TIME-DOMAIN PIANO MUSIC TRANSCRIPTION USING K-SVD

Zachary Neveu

Technical University of Denmark

ABSTRACT

This work presents a method for transcribing piano music in the time domain using k-means singular value decomposition (KSVD) given existing note onsets. The method presented allows for learning from small datasets and offers the potential to generalize to instruments with widely varying spectral characteristics and non-western musical scales.

***Index Terms*—** K-SVD, Factor Models, Music

1. INTRODUCTION

Music, in general, has a well defined structure that can be represented in simple forms such as sheet music or musical instrument digital interface (MIDI) scores. In these compressed forms, the salient features of music (note pitches, velocities, and timing) are stored directly, and can be searched and manipulated easily. Musical performance often deviates in interesting ways from written music, however live performances are most often recorded in audio form and require significant human effort to analyze or manipulate. Transforming music from a high-dimensional audio signal into a sparse symbolic form can be performed algorithmically which can save time and effort.

2. BACKGROUND

Traditional signal processing approaches can perform transcription for certain musical instruments and styles with a moderate level of success [1] . These approaches generally utilize the fact that a musical note is written as the lowest frequency (F_0) that occurs when this note is played. In the subtask of melody transcription, it is known that only one note can occur simultaneously, so simply finding the lowest prominent frequency in the time-frequency domain can often be adequate to transcribe a sample. In polyphonic transcription, however, multiple notes can be played at once. This problem can be approached well in a statistical context using maximum a-priori (MAP) estimation to determine the most likely combination of notes occurring given a set of frequencies [2]. Another approach to polyphonic transcription is to factor the magnitude spectrogram of the signal. Because of the nature of the short time fourier transform (STFT), the magnitude of each time-frequency bin must be non-negative, so non-negative matrix factorization has been applied by [3]. For piano music, another known constraint is that representations are sparse in the frequency domain. To utilize this, [4] used non-negative k-means singular value decomposition (NKSVD) in the spectrogram domain to constrain the sparsity of solutions along with the positivity constraint. More recently, neural networks have been applied to the polyphonic transcription task, most notably in [5]. Neural network approaches are able to benefit from learning musical structure, grasping the nonlinear relationships between notes over time, making them the most performant approach for polyphonic transcription. Neural architectures suffer from the lack of large-scale labeled data in this area, however, so highly general networks for multiple instruments and styles do not exist.

3. PROBLEM FORMULATION

For the purposes of this paper, it is assumed that a method for onset detection such as the one presented in [6] is available and performs well. The input for the problem is then time-domain frames of music each of which contains at least one note and at least one onset. The goal is to factor each frame F into an approximate linear combination Z of a dictionary of notes $\mathbf{D}$ as expressed by (1). The dimensionality of the variables are $F,\ Z \in R^{N\times n}$ and $D \in R^{n\times K}$ where N is the number of audio frames, n is the length of a frame, and K is the number of atoms in the dictionary. The constraint is imposed that the number of non-zero elements in $Z \leq k$ where $k \ll K$. For our specific purposes K was chosen to be 88 because a piano has 88 possible notes, and k was chosen to be 10 as the average pianist has 10 fingers.

$$F \approx \mathbf{D}Z \quad s.t. \|Z\|_0 \leq k \tag{1}$$

4. DATA

A subset of the MAESTRO dataset is used to fit the dictionary [7]. This data consists of audio files of piano music alongside MIDI files which contain the times, frequencies, and amplitudes of each note played. These audio files are oversampled to $44,100\,Hz$ to cover the full range of human hearing, however this is generally unnecessary for music transcription (the

```
# Initialization
D = random(n,K)
J = 1
# Sparse Coding
Z = matching_pursuit(F, D, k)
# Dictionary Update
for k=1:K
    E_k = F-(DZ-D[k]Z[k])
    w_k = ones(length(1:N))
    w_k[k] = 0
    E_k = E_k[wk,:]
    D[k], Z[k] = SVD(E_k)
    J += 1
```

Listing 1: K-SVD Algorithm [8]

```
# Input: signal F, dictionary
# Output: vector of coefficients Z
```

Listing 2: Matching Pursuit Algorithm [9]

highest piano note has an F_0 of $4186\,Hz$), so the audio files were downsampled by a factor of 4 to $11,025\,Hz$. Using the MIDI annotations, audio files were then split at note onsets. Each frame was then zero-padded or cropped so that the length of all frames was identical and of a tractable size of 512 samples. These frames were then concatenated into a $N \times 512$ matrix suitable for training the dictionary.

5. ALGORITHM

The KSVD algorithm was originally introduced in [8] and provides a way to learn an over complete dictionary with which to sparsely represent signals with a given sparsity.

6. REFERENCES

[1] Anssi Klapuri, *Signal Processing Methods for the Automatic Transcription of Music*, Ph.D. thesis, Tampere University of Technology, Tampere, 2004, OCLC: 76845615.

[2] Valentin Emiya, Roland Badeau, and Bertrand David, "Multipitch Estimation of Piano Sounds Using a New Probabilistic Spectral Smoothness Principle," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1643–1654, 2010.

[3] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," 2003, vol. 2003-, pp. 177–180, IEEE.

[4] N. Bertin, R. Badeau, and G. Richard, "Blind Signal Decompositions for Automatic Transcription of Polyphonic Music: NMF and K-SVD on the Benchmark," 2007, vol. 1, pp. I–65–I–68, IEEE.

[5] Curtis Hawthorne, Erich Elsen, Jialin Song, Adam Roberts, Ian Simon, Colin Raffel, Jesse Engel, Sageev Oore, and Douglas Eck, "Onsets and Frames: Dual-Objective Piano Transcription," *arXiv:1710.11153 [cs, eess, stat]*, June 2018.

[6] Sebastian Bock, Florian Krebs, and Markus Schedl, "EVALUATING THE ONLINE CAPABILITIES OF ONSET DETECTION METHODS," p. 6.

[7] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck, "Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset," *arXiv:1810.12247 [cs, eess, stat]*, Jan. 2019.

[8] Michal Aharon, Michael Elad, and Alfred M. Bruckstein, "K-SVD and its non-negative variant for dictionary design," in *Optics & Photonics 2005*, Manos Papadakis, Andrew F. Laine, and Michael A. Unser, Eds., San Diego, California, USA, Aug. 2005, p. 591411.

[9] S.G. Mallat and Zhifeng Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, Dec./1993.