GUITAR TRANSCRIPTION USING CONVOLUTION SPARSE CODING : A NEW DESIGN CONCEPT

Frank Cwitkowitz
University of Rochester
Electrical & Computer Engineering
500 Joseph C. Wilson Blvd.
Rochester, NY 14627

ABSTRACT

The automatic transcription of music is a notoriously challenging task which has received increasing attention in recent years. Most work has focused on transcribing solo piano recordings, of which there is an abundance of data. In contrast, the transcription of solo guitar recordings is often ignored, likely due to the severe lack of annotated data and the additional expressive dimensions of the instrument that must be modeled. In this work, we present a model-based approach which uses Convolutional Sparse Coding. In particular, we create a dictionary of waveform elements from real note observations within the training partition of a new guitar dataset. While the dictionary remains fixed, the activation of each element is computed, allowing us to generate latent note hypotheses for transcription. This information can also be used to separate the original audio by string. We finish by explaining how the method can be evaluated in terms of transcription accuracy and string separation quality, and by offering directions for future work.

1. INTRODUCTION

Transcription is the process by which the notation corresponding to music is realized through listening and understanding. Machines capable of Automatic Music Transcription (AMT) have applications beyond retrieving the notation for recordings. These include real-time instructional music scenarios which listen and provide feedback, mid-level music representations for database querying, or the improvement of methods for other music analysis problems. AMT approaches algorithmically recover the information sufficient to form a symbolic representation of the music inherent in an audio signal. Typically, notes are described with a fundamental frequency, an onset time, and optionally an offset time. AMT is a multi-faceted task which is challenging due to its complexity, overlap of harmonic energy from concurrent notes, variance in recording scenarios and instrumentation, lack of data, noise, etc. A good review of AMT can be found in [4] and [3].

Many instrument-specific approaches to music transcription have targeted solo piano music, of which there is an abundance of annotated data [6, 8] relative to solo guitar. This is also because the analysis of solo piano recordings is less complex than that of guitar, since a pianist can only control the duration and intensity of a note. Guitar recordings are harder to analyze, since the guitar offers more expressive and stylistic freedom. For instance, strings can be plucked at different points using alternate plucking styles. Notes can be played with slides, bends, hammer-ons or hammer-offs. Strings can also go out of tune when playing. There is plenty of interest in improving guitar transcription, but there has not been a large body of work similar to that of piano transcription.

In this work, we develop a novel approach to the automatic transcription of solo acoustic guitar music. A dictionary of waveforms is gathered from a training split of GuitarSet [13]. The dictionary elements correspond to the waveform that is expected when playing a note on each fret of each string. The element activations necessary to reproduce the observed signal are computed using Convolutional Sparse Coding (CSC) with sparsity and lateral inhibition [5]. For lateral inhibition, the groups are made up of all note observations across a string, including the various styles and durations which are included in the dictionary for each note.

The rest of the paper is as follows. We start by reviewing solo guitar transcription in Section 2. In Section 3, we outline the dictionary generation process, the CSC framework best suited for the computing activations, as well as our way of performing transcription and separation from the activations derived using CSC. In Section 4, we define the dataset and metrics used to evaluate our algorithm. Finally, we conclude in Section 5 and offer directions and insight for future work.

2. RELATED WORK

There have been several unique approaches to the transcription of solo guitar recordings. A simple peak-picking method with bio-feasibility constraints was proposed in [7]. A hidden Markov Model was used in [2] to estimate and model chord transitions based on detected fundamental frequencies and a learned musical model. Another early approach used basic filtering and signal processing techniques to estimate timing and pitch information jointly [1]. In [15] and [14], the results of a latent harmonic allocation algorithm are post-processing with bio-feasibility constraints, with the latter additionally using dynamic programming to weight possible fingerings in proportion to a player's estimated proficiency. In [10] non-negative matrix factorization is used to estimate the activations using a fixed dictionary of note spectra obtained during a preliminary performance. Another approach estimated plucking style, expression style, and the plucked string in addition to other timing and pitch information [9].

Most recently, a convolutional neural network architecture was proposed in [12], where baseline results were achieved on GuitarSet, the dataset chosen for our work [13]. One may argue that GuitarSet is quite homogeneous, and there is no clear split for training and testing. While still incredibly useful, GuitarSet is quite small compared to the datasets used for training state-of-the-art piano transcription models [8].

3. METHOD

3.1 Dictionary Acquisition

In order to create the dictionary, the dataset must be partitioned into training and testing splits. The waveforms from the hexaphonic pickup observed during the playing of notes on a given string are extracted from the training split. These are grouped by the string and fret that was held down to cause them. If the amount of examples for any string-fret combination is below a chosen threshold, the examples from the previous combination are pitch-shifted by one semitone and used to fill in the empty space. This assumes that there are always enough examples for the open note of each string, since we cannot go lower. Finally, a reduction step is performed, where the most dissimilar examples are chosen iteratively, until exceeding another fixed amount. In this work, dissimilarity is measured using the dot-product. Ideally, this gives a dictionary with the same amount of elements per string-fret combination, where each element represents a different duration and playing style of the note.

3.2 Fret-wise Activations

The dictionary element activations are estimated using the method presented in [5]. In this work, we group elements based on their string membership. Since we are dealing with guitar, only one element out of all belonging to every fret of one string can be active at once. Lateral inhibition is performed across the activation of elements within each group, and sparsity attempts to enforce that only one dictionary element be used to explain a single note during the factorization.

3.3 Inference

The inference step, which gives the final transcription and separation of a new solo guitar recording, can be performed by analyzing the latent activation space. The element which is maximally activated within a string group during a given frame can be said to comprise a note, if the activation exceeds a chosen threshold. The audio corresponding to each string can be reconstructed simply be recombining the dictionary elements and activations belonging to the string.

4. EXPERIMENTS

4.1 Dataset

GuitarSet, a recently proposed transcription dataset, features mono-channel and hex-channel recordings of experienced musicians playing solo guitar [13]. The dataset contains roughly 3 hours of approximately 30 second solo guitar recordings obtained using six different musicians playing a variety of genres. During each recording, a musician loosely plays an instructed chord progression. Then, a second version is created where the musician plays a second layer while listening to their previous recording. Both version are recorded with a mono-channel microphone and a hexaphonic pickup. The hexaphonic recording is subsequently post-processed with a debleeding algorithm, with produces the ground-truth audio by string used to evaluate separation quality. Ground-truth note annotations are provided in the dataset and used to evaluate transcription quality.

4.2 Metrics

We use the *mir_eval* package to evaluate both transcription quality and separation quality [11]. Relevant metrics for AMT include precision, recall, and F_1 score across all predictions. These metrics are used on a frame-based level and a note-based level. For frame-based evaluation, the frame-wise frequency activity inferred from the notes predictions are compared against the ground truth frame-wise frequency activity inferred from the ground-truth notes. The problem in this scenario is reduced to the detection of active frequencies across frames using multi-class binary classification. For note-based evaluation, we use two definitions for what constitutes a correct note. The most simple definition of a correct note prediction is one where the estimated fundamental frequency is within half a semitone interval of the true value, and there is a corresponding onset time estimation within 50 ms of the true value. Another more challenging definition builds upon the previous, but also requires that there exist a corresponding offset time estimation within the larger of either 50 ms or 20% of the ground truth duration. The metric used for separation is the average Signal-to-Distortion Ratio (SDR) across strings when comparing the inferred audio to the debleeded audio recorded with the hexaphonic pickup.

5. CONCLUSION

We presented a new design concept for the automatic transcription of solo guitar music. A dictionary is generated from a training partition of a solo guitar dataset, and CSC is used to obtain the corresponding activations. In the future, the method can be made to be unsupervised, by generating the dictionary iteratively using a parametric guitar model. The dictionary generation step can be improved by using clustering and nearest-neighbor to reduce the elements more effectively. In the future, a lot of effort must be devoted to tuning the process of CSC, as there are many parameters and considerations to make in this step.

6. REFERENCES

[1] Lance Alcabasa and Nelson Marcos. Automatic guitar music transcription. In *2012 International Conference on Advanced Computer Science Applications and Technologies (ACSAT)*, pages 197–202. IEEE, 2012.

[2] Ana M Barbancho, Anssi Klapuri, Lorenzo J Tardón, and Isabel Barbancho. Automatic transcription of guitar chords and fingering from audio. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3):915–921, 2011.

[3] Emmanouil Benetos, Simon Dixon, Zhiyao Duan, and Sebastian Ewert. Automatic music transcription: An overview. *IEEE Signal Processing Magazine*, 36(1):20–30, 2019.

[4] Emmanouil Benetos, Simon Dixon, Dimitrios Giannoulis, Holger Kirchhoff, and Anssi Klapuri. Automatic music transcription: challenges and future directions. *Journal of Intelligent Information Systems*, 41(3):407–434, 2013.

[5] Andrea Cogliati, Zhiyao Duan, and Brendt Wohlberg. Piano transcription with convolutional sparse lateral inhibition. *IEEE Signal Processing Letters*, 24(4):392–396, 2017.

[6] 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*, 18(6):1643–1654, 2009.

[7] Xander Fiss and Andres Kwasinski. Automatic real-time electric guitar audio transcription. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 373–376. IEEE, 2011.

[8] 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 preprint arXiv:1810.12247*, 2018.

[9] Christian Kehling, Jakob Abeßer, Christian Dittmar, and Gerald Schuller. Automatic tablature transcription of electric guitar recordings by estimation of score-and instrument-related parameters. In *DAFx*, pages 219–226, 2014.

[10] Masaki Otsuka and Tetsuro Kitahara. An on-line algorithm of guitar performance transcription using non-negative matrix factorization. In *2014 12th International Conference on Signal Processing (ICSP)*, pages 621–624. IEEE, 2014.

[11] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel. mir_eval: A transparent implementation of common mir metrics. In *ISMIR*, pages 367–372, 2014.

[12] Andrew Wiggins and Youngmoo Kim. Guitar tablature estimation with a convolutional neural network. In *ISMIR*, 2019.

[13] Qingyang Xi, Rachel M Bittner, Johan Pauwels, Xuzhou Ye, and Juan Pablo Bello. Guitarset: A dataset for guitar transcription. In *ISMIR*, pages 453–460, 2018.

[14] Kazuki Yazawa, Katsutoshi Itoyama, and Hiroshi G Okuno. Automatic transcription of guitar tablature from audio signals in accordance with player's proficiency. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3122–3126. IEEE, 2014.

[15] Kazuki Yazawa, Daichi Sakaue, Kohei Nagira, Katsutoshi Itoyama, and Hiroshi G Okuno. Audio-based guitar tablature transcription using multipitch analysis and playability constraints. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 196–200. IEEE, 2013.