TRANSCRIBING MULTI-INSTRUMENT POLYPHONIC MUSIC WITH TRANSFORMED EIGENINSTRUMENT WHOLE-NOTE TEMPLATES

Zhuo Chen
LabROSA, Columbia University
zc2204@columbia.edu

Graham Grindlay
LabROSA, Columbia University
grindlay@ee.columbia.edu

Daniel P.W. Ellis
LabROSA, Columbia University
dpwe@ee.columbia.edu

ABSTRACT

We present a system for the transcription of polyphonic music recordings to recover both the notes played and the instruments responsible for each note. In our framework, the spectrogram of the music is viewed as the superposition of note events, each characterized by an onset time and pitch, an instrument (described by a vector of eigeninstrument weights that combine instrument model bases to match a particular source in the mixture), and per-note transformation parameters that take a duration- and decay-normalized note template and extend it to match the actual duration and dynamics of each individual note. Transcription is achieved through an EM-like iterative estimation scheme. Initializing this estimation using a rough separation of sources from a frame-based transcription system gives stable and accurate results that directly describe the audio at a note, rather than a frame, level, with each note attributed to a particular instrument. This approach significantly improves transcription accuracy over a frame-level system, apparently because the transcription constrains each note to obey the dynamics encoded in the templates. Note-level transcription accuracy on real woodwind excerpts from the MIREX Multiple F0 evaluation improves from 64% (frame-level) to 67%; for the more temporally-structured notes in the RWC piano examples, accuracy improves from 70% to 79%, with dramatic reductions in false alarms.

Keywords: Polyphonic, Note-level transcription, Eigeninstruments

1. INTRODUCTION

Music transcription is one of the oldest and most heavily studied problems in Music Information Retrieval (MIR). Current work in this area is focused on two main areas: one is improving the signal processing and representation to better isolate notes as spectral peaks that can be associated with particular f_0 and instrument hypotheses (e.g., [1, 2]). The other thread is centered on machine learning, with systems that match unknown signals against labeled spectral templates that have been learned from training examples (e.g., [3–5]).

Almost all current systems are frame-based, performing the bottom level of analysis against individual spectral slices derived from short time frames. Since note events typically last for many frames, temporal continuity is introduced by some higher-level processing – most frequently via a hidden Markov model (HMM) [3, 4].

In this work, however, we perform transcription using note-level templates, which are made specific to individual instruments via an eigeninstrument decomposition. Such templates describe the behavior of each note across multiple frames, directly providing additional detail and constraints when compared to frame-by-frame analysis. This work brings together three techniques: first, different instruments in a recording of an ensemble are individually modeled as points in an eigeninstrument model space, as in [8]. The single-frame templates of that system are replaced by whole-note templates covering multiple frames, necessitating the introduction of convolutive NMF (CNMF, [6]). However, to allow notes that differ only in duration to map back to the same template, we introduce a parametric transformation step, controlling the spectral evolution and amplitude envelope. In addition to this basic model, we discussed how we can induce sparsity in the resulting solutions.

2. THE TRANSCRIPTION SYSTEM

2.1 Note-level eigeninstrument transcription system

The transcription system of [8] used the interpretation of a spectrogram as a distribution to perform instrument-dependent transcription using single spectral slices as templates, each formed from an eigeninstrument basis to match particular instruments in a recording. We extend this using note-level (multiple-slice) eigeninstrument templates:

$$\hat{P}(f,t) = \sum_{z,p,k,d,s,t_0,t_l} P(f|z,p,k)P(k|s)P(s)P(t_0|s) \cdot P(p|t_0,s)P(d|p,t_0,s)P(z,t_l|d,p,s)P(t_l,t|t_0) \quad (1)$$

$P(f|z,p,k)$ are the underlying note templates expressed as functions of a normalized time index z, one for every pitch p and eigenistrument basis k. s indexes the sources in the mixture, with $P(k|s)$ as the eigeninstrument coefficients matching that source. $P(p|t_0,s)$ is the "activation" of source s producing a note of pitch p starting at

time t_0, with $P(d|p,t_0,s)$ being the corresponding distribution across the set of transformations indexed by d, which directly specifies the duration of the note. Then, $P(z,t_l|d,p,s)$ provides the transformation, relating the normalized time index z to the true onset-relative time index t_l. $P(t_l,t|t_0) = \delta(t-t_0-t_l)$ provides the deterministic link between the time indices required for the convolution, and the remaining values are marginals needed to complete the equation.

2.1.1 From transformation to time-warping

Introducing transformations allows us to estimate transform-invariant templates for particular notes regardless of duration. However, estimating a full $P(z,t_l|d)$ for every d introduces a lot of parameters, leading to convergence problems and poor results. In fact, to connect real notes of different durations to a single template, we anticipate a fairly restricted range of mappings, corresponding to a time warp and an amplitude decay. We can thus devise a parametric form for the transformation matrix.

One possible model for musical notes consists of a relatively invariant onset, followed by a decay that depends on the overall duration of the note. Matching notes of varying durations can be accomplished by a simple time warp, but the warp needs to be nonlinear to provide nonuniformity across onset and decay; we use two parameters, described below. In addition to warping the timebase of each template, we also need a mechanism to achieve the decay in amplitude: the template is normalized in each time scale, and the note event $P(p|t_0,s)$ provides only a global amplitude scaling. Thus, we use a third parameter to control the amplitude decay. The parametric transformation is thus:

$$P(z,t_l|d,p,s) = P(z|d,t_l,s)\cdot P(t_l|d,p,s) \qquad (2)$$

Henceforth we will drop the dependence on source s, which is implicit in the choice of parameters. The first term provides the time warping. It has two parameters: γ controls how "sharply" the transitions between states occur, and a determines the overall warp within a duration d:

$$P(z|d,t_l) = \frac{1}{N_1}\exp\{-\gamma(z-4F_a(t_l))^2\} \qquad (3)$$

$$F_a(t_l) = \frac{1-a^{t_l/d}}{1-a} \qquad (4)$$

where $N_1 = \sum_z \exp -\gamma(z-4F_a(t_l))^2$ normalizes the result to a true distribution. See figure 1 for examples of transforms for several values of γ and a.

The second term provides an overall amplitude decay according to a single parameter, b:

$$P(t_l|d,p) = \frac{1}{N_2}\exp\{-F_b(p)\frac{t_l}{d}\} \qquad (5)$$

$$F_b(p) = -\ln(b) + (0.03b^3 - 0.12b^2 + 0.15b + 0.05)p \qquad (6)$$

where $N_2 = \sum_{t_l}\exp\{-F_b(p)t_l/d\}$ again normalizes the distribution. Blown and bowed instruments typically do not have an exponential decay of energy through the note, which results in a very small value for F_b (i.e., b close to 1). For plucked and struck instruments, even within a single instrument the decay time is generally inversely proportional to pitch; this is captured by the dependence on p of F_b, which is empirically fit to our data by the shown polynomial. Figure 1 also shows several values for the decay function, including two pitch values for each example.

2.1.2 Parameter estimation

We solve for the unknown terms in (1) and the parameters a, b and γ from (3)–(5) with an EM-like iterated algorithm. There are two steps in our algorithm: firstly, $P(d|p,t_0,s)$, $P(s)$, $P(t_0|s)$, $P(p|t_0,s)$ and $P(k|s)$ are updated following a standard EM procedure. In the E step, the expectation of hidden parameters $R(s,k,p,t_0,d|f,t)$ is calculated:

$$R(s,k,p,t_0,d|f,t) = \frac{1}{\hat{P}(f,t)}\sum_{z,t_l} P(f|z,p,k)P(k|s)P(s)P(t_0|s)P(p|t_0,s) \cdot P(d|p,t_0,s)P(z|d,t_l,s)P(t_l|d,p,s)P(t_l,t|t_0) \qquad (7)$$

The M step maximizes the complete likelihood:

$$\sum_{f,t} V(f,t)\sum_{s,k,t_0,p,d} R(s,k,p,t_0,d|f,t)log(\hat{P}(f,t)) \qquad (8)$$

giving the following update rules:

$$P(k|s) = \langle RV\rangle_{k|s} \qquad P(s) = \langle RV\rangle_{s} \qquad (9)$$

$$P(t_0|s) = \langle RV\rangle_{t_0|s} \qquad P(p|t_0,s) = \langle RV\rangle_{p|t_0,s} \qquad (10)$$

$$P(d|p,t_0,s) = \langle RV\rangle_{d|p,t_0,s} \quad P(f|z,p,k) = \langle RV\rangle_{f|z,p,k} \qquad (11)$$

In (9)–(11), $\langle\bullet\rangle_{x|y}$ means the summation over all variables except x and y, followed by the normalization on x. V is the spectrogram of the input signal.

In the second step, the parameters a, b and γ are found by maximizing likelihood

$$L = \sum_{f,t} V(f,t)log(\hat{P}(f,t)) \qquad (12)$$

through gradient descent:

$$a_s = a_s + \frac{\partial L}{\partial a_s} \quad b_s = b_s + \frac{\partial L}{\partial b_s} \qquad \gamma_s = \gamma_s + \frac{\partial L}{\partial \gamma_s} \qquad (13)$$

which converges in our experience after 30 iterations. Finally, note-level transcripts are obtained by thresholding

$$P(p,t_0|s) = P(t_0|s)P(p|t_0,s) \qquad (14)$$

Frame-level results can also be obtained by thresholding

$$P(p,t|s) = \sum_{z,f,k,d,t_0,t_l} P(f|z,p,k)P(k|s)P(t_0|s) \cdot P(p|t_0,s)P(d|p,t_0,s)P(z|d,t_l,s)P(t_l|d,p,s)P(t_l,t|t_0) \qquad (15)$$

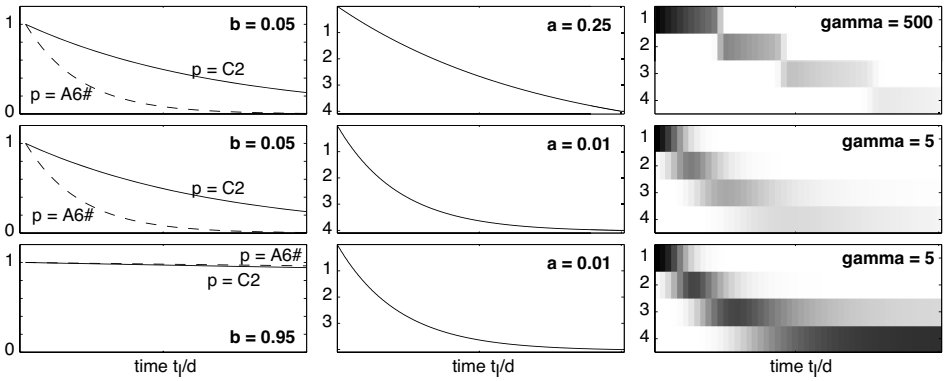

Figure 1. Examples of the parametric transformation matrix for several different values of γ, a, and b as indicated on the figure. The first column shows the decay envelope for two different pitches, the second column shows the time warp trajectory, and the third column shows the resulting transformation matrix $P(z, t_l|d, p, s)$.

2.2 Forming eigeninstruments

Following [8], we build a "super-vector" of instrument parameters for each training instrument using the templates learned by (11) and the parameters from (13). We then use NMF to extract a set of eigeninstrument model-space bases.

2.3 Sparsity

Since our system is note-based, we want each note represented by single note-template, rather than a combination of shorter ones. For example, if a note can be represented by the concatenation of ten short patches or single longer patch, we would prefer the latter since it has a much clearer physical meaning. We encourage this with sparsity constraints on three terms in our system: $P(d|p, s, t_0)$, $P(p|s, t_0)$, and $P(t_0|s)$. Sparsity on $P(d|p, s, t_0)$ discourages the overlapping of patches of different durations, which would otherwise tend to be equally-good matches at least for the onsets of each note. The sparsity on $P(p|s, t_0)$ tries to minimize the number of fundamental pitches used to explain a given set of harmonics (to avoid the introduction of notes rooted on harmonics). Sparsity on $P(t_0|s)$ simply prefers fewer onset events per instrument, to avoid the concatenation of shorter patches in the reconstruction.

[7] provides one route to sparsity. The update rules for $P(p|s, t_0)$ and $P(d|p, s, t_0)$ are changed to:

$$P(p|s, t_0) = \langle (RV)^{\alpha} \rangle_{p|s,t_0} \quad (16)$$

$$P(d|s, p, t_0) = \langle (RV)^{\beta} \rangle_{d|s,p,t_0} \quad (17)$$

Larger values of α and β above 1 result in sparser solutions. We found 1.2 to be a good value for both parameters.

For $P(t_0|s)$, we adopt the entropic prior from [9, 10]:

$$P(p_{s,t_0}|\beta) \propto \exp(\beta \sum_{t_0} p_{s,t_0} log(p_{s,t_0})) \quad \beta > 0 \quad (18)$$

Since the entropic prior is not conjugate to multinomial distribution, we follow [9] and obtain the final update rule for $P(t_0|s)$ by iterating the two relations:

$$h = \langle p_{s,t_0}^{\frac{\nu}{\nu-1}} \rangle_{t_0|s} \quad (19)$$

$$p_{s,t_0} = \langle \mu\nu h + \langle RV \rangle_{t_0|s} \rangle_{t_0|s} \quad (20)$$

where μ is a small value. We used $\nu = 50$, and set $\mu = 0.03$ for monophonic examples, or 0.02 for polyphonic transcription.

3. EXPERIMENTS

3.1 Data

Our eigeninstrument models were learned from synthesized instrument notes. Eight woodwind instruments (French Horn, Oboe, English Horn, Bassoon, Clarinet, PIccolo, Flute, Recorder) and two pianos (Acoustic Grand Piano, Bright Acoustic Piano) were synthesized by three soundfonts: *Papelmedia Final*, *FluidR3 GM* and *RealFont.2.1*. For each instrument, only the pitches within its natural playing range were used. We performed transcription on three kinds of material: piano excerpts, woodwind excerpts, and mixtures of piano and woodwind. We used two piano excerpts from RWC database [11], one synthesized piano excepts from J.S. Bach's Chromatic Fugue, and three woodwind excerpts from MIREX Multiple Fundamental Frequency Estimation and Tracking evaluation task, in both recorded and synthesized version. Mixtures were created by adding individual tracks. The *SGM V2:01* soundfont was used to produce the synthesized tracks. All tracks were down-sampled to 22kHz, and then a magnitude spectrogram was generated with a 1024-point STFT using a 46 ms Hamming window and 25% overlap. All experiments were repeated five times with random initialization.

4. REFERENCES

[1] K. Dressler, "Pitch estimation by the pair-wise evaluation of spectral peaks," *AES 42nd Conference*, Ilmenau, Germany, July 2011.

[2] Yeh, C., Roebel, A., Rodet, X., "Multiple fundamental frequency estimation and polyphony inference of poly- phonic music signals," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 6, August, 2010.

[3] Benetos, E., Dixon, S., "A temporally-constrained convolutive probabilistic model for pitch detection," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 133–136,New Paltz, NY 2011.

[4] E. Vincent and X. Rodet, "Music transcription with ISA and HMM," *Proc. 5th Int. Symp. on ICA and BSS (ICA04)*, Granada, Spain, 2004.

[5] E. Vincent, N. Bertin, and R. Badeau, "Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 109112, 2008.

[6] Smaragdis, P., "Convolutive Speech Bases and their Application to Supervised Speech Separation," *IEEE Tr. Audio, Speech and Lang. Proc.*, 15(1), 1–12, Jan. 2007.

[7] Smaragdis, P., B. Raj, and M. V. Shashanka, "Sparse and shift-invariant feature extraction from non-negative data," *IEEE International Conference on Audio and Speech Signal Processing*, Las Vegas, Nevada, USA. April 2008.

[8] G. Grindlay and D. Ellis, "Transcribing Multi-instrument Polyphonic Music with Hierarchical Eigeninstruments," *IEEE J. Sel. Topics in Sig. Process*, vol.5 no.6, pp. 1159–1169, October 2011.

[9] M. D. Hoffman, "Approximate Maximum A Posteriori Inference with Entropic Priors," *Technical Report*, September 2010. [Online]. Available: http://arxiv.org/abs/1009.5761

[10] R. J. Weiss and J. P. Bello, "Unsupervised Discovery of Temporal Structure in Music," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5(6), pp. 1240–1251, October 2011.

[11] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database:Classical Music," *Int. Conf. Music Information Retrieval*, Oct. 2003.

[12] Brett W. Bader, Tamara G. Kolda "MATLAB Tensor Toolbox Version 2.5" *January 2012 [Online]. Available http://www.sandia.gov/ tgkolda/TensorToolbox/*,