Methods for Polyphonic Music Transcription

Jeremy Nash[a], Paul Schroeder[a], Mark Liu[b]

[a] *Electrical Engineering Department, University of Michigan, Ann Arbor, MI 48109*
[b] *Computer Science Department, University of Michigan, Ann Arbor, MI 48109*

Abstract

Music transcription is the process of converting an audio recording into a musical score. Musicians have been transcribing music manually for centuries to learn new songs and gain musical insight. Transcription, however, is an arduous manual process for musicians, and it's a difficult problem for computers because the pitches in a musical score are the human perceptions of a time-varying and frequency-varying physical phenomenon. There has been a flurry of research on how to solve this problem. We investigate the best solutions to the automated transcription problem and evaluate and compare their performance.

Keywords: Transcription, sparse coding, non-negative matrix factorization, Bayesian non-parametrics, music information retrieval

1. Introduction

1.1. Terminology

Semitones are the smallest music interval in Western music. On a piano, semitones are the adjacent keys.

Timbre represents the spectral envelope, time envelope, noise, tonality, and onset of a musical note. Timbre captures the unique properties of a musical instrument.

Pitches are represented by the notes used in musical notation. Pitches are a psychological perception and not an objective physical property. While we can intuitively assign a linear order to pitches (by referring to pitches as lower or higher than other pitches), pitches are rarely a single frequency and are instead a function of their timbre.

Polyphony means multiple simultaneous notes. Monophony means only a single note at a time. These terms carry slightly different meanings outside of the automated transcription field.

2. Motivation

Applications of a music transcription are versatile. Transcription produces a compact and standardized parametric representation of music. Such representation is needed for content-based retrieval of music in most current musical databases. It is useful in music analysis systems for tasks such as melody extraction, music segmentation and rhythm tracking. Transcription aids musicologists in analyzing music that has never been written down, such as improvised or ethnical music. The conversion of an acoustical waveform into parametric description is also useful in the process of making music, as well as in newer coding standards, such as MPEG-4, which may include such descriptions.

3. Problem Statement

4. Related Work

5. Methodologies

There are currently two main approaches to polyphonic music transcription. The first approach learns a supervised learning model that uses frequency magnitudes to predict the presence of a note. The second approach uses unsupervised learning techniques to extract

5.1. Audio data

5.2. Spectrogram

5.3. Constant Q transform

5.4. Semitone filter bank

Unlike the linear frequency scale used in an FFT, semitones in Western music are spaced logarithmically. This logarithmic spacing derives from our roughly logarithmic perception of pitches, first captured by Stevens et al. [1] through the mel scale.

Transforming the spectra to a logarithmic frequency scale achieves two main benefits. It reduces the dimensionality of the training data, since most of the spectral energy occupies a logarithmic grid of frequencies. In practice, Bock and Schedl [2] were able to reduce the dimensionality of their training data from 5120 FFT magnitudes down to 183.

Email addresses: nashj@umich.edu (Jeremy Nash), pschro@umich.edu (Paul Schroeder), markmliu@umich.edu (Mark Liu)

The logarithmic frequency scale also decreases sensitivty to detuning. In a linear frequency scale, detuning, especially in higher semitones, will cause spectral energy to leak into neighboring frequency bins, creating more variability in the training data. On a logarithmic frequency scale, minor detuning would result in less spectral energy leakage and less unnecessary variability in the training data.

5.5. Note onset detection

This method proposes restricting the problem to determining note onsets. This is a simpler problem than detecting note durations because of the possibility for time-varying timbres in instruments. Most researchers in this field evaluate their music transcription methods on their note onset detection performance.

5.6. Discriminative models

5.7. SVM

5.8. Feed forward neural network

5.9. Non-negative matrix factorization

Non-negative matrix factorization (NMF) is an unsupervised, dimensionality reduction technique that factors a non-negative matrix V with dimensions FxN into two matrices,

$$V \approx WH$$

where W is an FxK matrix and H is a KxN matrix. K is typically chosen so that $F * K + K * N << F * N$.

In piano transcription, for example, V might be a spectrogram with dimensions $2048xN$, W might be $2048x88$ and encode the spectral envelope for each piano key in each column of the matrix, and H might be $88xN$ sparse piano roll matrix.

There are two cost functions commonly used for NMF. The first is a euclidean distance cost function, defined as the squared difference between the corresponding elements in each matrix:

$$||V - WH||^2 = \sum_{ij}(V_{ij} - (WH)_{ij})^2$$

Another commonly used measure is the KL divergence:

$$D(V||WH) = \sum_{ij}(V_{ij} \log \frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij})$$

The KL divergence cost function is more often used by transcription researchers because it XXX.

Lee and Seung [3] Seung and Lee [4]

5.10. Sparse non-negative matrix factorization

One of the great side-effects of NMF is that it often produces a sparse, parts-based representation. This is desirable in music transcription because the piano roll matrix, H, is a sparse matrix of note activations. However, the NMF method does not contain an explicit sparsity objective. Hoyer [5] extended NMF to contain a tunable sparsity objective.

The sparse NMF method in Hoyer [5] was adopted by Abdallah and Plumbley [6] in their transcription system.

5.11. Bayesian nonparametric models

5.12. Smoothing

5.12.1. HMM smoothing

5.12.2. Probabilistic spectral smoothness

5.13. Recurrent neural network

6. Evaluation

There is no standard metric to evaluate the performance of a transcription method. However, many researchers use the following methods:

6.1. MIDI databases

6.2. Evaluation measures

6.2.1.

6.2.2. Frame-level transcription error score

6.3. Method comparison

6.3.1. Reported measures

Algorithm	Accuracy	Score
SVM	21%	21
SVM with HMM	22%	22
NMF	23%	23
Sparse NMF	24%	24

6.3.2. Our evaluation

The following table captures our evaluation of these methods. We expect these accuracy measures to be lower than they were reported in the original papers because we are using a smaller training dataset and not attempting to optimize the methods.

Algorithm	Accuracy	Score
SVM	18%	18
SVM with HMM	19%	19
NMF	20%	20
Sparse NMF	21%	21

7. Conclusion

8. Individual Effort

- **Jeremy** wrote and evaluted the SVM method in Poliner and Ellis [7], the Bayesian nonparametric method in Blei et al. [8], and the sparse non-negative matrix factorization method in Abdallah and Plumbley [6]. Jeremy also wrote the paper.

- **Mark** implemented the hidden Markov model in Poliner and Ellis [7] and the LSTM network in Bock and Schedl [2].

- **Paul** implemented the LSTM network in Bock and Schedl [2].

References

[1] S. S. Stevens, J. Volkmann, E. B. Newman, A scale for the measurement of the psychological magnitude pitch, The Journal of the Acoustical Society of America 8 (1937) 185.

[2] S. Bock, M. Schedl, Polyphonic piano note transcription with recurrent neural networks, in: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, IEEE, pp. 121–124.

[3] D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (1999) 788–791.

[4] D. Seung, L. Lee, Algorithms for non-negative matrix factorization, Advances in neural information processing systems 13 (2001) 556–562.

[5] P. O. Hoyer, Non-negative matrix factorization with sparseness constraints, The Journal of Machine Learning Research 5 (2004) 1457–1469.

[6] S. A. Abdallah, M. D. Plumbley, Polyphonic music transcription by non-negative sparse coding of power spectra, in: Proc. 5th Intl Conf. on Music Information Retrieval (ISMIR), pp. 10–14.

[7] G. E. Poliner, D. P. Ellis, A discriminative model for polyphonic piano transcription, EURASIP Journal on Advances in Signal Processing 2007 (2006).

[8] D. M. Blei, P. R. Cook, M. Hoffman, Bayesian nonparametric matrix factorization for recorded music, in: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 439–446.