Networks of Adaptive Oscillators for Partial Tracking and Transcription of Music Recordings

Matija Marolt
Faculty of Computer and Information Science, University of Ljubljana, Slovenia
matija.marolt@fri.uni-lj.si

ABSTRACT

In this paper, we present a technique for tracking 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. The model is further extended to track groups of harmonically related partials by grouping oscillators into networks. We present the integration of the partial tracking model into a system for transcription of polyphonic piano music. The transcription system is based on a connectionist architecture that employs networks of adaptive oscillators for partial tracking and feed forward neural networks for associating partial groups with notes. We provide a short overview of our entire transcription system and present its performance on transcriptions of several synthesized and real piano recordings.

1 INTRODUCTION

Most human listeners would agree that music conveys some sort of meaning. While listening to a piece of music, a series of cues can be perceived, ranging from emotional characterizations (sad, happy…) to more objective elements, such as melody or types of instruments appearing in the piece. Although human listeners find it natural to extract, memorize and recreate the melody of a piece of music, transcription of polyphonic music is not a task inherent to human perception and can only be performed by experienced or trained musicians.

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 very 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.

A tool for transcription of polyphonic music would be useful in a wide range of applications. Even if we disregard the common desire of amateur musicians to have transcriptions of their favorite pieces available for reinterpretation, the compact and standardized parametric representation of music that transcription produces is useful in applications ranging from content-based retrieval of music (i.e. query by example systems) and music analysis systems to accompaniment systems. Transcription aids musicologists in analyzing music that has never been written down, such as improvised or ethnical music, it is useful in the process of making music, as well as in newer coding standards, such as MPEG-4, which may include note-level descriptions of music.

First researches into polyphonic music transcription have been made by Moorer (1977); he was experimenting with transcription of melodic lines of two voices of different timbres and frequency ranges. In recent years, several systems have been developed. Some of them are targeted to transcription of music played on specific instruments (Rossi, 1998; Sterian, 1999; Dixon, 2000; Bello, Daudet & Sandler, 2002; Ortiz-Berenguer & Casajus-Quiros, 2002; Monti & Sandler, 2002), while others are general transcription systems (Kashino *et al*, 1995; Klapuri, 2001). All authors, except for (Bello, *et al.*, 2002), 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 input signal. Various techniques, such as statistical frameworks (Kashino *et al*, 1995; Klapuri, 2001), blackboard architectures (Monti & Sandler, 2002), distance metrics (Rossi, 1998)... are used in the process of grouping the found cues into notes, relying on information such as harmonicity, common onset/offset times... Prior knowledge of tone sources is sometimes taken into account (Kashino, 1995; Rossi,

1998; Sterian, 1999; Ortiz-Berenguer & Casajus-Quiros, 2002), as well as higher-level knowledge of music, such as probabilities of chord transitions (Kashino, 1995). Good recognition rates have been achieved in identifying notes from individual frames of time-frequency representations (ignoring temporal features) (Klapuri, 2001; Ortiz-Berenguer & Casajus-Quiros, 2002); the downside of such approaches is that they do not scale well to transcription of continuous musical signals and may exhibit problems with noisy sounds (drums, percussive onsets) appearing in the signal. To reduce the complexity of the time-frequency 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 input signal (Sterian, 1999; Dixon, 2000).

In this paper, we present a connectionist approach to music transcription. Connectionist methods, such as neural networks, have been successfully applied in many pattern recognition domains and our incentive was to build a transcription system that would be based on connectionist principles. 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 synthesized and real piano recordings.

The organization of this paper is as follows. In Section 2 we propose a model for tracking partials in a polyphonic audio signal, based on adaptive oscillators. Section 3 presents an extension of this model to a model that tracks groups of harmonically-related partials. Section 4 presents a brief overview of our transcription system and in section 5 we present performance statistics of the system on transcriptions of synthesized and real recordings of piano music. Section 6 concludes this paper.

2 ADAPTIVE OSCILLATORS FOR PARTIAL TRACKING IN MUSICAL SIGNALS

A melodic sound can be roughly described as a sum of components with stable frequencies and time-varying amplitudes. These components are also called partials and can be recognized as prominent

horizontal structures in a time-frequency representation of an audio signal. By finding partials, one isolates the stable frequency components most likely belonging to tones, and discards noisy components, thus making the representation clearer and more compact. This is especially desirable in transcription systems, where the goal is to find all the tones (notes) present in the audio signal at any given moment in time. Currently, most partial trackers used in transcription systems are based on a procedure similar to the tracking phase vocoder (Roads, 1996). After the calculation of a time-frequency representation, peaks are computed in each frequency image. Only peaks with amplitude that is larger than a chosen (possibly adaptive) threshold are kept as candidate partials. Detected peaks are then linked over time according to intuitive criteria such as proximity in frequency and amplitude, and partial tracks are formed in the process. 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. Some systems therefore use additional heuristics for merging fragmented partial tracks. Another shortcoming of the "peak picking-peak connecting" approach is detection of frequency modulated partials. Here, the peak connecting algorithm can fail if it is not designed to tolerate frequency modulation. In transcription, where each missed or spurious partial may be important, errors of partial tracking algorithms made on signals that include beating, frequency modulation... may lead to a large number of missed or spurious notes. An innovative approach to partial tracking has been proposed by Sterian (1999), who still uses a peak picking procedure in the first phase of his transcription system, but later uses Kalman filters, trained on examples of instrument tones, to form partials from peaks.

The review of partial tracking methods used in current transcription systems has led us to the development of a different partial tracking model that would not be based on the standard peak-picking/peak connecting paradigm. In this section, we propose a partial tracking model based on connectionist principles. It is composed of two parts: an auditory model, which emulates the functionality of human ear, 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 many auditory models currently available, we decided to use a combination of the Patterson-Holdsworth gammatone filterbank (Patterson & Holdsworth, 1990; Slaney, 1993) and Meddis' hair cell model (Meddis, 1986), as their implementations are efficient and readily available.

The first stage of the auditory model emulates the movement of basilar membrane in the inner ear with a bank of bandpass IIR filters (gammatone filters). We are using a bank of 200 gammatone filters to split the signal into 200 frequency bands with center frequencies logarithmically spaced between 70 and 6000 Hz. Filter parameters are taken from (Moore & Glasberg, 1983).

Subsequently, output of the gammatone filterbank is processed by the Meddis' model of hair cell transduction. The hair cell model converts each gammatone filter 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).

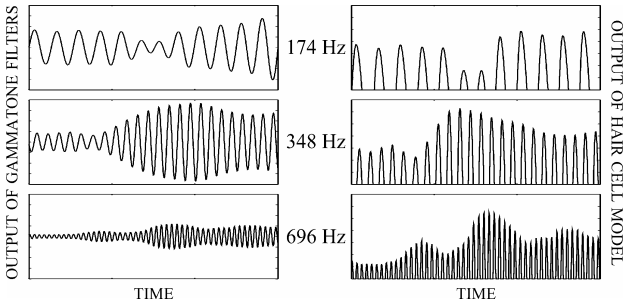

Fig. 1. Analysis of three partials of piano tone F3 with the auditory model.

2.2 Partial Tracking with Adaptive Oscillators

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 (see 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. Meddis and Hewitt (1991) have demonstrated that a summary autocorrelogram (autocorrelogram summed across frequency channels) explains the perception of pitch in a wide variety of stimuli and autocorrelogram is used in many psychoacoustic models that try to explain the mechanisms behind human pitch perception.

We decided to use a different approach to estimating periodicity in output channels of the auditory model. The approach is based on adaptive oscillators and their ability to synchronize with periodic input stimuli.

2.2.1 The Large-Kolen Adaptive Oscillator

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.

Various models of adaptive oscillators have been proposed, some have also found use in computer music researches for modeling rhythm perception (Large & Kolen, 1994; McAuley, 1995) and for simulation of various psychoacoustic phenomena (Wang, 1996). After reviewing several models, we

decided to use a modified version of the Large-Kolen adaptive oscillator (Large & Kolen, 1994) in our partial tracking model.

The Large-Kolen oscillator oscillates in time according to its period (frequency) and phase. Its input consists of a series of discrete impulses, representing events in the input signal. After each oscillation cycle, the oscillator adjusts its phase and period, trying to match the phase and period of events in the input signal. It also updates its output that reflects the level of synchronization achieved. If events occur in regular intervals (are periodic), the final effect of synchronization is alignment of oscillations with input events. Phase and period of the Large-Kolen oscillator are updated according to the modified gradient descent rule, minimizing an error function that describes the difference between input events and beginnings of oscillation cycles:

$$\begin{aligned} \Delta t_x &= \eta_1 s(t) \frac{p}{2\pi} \operatorname{sech}^2 \chi (\cos 2\pi\phi(t) - 1) \sin 2\pi\phi(t) \\ \Delta p &= \eta_2 s(t) \frac{p}{2\pi} \operatorname{sech}^2 \chi (\cos 2\pi\phi(t) - 1) \sin 2\pi\phi(t) \end{aligned} \tag{1}$$

Δt_x and Δp represent the changes in phase and period of the oscillator after each oscillator's cycle. p and ϕ are the phase and period of the oscillator and s the input signal (stimulus). η_1 and η_2 are parameters that control the strength of synchronization to the stimulus, while χ defines the width of the receptive field of the oscillator. An impulse in the stimulus only impacts the oscillator's period and phase if it falls within the width of the receptive field. The receptive field has been introduced to facilitate synchronization to simple rhythmical patterns in a rhythmically complex stimulus and its width is adapted after each cycle. The receptive field width also provides a measure of the level of synchronization achieved.

2.2.2 Modifications of the LK Oscillator

In our partial tracking model, we use Large-Kolen (LK) adaptive oscillators to detect periodicities in output channels of the auditory model. Each output channel of the auditory model is discretized by calculating centroids of individual half-waves and routed to the input of an adaptive oscillator. The initial frequency of the oscillator is set to the center frequency of its input channel. As in our partial

tracking model, oscillators are not processing complex polyrhythmic stimuli (oscillator's stimulus is either (quasi) periodic or not periodic at all), we simplified the LK oscillator model by choosing a constant width of receptive field χ (χ=1). Since the width of the receptive field is closely tied to the measure of level of synchronization, we introduced a new measure of level of synchronization *c*, which is also used as the output value of our oscillator model and is calculated each time an oscillator completes its oscillation cycle:

$$c = c\alpha \exp(-\beta \frac{\sum_{t_x - p < t \le t_x} s(t)\left|\Delta t_x(t)\right|}{\sum_{t_x - p < t \le t_x} s(t)}) \qquad (2)$$

s(t) represents the stimulus, composed of a series of discrete impulses. Oscillator's output *c* is calculated as an exponential function of the centroid of all phase corrections $\Delta t_x(t)$ that occurred during the last oscillator's cycle ($t_x - p < t \le t_x$), weighted with the strength of input impulses *s(t)*. α and β are parameters that control the scaling and the impact of phase corrections on the output function. After some experimenting, we chose to use the values 4.8 and 10 for the two constants. *c* is also averaged over time with a first order filter and limited to values between 0 and 1, higher value meaning a higher level of synchronization to the stimulus. If we look at equation (2), we can see that *c* is mostly dependant on the amount of phase corrections that occur in each oscillator's cycle. Large phase corrections indicate poor synchronization and consequently reduce the value of *c*; on the other hand, small phase corrections indicate good synchronization and thus increase the output function. If an oscillator stays completely out of sync (c<0.03) with the stimulus in over three consecutive cycles, its frequency is reset to an initial value.

2.2.3 Partial Tracking

The rationale behind the use of adaptive oscillators for partial tracking is simple. It is well known that a periodic output channel of an 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 two advantages: because oscillators constantly adapt to their stimuli, partials with slowly changing frequencies (vibrato…) can easily be tracked. Since the auditory model reduces the dynamic range of the input signal and thus boosts partials with low amplitudes, these can be easily tracked as well.

Four examples of partial tracking with the modified Large-Kolen oscillator are illustrated in Fig. 2. Example A presents a simple case of tracking a 440 Hz sinusoid. The oscillator (initial frequency 440 Hz) synchronizes successfully, as can be seen from its output, and after an initial 1 Hz rise, its frequency settles at 440 Hz. Example B shows how two oscillators with initial frequencies set to 440 and 445 Hz synchronize to a sum of 440 and 445 Hz sinusoids (5 Hz beating). Both oscillators synchronize successfully at 442.5 Hz, as can be seen from their outputs and frequencies. The behavior is consistent to human perception of the signal. Example C is shows the tracking of a frequency modulated 440 Hz sinusoid. The oscillator synchronizes successfully; its frequency follows that of the sinusoid. The last example (D) shows how two oscillators track two frequency components that rise/fall from 440 to 880 Hz. Tracking is successful; each oscillator tracks the component closest to its input frequency channel.

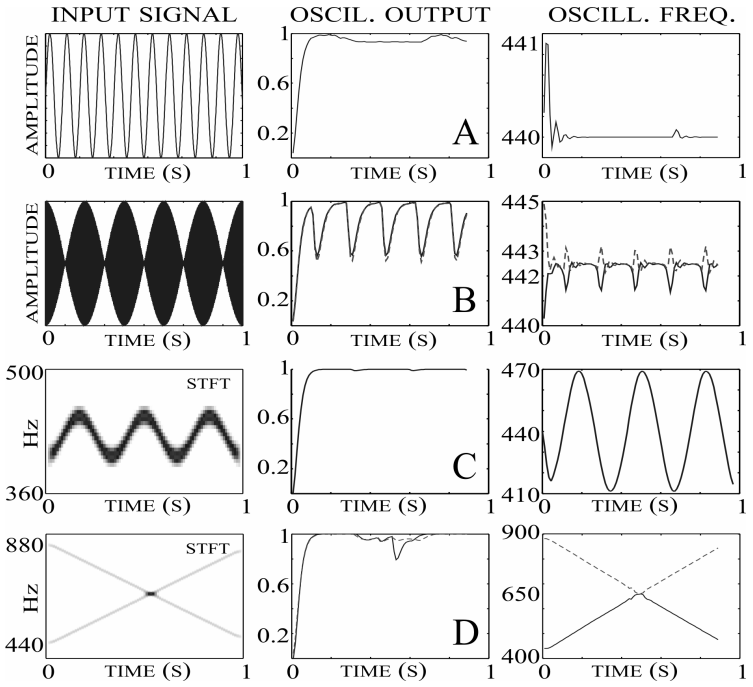

Fig. 2. Partial tracking with adaptive oscillators.

Example in Fig. 3 shows how oscillators track partials of piano tone A4 (pitch 440 Hz). 88 oscillators with initial frequencies logarithmically spaced between 55 and 8372 Hz were used in the example. Fig. 3A displays outputs of all 88 oscillators (these are independent on amplitudes of partials). Fig. 3B shows amplitude envelopes of frequency channels of the auditory model, calculated from outputs of the gammatone filterbank, while Fig. 3C shows the product of oscillator outputs and amplitude envelopes. The three bottom figures show a cross section of the upper time-frequency representations at 200 ms. In Fig. 3A we can see that oscillators successfully track the first seven and the tenth partial of tone A4; the 8^{th} and 9^{th} partial are missed, because their amplitudes were too low. One can also notice some noise in lower frequencies, which is mostly due to the sound of the hammer hitting the strings. Overall, the obtained representation is very compact and gives a clear picture of how partials of tone A4 develop over time.

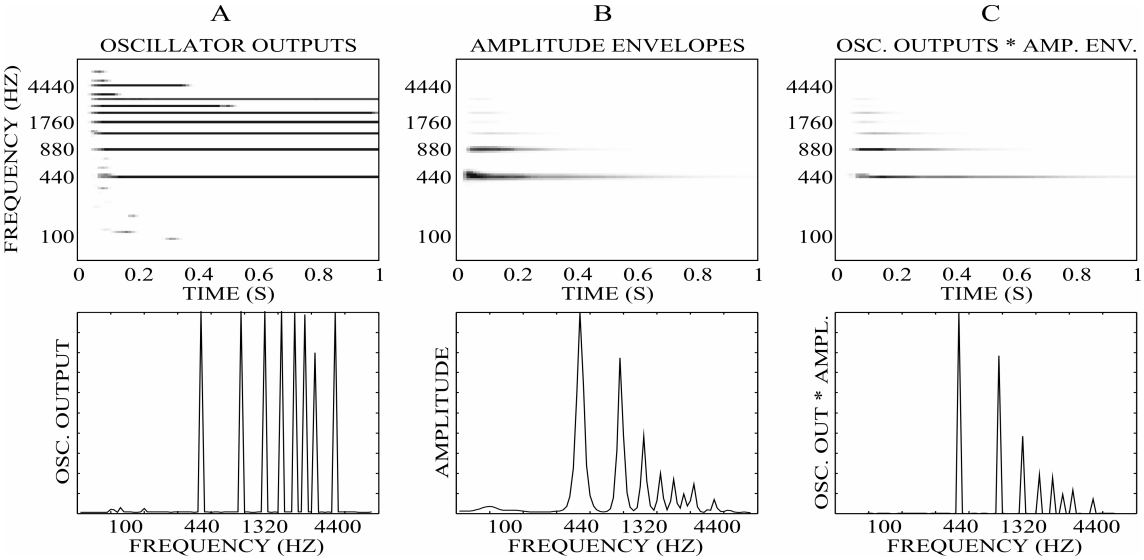

Fig. 3. Tracking of partials of piano tone A4

3 PARTIAL TRACKING WITH ADAPTIVE OSCILLATOR NETWORKS

The partial tracking model presented in the previous section is capable of successfully following partials in musical signals without explicit “peak picking/peak connecting” rules. We showed that the model is quite robust and that it can follow partials in cases of beating or vibrato. Although finding each individual partial in a signal may be useful in some applications, in transcription systems the goal is to obtain a representation that is as compact as possible.

As most tones are harmonic, we extended the presented model of tracking individual partials to a model of tracking groups of harmonically related partials by joining adaptive oscillators into networks. Networks of oscillators are fully connected, initial frequencies of oscillators in a network are set to integer multiples of the frequency of the first oscillator (see Fig. 4). 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.

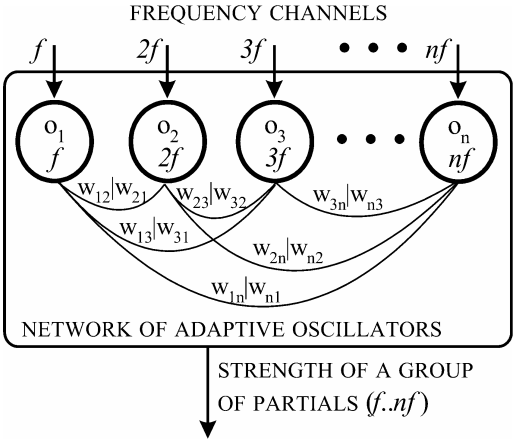

Fig. 4. A network of adaptive oscillators.

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. Weights are related to individual oscillator's number (lower-numbered oscillators have more influence on the overall output) and to individual oscillator's frequency (as the frequency of an oscillator drifts from the average network frequency, the oscillator loses its impact on the network's output). Larger frequency deviations are tolerated for higher-numbered oscillators to account for frequency stretching. Because the output of a network only depends on outputs of oscillators in the network, it is virtually independent of the amplitude of tracked partials.

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. Adjustments are made according to the following rules:

$$r_p = \exp(-1000(\frac{dp_d}{sp_s} - 1)^2) \qquad r_c = \exp(-2.3\frac{c_d^2}{c_s^2})$$
$$p_d = p_d + \frac{sp_s - dp_d}{d} r_p r_c w_{sd} \qquad c_d = c_d + c_d r_p r_c w_{sd} \tag{3}$$

d is the number of the destination (non-synchronized) oscillator in the network (starting from 1), while s represents the number of the source (synchronized) oscillator. The period of the destination

oscillator p_d and its output value c_d change according to two factors: r_p and r_c. These are two gaussians, representing the ratio of periods of the two oscillators (p_d - period of the destination oscillator, p_s - period of the source oscillator) and the ratio of outputs of the two oscillators (c_d - output of the destination oscillator, c_s output of the source oscillator). Factor r_p is a gaussian with maximum value, when periods of both oscillators are in a perfect harmonic relationship ($dp_d/sp_s = 1$). The value falls as periods drift away from this perfect ratio and approaches zero, when the ratio is larger than a semitone. r_c has the largest value, when a synchronized oscillator influences the behavior of a non-synchronized oscillator (c_s is large, c_d is small) and falls as c_d increases. Connection weights w_{sd} are calculated according to the oscillator's number; the first few partials are considered to be more important and consequently the influence of lower-numbered oscillators in the network is stronger than the influence of higher-numbered oscillators ($w_{1n}>w_{n1}$).

Adjustments (3) push the frequency of a non-synchronized oscillator closer to the frequency of the partial it should track and also increase its output value, which results in faster synchronization of all oscillators in the network and consequently leads to faster discovery of a group of partials.

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 (individual partials). Noise usually doesn't appear in the form of harmonically related frequency components, so networks of oscillators are more resistant to noise and provide a clearer time-frequency representation. Within a network, each oscillator is connected to all other oscillators with excitatory connections. 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.

Fig. 5 shows an example of tracking groups of partials of piano tone A4. The figure can be compared to Fig. 3, which shows partial tracking of tone A4 with individual oscillators. The representation was

obtained with 88 oscillator networks, each containing up to 10 oscillators. As in Fig. 3, results are divided into three parts; in A, we show outputs of all 88 oscillator networks, B shows amplitude envelopes of frequency channels of the auditory model, and C the product of network outputs and amplitude envelopes. Bottom figures show a cross section of upper figures at 200 ms. Comparison with Fig. 3 shows that networks give a clearer TF representation than individual oscillators. In A, we can see that three strong partial groups were found, corresponding to the first three partials of tone A4. There is less noise in the representation and when network outputs and amplitude envelopes are combined, only the three partial groups remain in the representation.

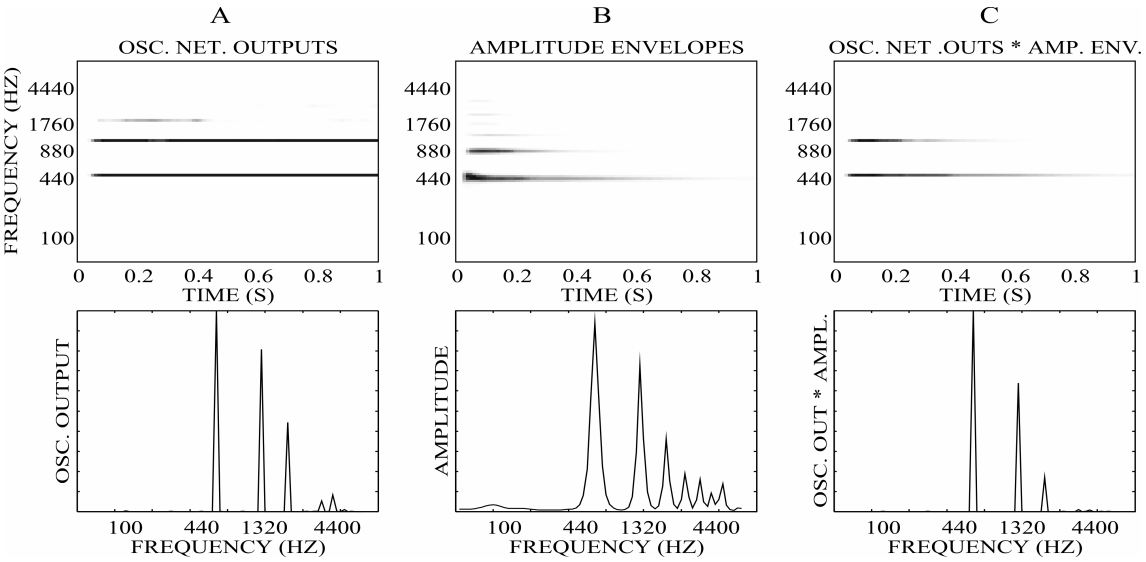

Fig. 5. Tracking groups of partials in piano tone A4

Another example is given in Fig. 6, which displays slices taken from three time-frequency representations of piano chord C3E3B4, calculated 100 ms after the onset: representation with uncoupled oscillators, representation 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. 6A. Fig. 6B shows outputs of 88 oscillator networks, tuned to the same pitches. Product of

networks' outputs and amplitudes of partials is shown in Fig. 6C. Fig. 6D displays the first 440 frequency bins of the Fourier transform calculated with a 100 ms Hamming window.

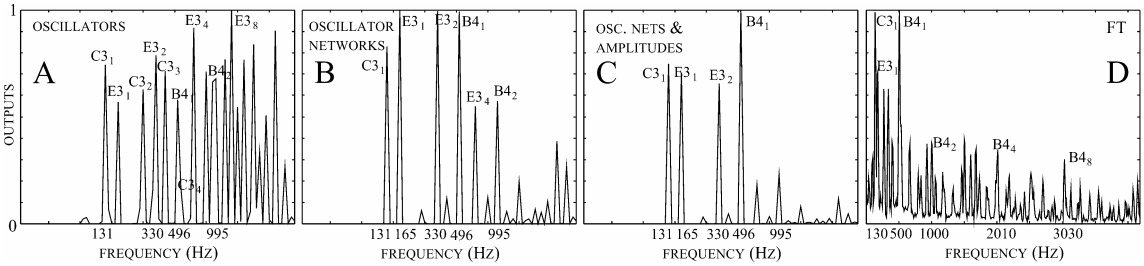

Fig. 6. Representations of piano chord C3E3B4.

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. 6C), 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. 6C with the Fourier transform in 6D, advantages of partial group tracking for transcription are obvious.

Overall, examples show that 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.

4 TRANSCRIPTION OF PIANO MUSIC

The partial tracking model presented in previous sections has been incorporated into our system for transcription of piano music, called SONIC (Marolt, 2001). The overall structure of the system is shown in figure 7. Next to partial tracking, the system also includes a note recognition module (briefly described in the next section), an onset detector (Marolt, 2002a), a module for resolving repeated notes (Marolt, 2002b) and simple algorithms for estimation of tuning, note length and loudness. Music transcription is a difficult task, so we put one major constraint on our transcription system: it only transcribes piano music, so piano should be the only instrument in the analyzed musical signal. We didn't make any other assumptions about the signal, such as maximal polyphony, minimal note length, style of transcribed music or the type of piano used. The system takes an acoustical waveform of a piano recording (44.1 kHz sampling rate, 16 bit resolution) as its input. Stereo recordings are converted to mono. The output of the system is a MIDI file containing the transcription. We present performance statistics of the system on several synthesized and real piano recordings in section 5.

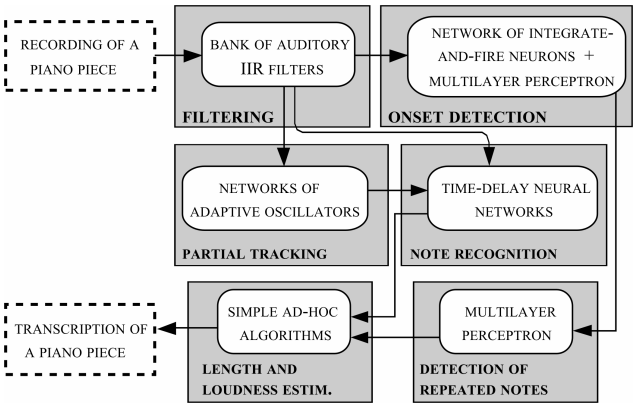

Fig. 7. Structure of SONIC.

4.1.1 Note recognition with neural networks

A note recognition module is the central part of every transcription system. Its input usually 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 neural networks.

We use a set of 76 neural networks to perform note recognition. Inputs of each network are taken from outputs of the partial tracking module presented in the previous sections. They contain one or more time frames (sampled at every 10 ms) of output values of oscillator networks, amplitude envelopes of signals in frequency channels of the auditory model (calculated by half-wave rectification and smoothing) and the combination of amplitude envelopes and oscillator networks' outputs. Each neural 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... Altogether 76 networks are used to recognize notes from A1 to C8. This represents the entire range of piano notes, except for the lowest octave from A0 to Ab1. We decided to ignore the lowest octave, because of poor recognition results. These notes are quite rare in piano pieces, so their exclusion does not have a large impact on the overall performance of the system. Because each neural network only recognizes one note (target note) in its input, it only has one output neuron; a high output value indicates the presence of the target note in the input signal, a low value indicates that the note is not present. After extensive testing of several neural network models, we decided to use time-delay neural networks (TDNNs) (Waibel *et al*, 1989) in our system, as they provided the best performance. Networks were trained on a database of approx. 120 synthesized piano pieces of various styles, combined with randomly generated chords.

4.1.2 Impact of Partial Tracking on Accuracy of Note Recognition with TDNNs

To assess the impact of the proposed partial tracking module on accuracy of note recognition with TDNNs, we compared the performance of TDNNs trained on patterns that consisted of outputs of the partial tracking module (as described previously) to the performance of TDNNs trained on patterns that consisted of outputs of a multiresolution time-frequency transform, similar to the constant-Q transform (Brown, 1992) with window sizes from 90 ms to 5 ms at frequencies from 60 Hz to 9 kHz. We tested the performance of both sets of TDNNs on transcriptions of several synthesized piano pieces. Table 1 lists average performance statistics of both sets of networks on seven synthesized piano pieces of different complexities and styles, containing over 20000 notes. Percentages of

correctly found notes, spurious notes (notes that were found but were not in the original score) and octave errors are given for both sets of networks.

Table 1. Average transcription accuracy of systems with and without partial tracking

	correct	spurious	oct. err.
No PT	92.8	27.9	39.5
With PT	94.4	11.1	77.9

The percentage of correctly found notes is similar in both systems; partial tracking improves accuracy by approximately 1.5%. Partial tracking significantly reduces the number of spurious notes, as it more than halves. Just as important is the change in the structure of errors. Almost 80% of all errors in the system with partial tracking are octave errors that occur when the system misses or finds a spurious note, because of a note an octave, octave and a half or two octaves apart. Octave errors are very hard to eliminate, but because the missed or spurious notes are consonant with other notes in the transcribed piece, they aren't very apparent if we listen to resynthesized transcriptions. Octave errors are therefore not as critical as some other types of errors (i.e. halftone errors), which make listening to resynthesized transcriptions unpleasant. We therefore consider the higher percentage of octave errors in the system with partial tracking to be a significant improvement. Overall, we can conclude that the partial tracking model significantly improves transcription accuracy with TDNNs.

5 PERFORMANCE ANALYSIS

To analyze the performance of our transcription system, we tested it on a number of synthesized recordings and on six real recordings that were transcribed by hand with the help of the original score. Originals and transcriptions of all presented pieces can be heard on *http://lgm.fri.uni-lj.si/SONIC*. Table 2 lists performance statistics of three synthesized and three real piano pieces: percentages of correctly found and spurious notes in transcriptions, as well as the distribution of errors into octave, repeated note and other errors are shown. Separate error distributions are given for missed and

spurious notes. An error can fall into several categories, so the sum of error percentages may be greater than 100. The total number of notes, as well as maximal and average polyphony of each piece are also shown.

Table 2. Performance Statistics of Transcriptions of 3 Synthesized and 3 Real Piano Recordings

	corr. notes	spur. notes	missed notes			spurious notes			num. notes	avg. poly	max. poly
			octave	repeat.	other	octave	repeat.	other			
1	98.1	7	31.4	23.6	56.4	84.4	22.3	7.9	6680	2.7	6
2	92.3	10.6	53.2	39.2	29.4	95.3	29.9	0	1008	4.1	12
3	86	9.5	80.8	25.6	9	96	8.2	5.1	1564	3.4	9
4	88.5	15.5	35.1	18.2	52.2	80.5	17.6	13.9	1351	2.6	6
5	68.3	13.6	30.3	2.1	75.3	79	6.4	20.7	457	4.4	11
6	85.9	15.2	70.3	10.8	27	87.4	7.1	12.3	1564	3.4	9

The transcribed synthesized recordings are: (1) J.S. Bach, Partita no. 4, BWV828, Fazioli piano; (2) A. Dvořak, Humoresque no. 7, op. 101, Steinway D piano; (3) S. Joplin, The Entertainer, Bösendorfer piano. Real recordings are: (4) J.S. Bach, English suite no. 5, BWV810, 1st movement, performer Murray Perahia, Sony Classical SK 60277; (5) F. Chopin, Nocturne no. 2, Op. 9/2, performer Artur Rubinstein, RCA 60822; (6) S. Joplin, The Entertainer, performer unknown, MCA 11836.

The average number of correctly found notes in synthesized recordings is around 90%. The average number of spurious notes is 9%. Most of the missed notes are either octave errors or misjudged repeated notes. Notes are also missed in very fast passages, such as arpeggios or thrills (most missed notes in Partita), when they are masked by louder notes (many notes in Humoresque) or due to other factors such as missed onsets and high polyphony. A majority of spurious notes are octave errors, often combined with misjudged repeated notes. These are especially common in pedaled music (Humoresque) or in loud chords (The Entertainer). Other reasons for spurious notes include missed and spurious onsets and errors due to high polyphony.

Some common errors can be seen in a transcription example taken from Humoresque and shown in Fig. 8A (table 2/2). Missed notes are marked with a - sign, spurious notes are marked with a + sign.

All three spurious notes are octave errors. Out of the two missed notes, A5 was missed, because it is masked by the louder E3C4 chord, while note E3 is a missed repeated note.

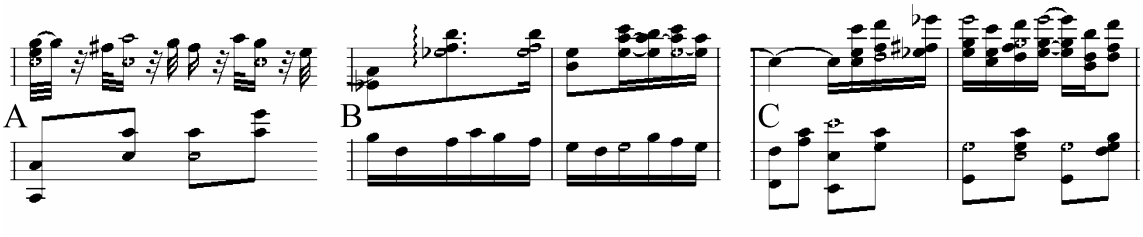

Fig. 8. Transcription examples: A: Humoresque (2/2), B: BWV810 (2/4), C: The Entertainer (2/6).

Results on real recordings are not as good as those on synthesized recordings. Poorer transcription accuracy is a consequence of several factors. Recordings contain reverberation and more noise, while the sound of real pianos includes beating and sympathetic resonance. Furthermore, performances of piano pieces are much more expressive, they contain increased dynamics, more arpeggios and pedaling. All of these factors make transcription more difficult.

The analysis of SONIC's performance on the real recording of Bach's English Suite (table 2/4, Fig. 8B) showed that besides octave and repeated note errors, most of the missed notes are either quiet low pitched notes (E3 in measure 2, Fig. 8B) or notes in arpeggios and thrills. Chopin's Nocturne (table 2/5) proved to be the greatest challenge for our system. The recording is a good example of very expressive playing, where a distinctive melody is accompanied by quiet, sometimes barely audible left hand chords. The system 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). We compared transcriptions of the real and synthesized version of The Entertainer (table 2/3 and 2/6, Fig. 8C) and 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 types of errors (see Fig. 8C).

When we set out to make a comparison of our results to results of other systems, we found that the task is not an easy one. The lack of a standard set of test examples makes comparison of different transcription systems a difficult task, at best. It is further complicated by the fact that systems put very different constraints on the type or style of music they transcribe. Some authors have published performance statistics of their systems on piano pieces, but as we were unable to obtain the same recordings, we chose to avoid direct comparison, as our belief is that results would not be very relevant.

6 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 partials. By using a connectionist approach, we avoided some of the pitfalls of classical partial tracking approaches and showed that our model successfully tracks partials in cases of beating and frequency modulation. An additional advantage of the presented partial tracking model is that it can be easily extended to a model for tracking groups of harmonically related partials by joining oscillators into networks. Oscillator networks provide a clearer time-frequency representation of a signal that is especially suitable for transcription purposes and we showed that partial tracking with networks of adaptive oscillators significantly improves the accuracy of transcription with time-delay neural networks. We presented an overview of our transcription system called SONIC and presented performance statistics on transcriptions of several synthesized and real piano recordings. Overall, results are very promising and we believe that connectionist approaches to transcription should be further studied.

We are currently extending our partial tracking model to include an algorithm for dynamic self-organizing grouping of oscillators into networks (as opposed to the current static structure of networks), based on competition between various oscillator groups. This will hopefully reduce the number of octave-related partial groups and lead to a smaller number of errors in the transcription

process. Since results of the system on transcriptions of non-piano music (some examples are given on *http://lgm.fri.uni-lj.si/SONIC*) are also quite promising, even though the system was not intended to be used on non-piano music, we plan to extend the system to partial transcription of any kind of melodic music, where our goal is to extract only the melody and the basic harmonic structure of a piece.

7 REFERENCES

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

Brown. J.C. (1992). Calculation of a constant Q spectral transform. *Journal of Acoustical Society of America*, *vol. 89*, no. 1, 425-434.

Dixon. S. (2000). On the computer recognition of solo piano music. *Proceedings of Australasian Computer Music Conference*, Brisbane, Australia.

Kashino, K., Nakadai, K., Kinoshita, T. & Tanaka, H. (1995). Application of Bayesian probability network to music scene analysis. *Proceedings of International Joint Conference on AI, Workshop on Computational Auditory Scene Analysis*, Montreal, Canada.

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.

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

Marolt, M. & Divjak, S. (2002). On detecting repeated notes in piano music. In Fingerhut, M. (ed.). *ISMIR 2002: conference proceedings*. Paris: IRCAM - Centre Pompidou, 273-274.

Marolt. M. (2001). SONIC : transcription of polyphonic piano music with neural networks. *Workshop on Current Research Directions in Computer Music*, Barcelona, Spain, 217-224.

Marolt, M., Kavcic, A., Privosnik, M., Divjak, S. (2002). On detecting note onsets in piano music. *Proceedings of MELECON 2002*, Cairo, Egypt, 385-389.

McAuley. J.D. (1995). Perception of time as phase: toward an adaptive-oscillator model of rhythmic pattern processing. Ph.D. Thesis, Indiana University.

Meddis R. & Hewitt M.J. (1991). Virtual pitch and phase sensitivity of a computer model of the auditory periphery I: pitch identification. *Journal of Acoustical Society of America*, *vol. 89*, no. 6, 2866-2882.

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

Monti, G. & Sandler, M. B. (2002). Automatic Polyphonic Piano Note Extraction using Fuzzy Logic in a Blackboard System. *Proceedings of the 5th Conference on Digital Audio Effects*, Hamburg, Germany.

Moore, B.C.J. & Glasberg, B.R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America, vol. 74*, no. 3, 750-753.

Moorer. J.A. (1977). On transcription of musical sound by computer. *Computer Music Journal vol. 1,* no. 4, 32-38.

Ortiz-Berenguer, L.I. & Casajus-Quiros, F.J. (2002). Polyphonic Transcription Using Piano Modeling for Spectral Pattern Recognition. *Proceedings of the 5th Conference on Digital Audio Effects*, Hamburg, Germany.

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.
Roads. C. (1996). The Computer Music Tutorial. Cambridge, MA: MIT Press.
Rossi. L. (1998). Identification de Sons Polyphoniques de Piano. Ph.D. Thesis, L'Universite de Corse, France.
Slaney. M. (1993). An efficient implementation of the Patterson-Holdsworth auditory filterbank. *Apple Computer Technical Report* #35.
Sterian. A.D. (1999). Model-based Segmentation of Time-Frequency Images for Musical Transcription. Ph.D. Thesis, Univesity of Michigan.
Waibel, A.T., Hanazawa, T., Hinton, G., Shikano, K. & Lang, K.J. (1989). Phoneme recognition using time-delay neural networks. *IEEE Trans. on Acoustics, Speech and Signal Processing vol. 37*, no. 3, 328-339.
Wang. D. (1996). Primitive Auditory Segregation Based on Oscillatory Correlation. *Cognitive Science*, *no. 20*, 409-456.