A Connectionist Model of Finding Partial Groups in Music Recordings with Application to Music Transcription

Matija Marolt

Faculty of Computer and Information Science, University of Ljubljana, Slovenia
E-mail: matija.marolt@fri.uni-lj.si

Abstract

In this paper, we present a technique for tracking groups of partials in musical signals, based on networks of adaptive oscillators. We show how synchronization of adaptive oscillators can be utilized to detect periodic patterns in outputs of a human auditory model and thus track stable frequency components (partials) in musical signals. We present the integration of the partial tracking model into a connectionist system for transcription of polyphonic piano music. We provide a short overview of our transcription system and present its performance on transcriptions of several real piano recordings.

1 Introduction

Music transcription could be defined as a process of converting an audio signal into a note-level (parametric) representation, where notes (pitches), their starting times and durations are extracted from the signal. Transcription is a challenging problem for current computer systems; separating notes from a mixture of other sounds, which may include other notes played by the same or different instruments or simply background noise, requires robust algorithms with performance that should degrade gracefully when noise increases.

Automatic transcription of polyphonic music would be useful in a variety of applications, ranging from content-based retrieval of music (i.e. query by example systems) and music analysis systems to accompaniment systems and musicological studies.

In recent years, several transcription systems have been developed [1-4]. All authors, except for Bello [2], base their systems on frequency domain analysis of the musical signal. Cues, such as local energy maxima, are extracted from the time-frequency representation of the signal and used in subsequent processing stages to find notes that are present in the signal. Various techniques, such as statistical frameworks, blackboard architectures, distance metrics or ICA are used in the process of grouping the found cues into notes, relying on information such as harmonicity and common onset/offset times. To reduce the complexity of the TF representation, to reduce noise and to incorporate some kind of temporal processing, partial tracking has been used in some systems to locate stable frequency components in the audio signal [4,5].

In this paper, we present a connectionist approach to music transcription. Transcription is a challenging task, so we limited the domain of our system to transcription of polyphonic piano music. The paper focuses on our approach to partial tracking with networks of adaptive oscillators, provides a short description of our entire transcription system, and presents some results obtained on transcriptions of real piano recordings.

2 Partial tracking with networks of adaptive oscillators

A melodic sound can be roughly described as a sum of components with relatively stable frequencies and time-varying amplitudes, called partials. By finding partials in a signal, one isolates the stable frequency components most likely belonging to tones, and discards noisy components. This is especially desirable in transcription systems, where the goal is to find all the tones present in the audio signal. Currently, most partial trackers used in transcription systems are based on a procedure similar to the tracking phase vocoder [6], where peaks are computed in each frame of the time-frequency representation. Detected peaks are then linked over time according to intuitive criteria such as proximity in frequency and amplitude to form partial tracks. Such approach is quite susceptible to errors in the peak peaking procedure, where missed or spurious peaks can lead to fragmented or spurious partial tracks.

We propose an alternative partial tracking approach that is not based on the standard peak-picking/peak connecting paradigm, but on connectionist principles. It is composed of two parts: an auditory model, and adaptive oscillators that extract partials from outputs of the auditory model.

2.1 Auditory Model

The auditory model emulates the functionality of human ear and transforms the audio signal into a probabilistic representation of firing activity in the auditory nerve. Amongst the several auditory models available, we chose to use a combination of the Patterson-Hodsworth gammatone filterbank [7] and Meddis' model of hair cell

transduction [8], as their implementations are readily available. The gammatone filterbank emulates the movement of basilar membrane in the inner ear. Its outputs are processed by the hair cell model, which converts each output into a probabilistic representation of firing activity in the auditory nerve. Its operations are based on a biological model of the hair cell and it simulates several of the cell's characteristics, most notably half-wave rectification, saturation and adaptation. Saturation and adaptation are very important to our model, as they reduce the dynamic range of the signal, and in turn enable our partial tracking model to track partials with low amplitude. These characteristics can be observed in Fig. 1, displaying outputs of three gammatone filters and the hair cell model on the 1., 2., and 4. partial of piano tone F3 (pitch 174 Hz).

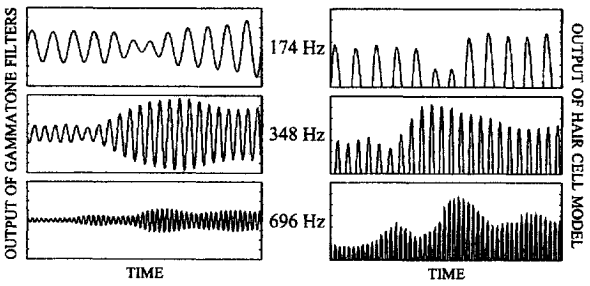

Fig. 1. Auditory analysis of three partials of piano tone F3.

2.2 Partial Tracking

Output of the auditory model consists of a set of quasi-periodic functions describing firing activities of inner hair cells in different parts of the basilar membrane (Fig. 1). Temporal models of pitch perception are based on the assumption that detection of periodicity in output channels of the auditory model forms the basis of human pitch perception. Periodicity is usually calculated with autocorrelation, resulting in a three-dimensional time-frequency representation of the signal called autocorrelogram, with time, channel center frequency and autocorrelation lag represented on orthogonal axes. In contrast, we propose to use a set of adaptive oscillators to estimate periodicity in output channels of the auditory model.

An oscillator is a system with periodic behavior. It oscillates in time according to its two internal parameters: phase and frequency. An adaptive oscillator adapts its phase and frequency in response to its input (driving) signal. When a periodic signal is presented to an adaptive oscillator, it adjusts its phase and frequency to match that of the driving signal and thus synchronizes with the signal. By observing the frequency and phase of a synchronized oscillator, an accurate estimate of the frequency and phase of its driving signal can be made. After reviewing several models, we decided to use a modified version of the Large-Kolen adaptive oscillator [9] in our partial tracking model.

The rationale behind the use of adaptive oscillators for partial tracking is simple. As periodicity in an output channel of the auditory model points to the presence of a frequency component (partial) in the input signal, analysis of periodicity in the channel indicates the exact frequency of the partial. In our model, periodicity is detected by a set of adaptive oscillators. If these synchronize with their stimuli (outputs of the auditory model), this indicates that the stimuli are periodic, and consequently that partials are present in the input signal. Frequencies of partials can be estimated by observing the frequencies of synchronized oscillators. Such a model has several advantages, when compared to standard approaches: it produces a continuous estimate of partials in a signal; because oscillators constantly adapt to their stimuli, partials with slowly changing frequencies (vibrato...) can be tracked; and as the auditory model reduces the dynamic range of the input signal and thus boosts partials with low amplitudes, these can be tracked as well.

2.3 Oscillator networks

As most tones are harmonic, we extended the model of tracking individual partials to a model of tracking groups of harmonically related partials by joining adaptive oscillators into fully-connected networks. Each network contains oscillators that track a series of harmonically related partials, so the frequencies of oscillators in a network are set to integer multiples of the frequency of the first oscillator (Fig. 2). As each oscillator in the network tracks a single partial close to its initial frequency, a network of oscillators tracks a group of harmonically related partials, which may belong to one tone with pitch equal to the frequency of the first oscillator.

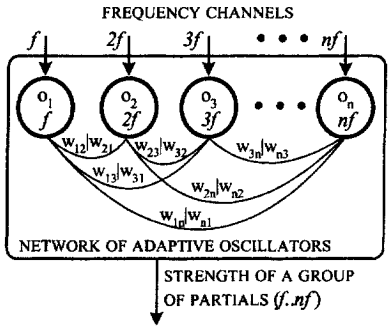

Fig. 2. A network of adaptive oscillators.

Within a network, each oscillator is connected to all other oscillators with excitatory connections. These connections are used to adjust frequencies and outputs of non-

synchronized oscillators in the network with the goal of speeding up their synchronization. Only a synchronized oscillator can affect frequencies and outputs of other oscillators in the network. Output of a network is calculated as a weighted sum of outputs of individual oscillators in the network and represents the strength of a group of partials tracked by oscillators in the network.

Connecting oscillators into networks has several advantages if the goal is to obtain a compact representation of a signal, suitable for transcription. Output of a network represents the strength of a group of harmonically related partials tracked by its oscillators. Such output provides a better indication of presence of a harmonic tone in the input signal than do outputs of individual oscillators. Noise usually doesn't appear in the form of harmonically related frequency components, so networks are more resistant to noise and provide a clearer time-frequency representation. Network connections are used by synchronized oscillators to speed up synchronization of non-synchronized oscillators, leading to a faster network response and faster discovery of a group of partials. Missing partials (even missing fundamental) are tolerated, if enough partials are found by other oscillators in the network.

An example is given in Fig. 3, which displays slices taken from three time-frequency representations of piano chord C3E3B4, calculated 100 ms after the onset: representation with uncoupled oscillators, repre-sentation with networks of adaptive oscillators and short-time Fourier transform. The representation with uncoupled oscillators was calculated with 88 oscillators tuned to pitches of piano tones A0-C8. Oscillator outputs (independent of partial amplitudes) are presented in Fig. 3A. Fig. 3B shows outputs of 88 oscillator networks, tuned to the same pitches. Product of networks' outputs and amplitudes of partials is shown in Fig. 3C. Fig. 3D displays the first 440 frequency bins of the Fourier transform calculated with a 100 ms Hamming window.

Individual oscillators have no difficulty in finding the first few partials of all tones (A). Some of the higher partials are not found, as they are masked by louder partials of other tones (we use only one oscillator per semitone). Oscillator networks (B) produce a clearer representation of the signal; the first two or three partial groups of each tone stand out. Networks coinciding with tones E3 and B4 produce the highest outputs, because almost all partials in the networks are found. When amplitudes are combined with network outputs (Fig. 3C), only four partial groups stand out, corresponding to first partials of all three tones (C3, E3, B4) and the second partial of tone E3. If we compare Fig. 3C with the Fourier transform in 3D, advantages of partial group tracking for transcription are obvious.

Overall, oscillator networks produce a compact and clear representation of partial groups in a musical signal. The main problem of this representation lies in occasional slow synchronization of oscillators in networks, which can lead to delayed discovery of partial groups. This is especially true at lower frequencies, where delays of 40-50 ms are quite common, because synchronization only occurs once per oscillator cycle; an oscillator at 100 Hz synchronizes with the signal every 10 ms, so several 10s of milliseconds are needed for synchronization. Closely spaced partials may also slow down synchronization, although it is quite rare for a group of partials not to be found.

3 Transcription of piano music

The described partial tracking model has been incorporated into our system for transcription of piano music, called SONIC [10]. Next to partial tracking, the system also includes a note recognition module, an onset detector based on a network of integrate-and-fire neurons, a module for resolving repeated notes, based on multilayer perceptrons and simple algorithms for estimation of tuning, note length and loudness.

A note recognition module is the central part of every transcription system. Its input consists of a set of cues extracted from the time-frequency representation of the input signal and its task is to associate the found cues with notes. Statistical methods are frequently used for this task; in our transcription system the task is performed by a set of neural networks. Inputs of each network are taken from outputs of the partial tracking module presented in

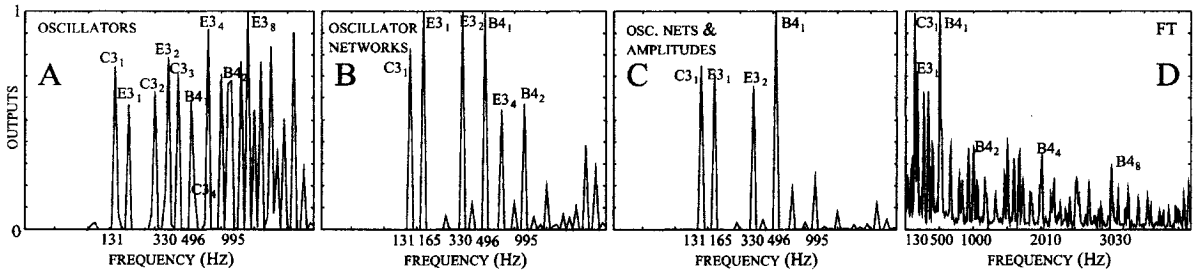

Fig. 3. Representations of piano chord C3E3B4

previous sections. Each network is trained to recognize one piano note in its input; i.e. one network is trained to recognize note A4, another network recognizes note G4... After extensive testing of several neural network models, we decided to use time-delay neural networks (TDNNs) in our system, as they provided the best performance. Networks were trained and tested on a database of approx. 150 synthesized piano pieces of various styles, combined with randomly generated chords. Tests showed that the incorporation of the partial tracking module significantly improved overall accuracy of transcription, halving the number of spurious notes [10].

4 Performance analysis

To analyze the performance of our transcription system, we tested it on a number of synthetisized and real recordings. The real recordings were transcribed by hand with the help of the original score. Originals and transcriptions of several pieces can be found on http://lgm.fri.uni-lj.si/SONIC. Table 1 lists performance statistics of three real piano performances: percentages of correctly found and spurious notes in transcriptions, as well as percentages of octave errors for missed and spurious notes are given.

	corr. notes	spur. notes	missed octave err.	spurious octave err.
1	88.5	15.5	35.1	80.5
2	68.3	13.6	30.3	79
3	85.9	15.2	70.3	87.4

Table 1. Performance statistics on real recordings

The transcribed recordings are: (1) J.S. Bach, English suite no. 5, 1st mvt, performer Murray Perahia, Sony Classical SK 60277; (2) F. Chopin, Nocturne no. 2, Op. 9/2, perf. Artur Rubinstein, RCA 60822; (3) S. Joplin, The Entertainer, performer unknown, MCA 11836.

Overall, most transcription errors are either due to octave errors or errors related to missed or spurious repeated notes. In Bach's English Suite, next to octave and repeated note errors, most of the missed notes are either quiet low pitched notes or notes in arpeggios and thrills. Chopin's Nocturne is a good example of very expressive playing, where a distinctive melody is accompanied by quiet, sometimes barely audible left hand chords. The system therefore misses over 30% of all notes, but even so the resynthesized transcription sounds quite similar to the original (listen to the example on the aforementioned URL address). When we compared transcriptions of the real and a synthesized version of The Entertainer, both turned out to be very similar. Transcription of the real recording contains more spurious notes, mostly occurring because of pedaling, which was not used in the synthesized version. The number of correctly found notes is almost the same in both pieces. Octave errors are the main cause of both missed and spurious notes. For a more detailed analysis, see [10].

5 Conclusion

In this paper, we presented a connectionist approach to partial tracking in musical signals. Our approach is based on a human auditory model and on adaptive oscillators for discovery and tracking of partial groups. By using a connectionist approach, we avoided some of the pitfalls of classical partial tracking approaches. We presented a brief overview of our transcription system and presented performance statistics on transcriptions of several real piano recordings. Overall, results are very promising and we believe that connectionist approaches to transcription should be further studied.

References

[1] Abdallah S., Plumbley M. (2004). Polyphonic transcription by non-negative sparse coding of power spectra. Proceedings of ISMIR'2004, Barcelona, Spain.

[2] Bello, J. P., Daudet, L. & Sandler, M. B. (2002). Time-Domain Polyphonic Transcription using Self-Generating Databases. In Proceedings of the 112th AES Convention. Munich, Germany.

[3] Klapuri A., Virtanen, T., Eronen, A. & Seppänen, J. (2001). Automatic transcription of musical recordings. Proceedings of Consistent & Reliable Acoustic Cues Workshop, CRAC-01, Aalborg, Denmark.

[4] Sterian. A.D. (1999). Model-based Segmentation of Time-Frequency Images for Musical Transcription. Ph.D. Thesis, Univesity of Michigan.

[5] Marchand, S. (2001). An efficient pitch-tracking algorithm using a combination of Fourier transforms, Proceedings of DAFX-01, Limerick, Ireland.

[6] Roads. C. (1996). The Computer Music Tutorial. Cambridge, MA: MIT Press.

[7] Patterson, R. D. & Hodsworth J. (1990). A functional model of neural activity patterns and auditory images. Advances in speech, hearing and auditory images 3, W.A. Ainsworth (ed.), London: JAI Press.

[8] Meddis. R. (1986). Simulations of mechanical to neural transduction in the auditory receptor. J.Acoust.Soc.Amer., vol. 79, no. 3, 702-711.

[9] Large E.W., Kolen, J.F. (1994). Resonance and the perception of musical meter. Connection Science, vol. 2, no. 6, 177-208.

[10] Marolt M. (2004). A Connectionist Approach to Transcription of Polyphonic Piano Music. IEEE Transactions on Multimedia, June 2004, Vol.6, Issue 3.