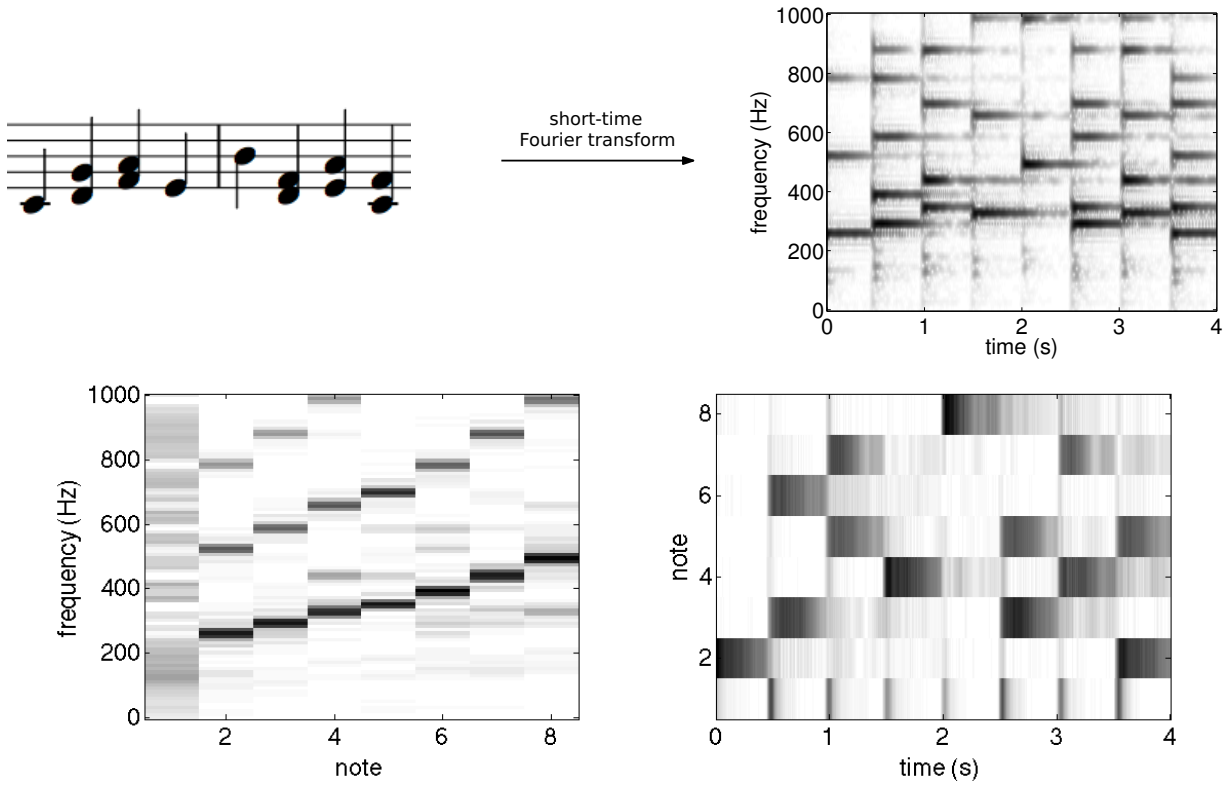

Figure 1. Top: The short-time Fourier transform of an audio signal is taken to obtain the matrix X. Bottom: NMF produces note matrix B (left) and note activation matrix G (right).

We make use of a precomputed B matrix learned from synthetically generated piano music, and hence we must only calculate the G (transcription) matrix for a given music sample. The B matrix has 88 columns corresponding to the possible keys on a piano. Once we have found our G we threshold its values to obtain a binary matrix where values of one indicate a note is activated.

Once we have our binary transcription matrix G we can easily output either frame level activations in the case of subtask 1, or note onset and offsets for subtask 2. An in-depth explanation of our methodology can be found in [6].

3. REFERENCES

[1] N. Bertin, R. Badeau, and E. Vincent. Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(3):538–549, March 2010.

[2] Arshia Cont. Realtime multiple pitch observation using sparse non-negative constraints. In *International Conference on Music Information Retrieval*, 2006.

[3] Sen Jia and Yuntao Qian. Constrained nonnegative matrix factorization for hyperspectral unmixing. *Geoscience and Remote Sensing, IEEE Transactions on*, 47(1):161–173, Jan 2009.

[4] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[5] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562. MIT Press, 2000.

[6] Daniel Recoskie. Constrained nonnegative matrix factorization with applications to music transcription. Master's thesis, University of Waterloo, Canada, 2014. http://hdl.handle.net/10012/8639.

[7] P. Smaragdis and J.C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, pages 177–180, Oct 2003.

[8] E. Vincent, N. Bertin, and R. Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(3):528–537, March 2010.

[9] T. Virtanen. Sound source separation using sparse coding with temporal continuity objective. In *Proc. Int. Comput. Music Conf*, pages 231–234, 2003.

[10] T. Virtanen. Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(3):1066–1074, March 2007.