Multiple instrument tracking based on reconstruction error, pitch continuity and instrument activity

Holger Kirchhoff[1]*, Simon Dixon[1], and Anssi Klapuri[2]

[1] Centre for Digital Music, Queen Mary University of London, UK
[2] Ovelin, Helsinki, and Tampere University of Technology, Finland
{holger.kirchhoff, simon.dixon}@eecs.qmul.ac.uk, anssi@ovelin.com

Abstract. We present an algorithm for tracking individual instruments in polyphonic music recordings. The algorithm takes as input the instrument identities of the recording and uses non-negative matrix factorisation to compute an instrument-independent pitch activation function. The Viterbi algorithm is applied to find the most likely path through a number of candidate instrument and pitch combinations in each time frame. The transition probability of the Viterbi algorithm includes three different criteria: the frame-wise reconstruction error of the instrument combination, a pitch continuity measure that favours similar pitches in consecutive frames, and the activity status of each instrument. The method was evaluated on mixtures of 2 to 5 instruments and outperformed other state-of-the-art multi-instrument tracking methods.

Keywords: automatic music transcription, multiple instrument tracking, Viterbi algorithm

1 Introduction

The task of automatic music transcription has been studied for several decades and is regarded as an enabling technology for a multitude of applications such as music retrieval and discovery, intelligent music processing and large-scale musicological analyses [1]. In a musicological sense, a transcription refers to a manual notation of a music performance which can include the whole range of performance instructions ranging from notes and chords over dynamics, tempo and rhythm to specific instrument-dependent playing styles. In scores of Western music each instrument or instrument group is usually notated on its own staff.

Computational approaches to music transcription have mainly focussed on the extraction of pitch, note onset and note offset information from a performance (e. g. [2], [3], [4]). Only few approaches have addressed the task of additionally assigning the notes to their sound sources (instruments) in order to obtain a parts-based transcription. The transcription of individual instrument parts, however, is crucial for many of the above mentioned applications.

* This work was funded by a Queen Mary University of London CDTA studentship.

In an early paper, Kashino et al. [5] incorporated a feature-based timbre model in their hypothesis-driven auditory scene analysis system in an attempt to assign detected notes to instruments. Vincent and Rodet [6] combined independent subspace analysis (ISA) with 2-state hidden Markov models (HMM). Instrument spectra were learned from solo recordings and the method was applied to duet recordings. The harmonic-temporal clustering (HTC) algorithm by Kameoka et al. [7] incorporates explicit parameters for the amplitudes of harmonic partials of each source and thus enables an instrument-specific transcription. However, no explicit instrument priors were used in the evaluation and the method was only tested on single-instrument polyphonic material. Duan et al. [8], [9] proposed a tracking method that clusters frame-based pitch estimates into instrument streams based on pitch and harmonic structure. Grindlay and Ellis [10] used their eigeninstruments method as a more generalised way of representing instruments to obtain parts-based transcriptions.

The standard non-negative matrix factorisation (NMF) framework with instrument-specific basis functions is capable of extracting parts-based pitch activations. However, it only relies on spectral similarity and does not involve pitch tracking and other explicit modelling of temporal continuity. Bay et al. [11] therefore combined a probabilistic latent component analysis model and a subsequent HMM to track individual instruments over time.

In this paper we follow a similar approach as in [11]. We also employ the Viterbi algorithm to find the most likely path through a number of candidate instrument combinations at each time frame. However, we use a more refined method for computing the transition probabilities between the states of consecutive time frames. The proposed transition probability is based on the reconstruction error of each instrument combination and the continuity of pitches across time frames. Additionally we address the fact that one or more instruments might be inactive in any time frame by an explicit activity model. For this work we assume that all instruments are monophonic but the method could be extended to include polyphonic instruments.

The paper is structured as follows: In Section 2 we describe our multiple instrument tracking method. We explain the preliminary steps of finding candidate instrument combinations and illustrate the details of the Viterbi algorithm. Section 3 outlines the evaluation procedure and presents the experimental results. We conclude the paper in Section 4.

2 Multiple instrument tracking

2.1 Overview

Given the identities of the I instruments, we learn protoype spectra for the instruments in the mixture from a musical instrument database. These spectra are used as basis functions in an NMF framework in order to obtain pitch activations for each instrument individually. Instrument confusions are likely to happen in the NMF analysis. We therefore sum all instrument activations at the same pitch

into an overall pitch activation matrix from which we can obtain more reliable pitch information in each time frame.

In the resulting pitch activation matrix, we identify the P most prominent peaks in each time frame ($P \geq I$) and consider all possible assignments of each peak to each of the I instruments. For each of these instrument-pitch combinations, the reconstruction error is determined and the combinations are sorted in ascending order of their reconstruction error. The N combinations with the lowest reconstruction errors at each time frame are selected as candidates for the Viterbi algorithm. We then find the most likely path through the Viterbi state sequence by applying a transition probability function that takes into account the reconstruction error of each instrument combination, the pitch continuity as well as the fact that instruments might be inactive in each time frame.

2.2 Pitch activation function

To obtain the pitch activation function, we apply the non-negative matrix factorisation algorithm with a set of fixed instrument spectra on a constant-Q spectrogram and use the generalised Kullback-Leibler divergence as a cost function. The instrument spectra for each instrument type in the target mixture were learned from the RWC musical instrument database [12]. The constant-Q spectrogram was computed with a sub-semitone resolution of 4 bins per semitone, and in order to detect pitch activations with the same resolution, additional shifted versions of the instrument spectra up to $\pm$ 0.5 semitones were employed.

The NMF analysis with instrument-specific basis functions actually provides instrument-specific pitch activation functions, however, we realised that instrument-confusions do occur occasionally which introduce errors at an early stage. We therefore compute a combined pitch activation matrix by summing the activations of all instruments at the same pitch, which provides more reliable estimates of the active pitches. It should be pointed out here that numerous other ways of computing pitch activations have been proposed (e. g. [3]) which might equally well be used for the initial pitch analysis.

2.3 Instrument combinations

From the pitch activation function, the P highest peaks are extracted and all assignments of peaks to instruments are considered. To make this combinatorial problem tractable we make the assumptions that each instrument is monophonic and that no two instruments will play the same pitch at the same time. An extension to polyphonic instruments is discussed in Section 4. The total number of pitch-to-instrument assignments is given by

$$C(P, I) = \frac{P!}{(P - I)!}, \tag{1}$$

where P denotes the number of extracted peaks per frame and I the number of instruments. Depending on both P and I, this can lead to a large number of

combinations. In practice, however, we can discard all combinations for which a peak lies outside the playing range of one of the instruments. In our experiments this reduced the number of combinations considerably. If *all* peaks lie outside the range of an instrument, however, the case in which the instrument is inactive has to be included.

In order to determine the reconstruction error for each instrument-pitch combination we computed another NMF with fixed instrument spectra. Here, only a single spectrum per instrument at the assigned pitch was used and we applied only 5 iterations of the NMF update rules for the gains. Due to the small number of basis functions and iterations, this can be computed reasonably fast. Given the reconstruction errors for each combination at each time frame, we select the N combinations with the lowest reconstruction errors as our candidate instrument-pitch combinations. The gains obtained from these NMF analyses are used for the activity modelling as described in the following section.

2.4 Viterbi algorithm

We employ the Viterbi algorithm to find the most likely sequence of instrument-pitch combinations over time. A general description of the Viterbi algorithm can be found in [13]. In our framework, a state j at time frame t can mathematically be described as $S_{j,t} = (\phi_{j,t,i}, a_{j,t,i})$ with $i \in \{1, \ldots, I\}$. In this formulation, $\phi_{j,t,i}$ denotes the pitch of instrument i and $a_{j,t,i}$ is a binary activity flag that indicates whether the instrument is active at that time frame. The observed gain values for the instruments i of a state $S_{j,t}$ are denoted by $g_{j,t,i}$ and the reconstruction error of the state is given by $e_{j,t}$.

The states of the Viterbi algorithm are obtained by considering all possible combinations of instruments being active ($a_{j,t,i} = a$) and inactive ($a_{j,t,i} = \overline{a}$) for each of the selected instrument-pitch combinations from Section 2.3. These can be seen as activity hypotheses for each combination. Note that in this process, a large number of duplicates are produced when the pitches of all active instruments agree between the selected instrument-pitch combinations. As an example, consider a 2-instrument mixture with the following two candidate instrument-pitch combinations at time t: $(\phi_{1,t,1} = x, \phi_{1,t,2} = y)$ and $(\phi_{2,t,1} = x, \phi_{2,t,2} = z)$. The activity hypothesis in which $a_{1,t,1} = a_{2,t,1} = a$ and $a_{1,t,2} = a_{2,t,2} = \overline{a}$ produce identical Viterbi states, that both assume that instrument 1 is responsible for pitch x and that instrument 2 is inactive. For all identical Viterbi states, we only consider the one with the lowest reconstruction error $e_{j,t}$.

For the transition probability from state $S_{k,t-1}$ at the previous frame to state $S_{j,t}$ at the current frame, we consider 3 different criteria:

1. States with lower reconstruction errors $e_{j,t}$ should be favoured over those with higher reconstruction errors. We therefore model the reconstruction error by a one-sided normal distribution with zero mean: $p_e(e) = \mathcal{N}(0, \sigma_e^2)$ (Fig. 1a), where σ_e is set to a value of 10^{-3}.
2. We employ a pitch continuity criterion in the same way as [11]: $p_d(\phi_{j,t,i}|\phi_{k,t-1,i}) = \mathcal{N}(0, \sigma_d^2)$, with $\sigma_d = 10$ semitones (see Fig. 1b). Large

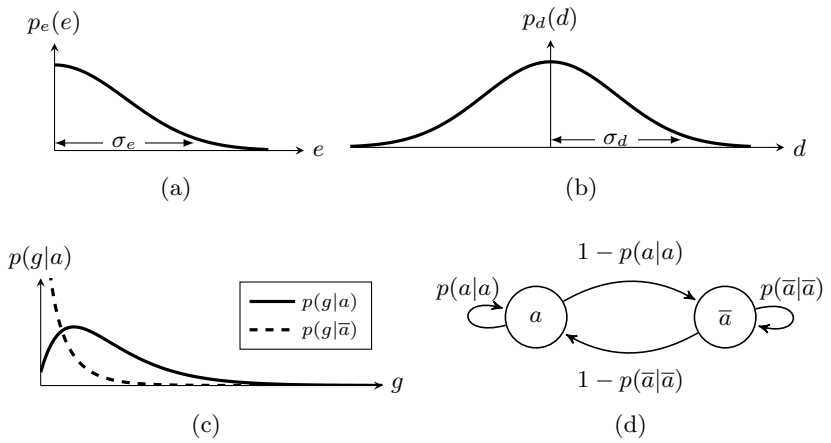

Fig. 1. Components of the transition probability for the Viterbi algorithm.

jumps in pitch are thereby discouraged while continuous pitch values in the same range in successive frames are favoured. This probability accounts for both the within-note continuity as well as the continuity of the melodic phrase.

3. An explicit activity model is employed that expresses the probability of an instrument being active at frame t given its gain at frame t and its activity at the previous frame. With Bayes rule, this probability can be expressed as

$$p_a(a_{j,t,i}|g_{j,t,i}, a_{k,t-1,i}) \quad = \quad \frac{p(g_{j,t,i}|a_{j,t,i}, a_{k,t-1,i}) \cdot p(a_{j,t,i}|a_{k,t-1,i})}{p(g_{j,t,i}|a_{k,t-1,i})}. \quad (2)$$

We furthermore assume that the gain only depends on the activity status at the same time frame and obtain the simpler form

$$p_a(a_{j,t,i}|g_{j,t,i}, a_{k,t-1,i}) = \frac{p(g_{j,t,i}|a_{j,t,i}) \cdot p(a_{j,t,i}|a_{k,t-1,i})}{p(g_{j,t,i})}. \quad (3)$$

We model the probability $p(g_{j,t,i}|a_{j,t,i})$ by two Gamma distributions with shape and scale parameters (2.02, 0.08) for active frames and (0.52, 0.07) for inactive frames. These distributions are illustrated in Fig. 1c. The probability $p(a_{j,t,i}|a_{k,t-1,i})$ for transitions between active and inactive states is illustrated in Fig. 1d. In this model, $p(a|a)$ was set to 0.986 and $p(\overline{a}|\overline{a})$ was set to 0.976 at the given hopsize of 4 ms. The term $p(g_{j,t,i})$ can be discarded in the likelihood function as it takes on the same value for all state transitions to state j at time t.

All parameter values above were obtained from analyses of the test set. Even though the parameters could have been obtained from other data sources, we believe that the parameter values and distributions and are reasonably generic to obtain similar results on other datasets.

Based on these criteria the overall log transition probability from state $S_{k,t-1}$ at time $t-1$ to state $S_{j,t}$ at time t can be formulated as

$$\ln(p(S_{j,t}|S_{k,t-1})) = \sum_{i=1}^{I} \ln[p(g_{j,t,i}|a_{j,t,i})] + \ln[p(a_{j,t,i}|a_{k,t-1,i})] + \sum_{\substack{\{i|a_{j,t,i}= \\ \{a_{k,t-1,i}=a\}}} \ln[p_d(\phi_{j,t,i}|\phi_{k,t-1,i})] + \ln[p_e(e_{j,t})] \quad (4)$$

3 Evaluation

3.1 Dataset

The multi-instrument note tracking algorithm described above was evaluated on the development dataset for the *MIREX Multiple fundamental frequency & estimation task*[3], which consists of a 54 s excerpt of a Beethoven string quartet arranged for wind quintet. We created all mixtures of 2 to 5 instruments from the separate instrument tracks, which resulted in 10 mixtures of 2 and 3 instruments, 5 mixtures with 4 instruments and a single mixture containing all 5 instruments. A MIDI file associated with each individual instrument provides the ground truth note data, that is, the pitch, onset time and offset time of each note.

3.2 Metrics

For the evaluation we did not use the common multiple-f0 estimation metrics because these do not take into account the instrument label of a detected pitch. Instead, we employed the same metrics as in the *MIREX Audio Melody Extraction* task, which evaluates the transcription of individual voices[4].

The metrics are frame-based measures and contain a voicing detection component and a pitch detection component. The voicing detection component compares the voice labels of the ground truth to those of the algorithmic results. Frames that are labelled as *voiced* or *unvoiced* in both the ground truth and the estimate are denoted as *true positives* (TP) and *true negatives* (TN), respectively. If labels differ between ground truth and estimate, they are denoted as *false positives* (FP) or *false negatives* (FN). The pitch detection component only looks at the *true positives* and measures how many of the pitches were correctly detected. Correctly detected pitches are denoted by TPC, incorrect pitches by TPI, with $TP = TPC + TPI$.

[3] available from: `http://www.music-ir.org/evaluation/MIREX/data/2007/multiF0/index.htm`

[4] MIREX 2012 Audio melody extraction task, `http://www.music-ir.org/mirex/wiki/2012:Audio_Melody_Extraction#Evaluation_Procedures`

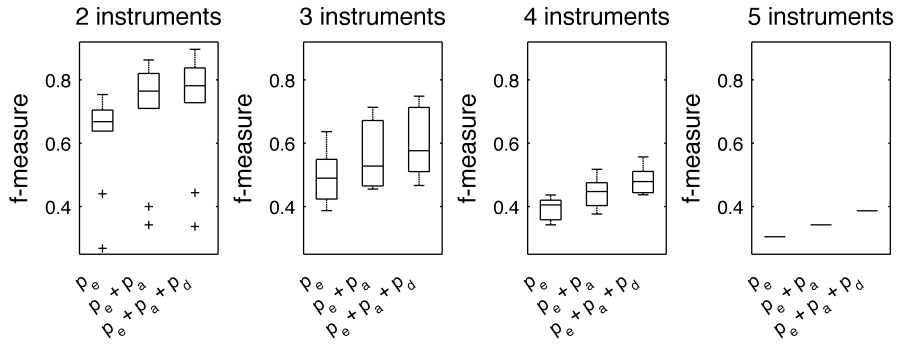

Fig. 2. Experimental results of the Viterbi note tracking method for different combinations of the transition probability components.

From these numbers, precision, recall and f-measure are computed in the following ways:

$$\text{precision} = \frac{\sum_{i=1}^{I}\sum_{t=1}^{T} TPC_{i,t}}{\sum_{i=1}^{I}\sum_{t=1}^{T} TP_{i,t} + FP_{i,t}} \tag{5}$$

$$\text{recall} = \frac{\sum_{i=1}^{I}\sum_{t=1}^{T} TPC_{i,t}}{\sum_{i=1}^{I}\sum_{t=1}^{T} TP_{i,t} + FN_{i,t}} \tag{6}$$

$$\text{f-measure} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}, \tag{7}$$

The precision measure indicates what percentage of the detected pitches were correct whereas the recall measure specifies the number of correctly detected pitches in relation to the overall number of correct pitches in the ground truth.

3.3 Results

The results were computed for each file in the test set individually and are here reported for each polyphony individually. We were also interested in the contributions of the different parts of the transition probability (Eq. 4) on the results, that is, the reconstruction error, the activity detection part as well as the pitch continuity criterion. To that end we first computed the results by using only the probability of the reconstruction error p_e, then we added the activity detection part p_a and finally we also considered the pitch continuity part p_d. Figure 2 shows the results as boxplots. Each boxplot summarises the results for all the instrument mixtures at a specific polyphony level.

When we successively combine the different parts of the transition probability, an increase in performance is apparent. Adding the activity detection part p_a (see middle plot in each panel) consistently improves the f-measure and likewise adding the pitch continuity criterion p_d (right plot in each panel) leads to

another leap in performance. Both parts roughly contribute the same amount of performance improvement. The activity detection part mainly improves the precision measure because it considerably reduces the false positives (FP) rate. The pitch continuity part on the other hand improves the correct true positives (TPC) and thus affects both precision and recall. The median f-measure goes up to 0.78 for the 2-instrument-mixture, to 0.58 for mixtures of 3 instruments and up to 0.48 and 0.39 for 4 and 5 instrument mixtures, respectively.

In terms of the absolute performance of the tracking method, we compared our results to the results reported in [10] and [11]. These methods likewise apply their algorithms to the wind quintet dataset. The results in [10] were computed on the same dataset, however, ground truth data was only available for the first 22 seconds of the recording. The authors in [11] reported their results on other excerpts from the wind quintet recording that are not publicly available, and five 30s excerpts were used in the evaluation. A comparison of the results can be found in table 1. Both algorithms use the same metrics as the ones described above, and report the *mean* of the results for the different instrument mixtures. To enable a comparison, we likewise compute the *mean* values of our results in table 1. Note that these values differ slightly from the *median* values in the boxplots in Fig. 2.

Table 1. Comparison of average f-measure with other multi-instrument tracking methods on similar datasets.

	2 instr.	3 instr.	4 instr.	5 instr.
Grindlay et al. [10]	0.63	0.50	0.43	0.33
Bay et al. [11]	0.67	**0.60**	0.46	0.38
Viterbi tracking	**0.72**	**0.60**	**0.48**	**0.39**

The comparison shows that the proposed algorithm outperforms the previous methods at almost all polyphony levels. While the results are only slightly better than the results reported in [11], the difference compared to the method proposed [10] is significantly larger. In [10], pitch activations were thresholded and no temporal dependencies between pitch activations were taken into account which underlines the fact that both an explicit activity model as well as a pitch continuity criterion are useful improvements for instrument tracking methods.

4 Conclusion

In this paper we presented an algorithm that tracks the individual voices of a multiple instrument mixture over time. After computing a pitch activation function, the algorithm identifies the most prominent pitches in each time frame and considers assignments of these pitches to the instruments in the mixture. The reconstruction error is computed for all candidate pitch-instrument combinations. Those combination with the lowest reconstruction errors are combined

with instrument activity hypotheses to form the states of a Viterbi framework in order to find the most likely sequence of pitch-instrument combinations over time. The transition probabilities for the Viterbi algorithm are defined based on three different criteria: the reconstruction error, pitch continuity across frames and an explicit model of active and inactive instruments.

The evaluation results showed that the algorithm outperforms other multi-instrument tracking methods which indicates that the activity model as well as the pitch continuity objective are useful improvements over systems which are based solely on the reconstruction error of the spectrum combinations.

Although in this paper we restricted the instruments to be monophonic, the method could be extended to incorporate polyphonic instruments. In this case a maximum number of N simultaneous notes would have to be specified for each polyphonic instrument. Instead of assigning each peak of the pitch activation function to a *single* instrument, we would allow up to N peaks of the pitch activation function to be assigned to the polyphonic instrument. If the number of simultaneously played notes of the polyphonic instrument remains constant over time, the Viterbi algorithm would combine the notes closest in pitch into individual note streams associated with the polyphonic instrument. If the polyphony increases, one or more of the inactive note streams would transition from a rest state to an active state. In the same way, if the polyphony decreases, one or more of the active streams would transition to a rest state.

A potential improvement could address the complexity of the method, that is, reducing the number of peak-to-instrument assignments which leads to a high computational cost for larger polyphonies. Instead of allowing each peak to be assigned to each instrument, peaks could be assigned to a subset of instruments only based on the highest per-instrument pitch activations in the initial NMF analysis.

References

1. Klapuri, A., Davy, M.: Signal Processing Methods for Music Transcription. Springer (2006)
2. Goto, M.: A Real-Time Music-Scene-Description System: Predominant-F0 Estimation for Detecting Melody and Bass Lines in Real-World Audio Signals. Speech Communication 43(4), 311–329 (2004)
3. Klapuri, A.: Multiple Fundamental Frequency Estimation by Summing Harmonic Amplitudes. In: Proceedings of the 7th International Conference on Music Information Retrieval, pp. 216–221 (2006)
4. Yeh, C., Roebel, A., Rodet, X.: Multiple Fundamental Frequency Estimation and Polyphony Inference of Polyphonic Music Signals. IEEE Transactions on Audio, Speech, and Language Processing 18(6), 1116–1126 (2010)
5. Kashino, K., Nakadai, K., Kinoshita, T., Tanaka, H.: Organization of Hierarchical Perceptual Sounds. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, pp. 158–164 (1995)
6. Vincent, E., Rodet, X.: Music transcription with ISA and HMM. In: 5th International Conference on Independent Component Analysis and Blind Signal Separation, pp. 1197–1204, Springer (2004)

7. Kameoka, H., Nishimoto, T., Sagayama, S.: A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering. IEEE Transactions on Audio, Speech, and Language Processing 15(3), 982–994 (2007)
8. Duan, Z., Han, J., Pardo, B.: Harmonically informed multi-pitch tracking. In: Proceedings of the 10th International Conference on Music Information Retrieval, pp. 333–338, Kobe, Japan (2009)
9. Duan, Z., Han, J., Pardo, B.: Song-level multi-pitch tracking by heavily constrained clustering. In: Proceedings of the 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 57–60 (2010)
10. Grindlay, G., Ellis, D.P.W.: Transcribing multi-instrument polyphonic music with hierarchical eigeninstruments. IEEE Journal of Selected Topics in Signal Processing 5(6), 1159–1169 (2011)
11. Bay, M., Ehmann, A., Beauchamp, J., Smaragdis, P., Downie, J.S.: Second fiddle is important too: Pitch tracking individual voices in polyphonic music. In: Proceedings of the 13th International Conference on Music Information Retrieval (ISMIR), pp. 319–324, Porto, Portugal (2012)
12. Goto, M., Hashiguchi, H., Nishimura, T., Oka, R.: RWC Music Database: Popular, Classical, and Jazz Music Databases. In: Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR), pp. 287–288, (2002)
13. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)