AUTOMATIC TRANSCRIPTION OF MUSIC

Anssi Klapuri[1], *Antti Eronen*[1], *Jarno Seppänen*[2], *Tuomas Virtanen*[1]

[1]Tampere University of Technology, P.O.Box 553, FIN-33101 Tampere, Finland
[2]Nokia Research Center, P.O.Box 100, FIN-33721 Tampere, Finland
{klap,eronen,tuomasv}@cs.tut.fi, jarno.seppanen@nokia.com

ABSTRACT

A system for the automatic transcription of music is described. Signal processing methods are introduced that solve different facets of the overall problem. Main emphasis is laid on finding the multiple pitches of concurrent musical sounds. Sound onset detection and musical meter estimation are described to some extent. Other topics discussed are noise robustness, estimation of the number of concurrent voices, sound separation, and musical instrument recognition. The presented system is evaluated using a database of musical sounds, synthesized MIDI-songs, and CD-recordings. Also, the performance of the system is compared to that of human listeners.

1. INTRODUCTION

Transcription of music is defined to be the act of listening to a piece of music and of writing down the musical notation for the sounds that constitute the piece. In other terms, this means transforming an acoustic signal into a symbolic representation, which comprises musical events and their parameters. The scope of this paper is the automatic transcription of the harmonic and melodic parts on real-world musical recordings.

A person without a musical education is usually not able to transcribe polyphonic music. The richer is the complexity of a musical composition, the more experience is needed in musical ear training, instruments involved, and in music theory. However, skilled musicians are able to resolve even rich polyphonies with such a flexibility and accuracy that computational transcription systems fall clearly behind humans in performance.

Attempts toward polyphonic transcription date back to 1970s. However, the earliest systems were severely limited in regard to the number of simultaneous sounds, pitch range, or variety of sound sources involved [1,2,3]. Relaxation of these constraints was first tried by limiting to a one carefully modeled instrument [4,5], or by allowing some more errors to occur in the output [6].

More recently, Kashino et al. applied psychoacoustic processing principles in the framework of a Bayesian probability network, where bottom-up signal analysis could be integrated with temporal and musical predictions [7]. Martin proposed a system that was able to utilize musical rules in transcribing four-voice piano compositions [8]. Brown and Cooke addressed the auditory grouping and streaming of musical sounds according to common acoustic properties [9]. Godsmark and Brown proposed a blackboard architecture to integrate evidence from different auditory organization principles and demonstrated that the model could segregate melodic lines from polyphonic music [10]. Goto introduced the first pitch analysis method that works quite reliably for real-world complex musical signals, finding the melody and bass lines from complex audio signals [11].

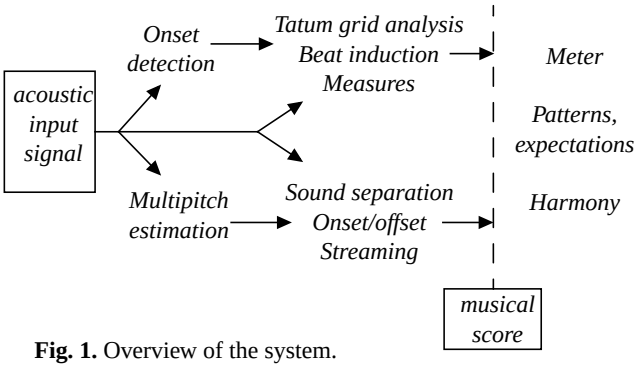


Fig. 1. Overview of the system.

2. SYSTEM OVERVIEW

Figure 1 shows the overview of the system to be presented in this paper. Transformation of an acoustic signal into a musical score goes through data representations of increasing level of abstraction from left to right. Processing takes place in two parallel lines, one for the rhythmic issues (top) and another for the harmony and melody (bottom). The partial results of these two lines are combined to yield the results of the bottom-up transcription: a raw musical score.

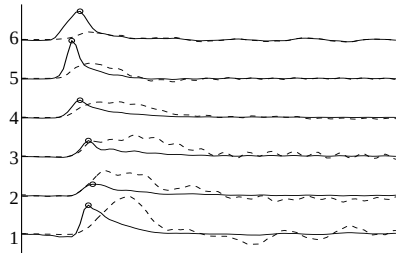
The implementation to be described here addresses only the issues of bottom-up signal analysis, up to the point of musical score. Beyond that line of abstraction, there are still higher-level musical constructs which represent the "internal state" of a musical performance. By implementing musicological and statistical models governing the progression of these musical parameters, predictions of the internal model can be used as a source of information along with the acoustic input signal. Analogous to automatic speech recognition, a "language model" for music seems to be indispensable to achieve reliable transcription. However, these issues are above the scope of this paper.

The different parts of the system are now described in more detail.

3. SOUND ONSET DETECTION

The term *onset detection* refers to the detection of the beginnings of discrete events in acoustic signals. A percept of an onset is caused by a noticeable change in the intensity, pitch or timbre of a sound [12]. A fundamental problem in the design of an onset detection system is distinguishing genuine onsets from gradual changes and modulations that take place during the ringing of a sound. Robust one-by-one detection of onsets has proved to be hard to attain without significantly limiting the target signals.

Only few published systems have set out to solve the problem of one-by-one onset detection [12,13,14]. Instead, most systems



Dashed line: $\frac{d}{dt}(A(t))$
Solid line: $\frac{d}{dt}(\log(A(t)))$
where $A(t)$ denotes the amplitude envelope function.

Fig. 2. Onset of a piano sound. First order *absolute* (dashed) and *relative* (solid) difference functions of the amplitude envelopes at six different frequency channels.

aim directly at higher-level information, such as the perceived beat of a musical signal [15,16,17,18], in which case the longer-term regularities of music can be used to remove single errors and to tune the sensitivity of the low-level detection process.

For us, it was appropriate to detect onsets one-by-one and to use that information along with the acoustic signal to find the musical meter later on. This approach has the advantage of providing information whether events occur at the different positions of the meter or not, and of indicating events which are ornamental or otherwise deviate from the regular temporal framework.

The onset detector has been originally proposed in [19]. The algorithm employs bandwise processing, building upon the idea that incoming energy at some frequency band indicates the beginning of a physical event that is producing the energy. The problem of distinguishing genuine onsets from modulations during the ringing of a sound was solved by processing the relative difference functions of the amplitude envelopes at each frequency band, i.e., by differentiating the logarithm of the amplitude envelopes at each band instead of the amplitude envelope as such. In this case, oscillations in the amplitude envelope do not matter too much after the sound has set on, as illustrated in Fig. 2. The method is able to perform reliable onset detection in musical signals without assuming regularity in the onset positions.

Up-to-date version of the algorithm is merely a simplified version of that described in [19]. The basic structure is important: detecting onsets in the logarithmic amplitude envelopes at distinct frequency bands, and then combining the results across channels. On the contrary, most psychoacoustic details in the original article turned out to be less critical. The current version uses only eight octave-wide frequency channels between 45 Hz and 19 kHz and replaces perceptual models of loudness with an ERB-frequency scale integral over log-magnitude spectrum.

4. MUSICAL METER ESTIMATION

Musical signals usually exhibit temporal regularity, a *meter*. Perceiving musical meter is characterized as the process of detecting and filtering musical events so as to discover underlying periodicities [20]. Musical meter is a hierarchical structure which consists of pulse sensations at different levels (different rates). Moments of musical stress serve as cues from which the listener attempts to extrapolate a regular pattern. Perceiving meter is an essential part of making sense of music.

The most prominent pulse sensation in the metrical hierarchy is *tactus*, often called beat, or the “foot tapping rate”. The beat pulse tends to remain the most regular and aurally prominent through music [20]. *Tatum*, the temporally shortest and perceptually lowest level pulse, is another well-defined and important metrical level for computational processing of music. It acts as a time quantum, integer multiples of which appear as pulse intervals on the other metrical levels. A third important metrical level in Western music is the *measure* rate, which can often be deduced from the rate of harmonic changes and the length of a rhythmic pattern.

The mentioned three relatively well-defined metrical levels together span the rest of the hierarchy. Thus we have taken a four-stage approach to estimating musical meter. Onsets are first detected one-by-one, and that information is then used along with the acoustic signal to estimate the tatum, tactus, and measures. These span the musical meter.

4.1 Tatum grid analysis

A system for finding the tatum grid from acoustic musical signals has been proposed in [21]. The tatum period turned out to be best determined from *event timing* information only. The problem can be seen as finding the greatest common divisor for inter-onset time intervals.

The analysis consists of the following steps. First, inter-onset intervals (IOI) are causally accumulated into a histogram. Robustness against tempo changes is built into the system by letting the histogram values have an exponential decay through time. IOI’s are computed between all onset pairs, not adjacent only. In the second step, we try to find the tatum period, i.e., the largest interval which divides all IOI’s exactly. A problem in doing this is that IOI’s are contaminated with noise. This is solved by evaluating a remainder error function, resulting in an “approximate greatest common divisor” [21].

4.2 Beat induction

Beat induction is the most studied subproblem of musical meter estimation [15,16,17,18]. Beat (tactus) induction embodies the part of rhythm which is most useful for musical interaction. Due to its close connection to movement along with music, absolute speed is one criterion in choosing the tactus. It varies between 40 and 160 beats per minute, and is often close to 70 beats per minute.

The concept of *accent* is central in defining tactus: accentuated events, i.e., moments of musical stress, should preferably coincide with the beat. Beat induction is the result of perceiving accentuated musical events and discovering underlying periodicities in them. Contrary to tatum, timing information is not sufficient for determining the tactus. Thus a model for an accent is needed: what does a musical accent or stress mean in terms of acoustic features?

We took a statistical approach to find a model for a musical accent. A number of 395 musical pieces from different musical genres were collected, and beat was tapped manually for one-minute sections of each piece, recording the beat positions. The acoustic features of the events at the beat positions were then analyzed, with the assumption that they are likely to be more accentuated than other events. The statistical distribution of different features: cepstrum coefficients, delta cepstrum coefficients, spectral centroid, bass level, attack time, etc. were stored. The distributions of those features were then compared for event that

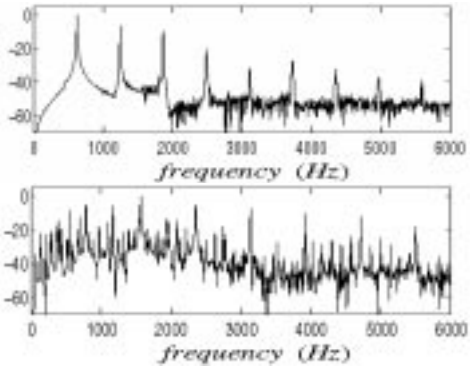


Fig. 3. The magnitude spectrum of a single harmonic sound (top) and that of four sounds (bottom).

coincide with the beat, and for those that do not. This resulted in a model which is able to discriminate events that should be aligned with the tactus grid.

4.3 Measure level

At the time of writing this paper, metrical analysis at the measure level was not yet implemented. Multipitch analysis provides material from which harmonic changes could be detected. The rate and positions of these changes, in turn, provide a cue for finding the measures. Another potential source of information is the length of rhythmic patterns in music.

5. MULTIPITCH ESTIMATION

Pitch perception plays an essential part in experiencing and understanding music. As human listeners, we are able to perceive the pitches of several simultaneous sounds and make efficient use of the pitch to "hear out" a sound in a mixture [22]. Computational modeling of this function, multipitch estimation, has been relatively little explored in comparison to the availability of algorithms for single pitch estimation in monophonic speech signals [23]. It is generally admitted that single pitch estimation methods are not appropriate as such for multipitch estimation. The complexity difference between the spectrum of a single harmonic sound and that of four sounds is illustrated in Fig. 3.

Different parts of the multipitch estimation system have been originally proposed by us in [24,25,26]. The method finds the pitches and separates the spectra of concurrent musical sounds at the level of a single time frame, without temporal features available. The method operates at a wide pitch range and does not require *a priori* knowledge of the sound sources involved.

Overview of the multipitch estimation system is shown in Figure 4. The algorithm consists of two main parts that are applied in an iterative succession, as illustrated in Fig. 1. The first part, predominant pitch estimation, refers to the crucial stage where the pitch of the most prominent sound is estimated in the interference of other harmonic and noisy sounds. This is achieved by utilizing the harmonic concordance of simultaneous spectral components [24]. In the second part, the spectrum of the detected sound is estimated and linearly subtracted from the mixture. This stage utilizes the fact that the spectral envelopes of real sound sources tend to be continuous [25]. The estimation and subtraction steps are then repeated for the residual signal.

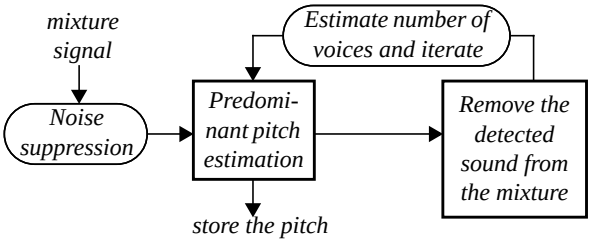


Fig. 4. Parts of the multipitch estimation method.

To control the stopping of the iterative multipitch detection system, the number of concurrent voices must be estimated together with the extracted pitch values. Somewhat surprisingly, the difficulty of estimating the number of voices is comparable to that of finding the pitch values themselves. Huron has studied musician's ability to identify the number of concurrently sounding voices in polyphonic textures [27]. According to his report, the accuracy in performing the task drops markedly already in four-voice polyphony, where the test subjects underestimated the number of voices present in more than half of the cases. Musical mixtures often blend well enough to virtually bury one or two sounds under the others.

Again, we took a statistical approach to solve the problem. We ran the iterative multipitch estimation system for generated mixtures of known polyphonies and measured different characteristics of the signal in the course of the iteration – in search for a feature which would indicate the stopping of the iteration after all sounds have been extracted. A number of features was measured, reflecting the level of the extracted sound, residual spectrum, estimated signal-to-noise ratio etc. The found model for polyphony estimation is described in detail in [26].

Noise suppression is a final necessary part which is needed to apply the multipitch estimation system to the analysis of real musical recordings. Here, "noise" refers to all signal components that do not belong to the harmonic and melodic parts. This definition differs considerably from that in speech processing. Modern musical recordings practically never have continuous noise that could be estimated over a longer period of time. Instead, non-harmonic parts are due to drums and percussive instruments which are transient-like in nature and short in duration.

Due to the non-stationary nature of the noise, we employed an algorithm which estimates and removes noise independently in each analysis frame. Successful noise suppression was achieved by removing both additive and convolutive noise simultaneously, following the lines of RASTA spectral processing [28]. Details of the process are given in [26].

6. SOUND SEPARATION AND STREAMING

Multipitch estimation and sound separation are intimately linked. If the pitch of a sound can be determined without getting confused by other co-occurring sounds, the pitch information can be used to organize spectral components to their sources of production. Or, vice versa, if the spectral components of a source can be separated from the mixture, multipitch estimation reduces to single pitch estimation.

In this section, we consider the separation of the time domain waveform of an individual musical event from the mixture. This is useful for determining the onset and offset times of that particular event, and for streaming events according to their recognized sources of production.

6.1 Sound separation

In [29], we have presented a method for the separation of concurrent harmonic sounds. The method is based on a two-stage approach, where the described multipitch estimator is applied to find initial sound parameters, and in a second stage, more accurate and time-varying sinusoidal parameters are estimated.

For real musical signals, sound separation is significantly more difficult than for artificial mixtures of clean harmonic sounds. However, provided that the correct sounds are detected by the multipitch estimator and that drums do not dominate a musical signal too badly, separation works rather well.

Determination of the onset and offset of each individual sound is based on sound separation. The parameters are estimated in two "sweeps". The first sweep proceeds forwards and tracks the sound until its amplitude envelope indicates the offset. The second sweep goes backwards in time and estimates the onset in a similar manner.

6.2 Streaming

The term *streaming* is here used to refer to the process of classifying separated sounds into distinct streams according to their common sources of production. A preliminary attempt towards stream formation from the separated notes was performed by utilizing acoustic features used in musical instrument recognition research [30]. Mel-frequency cepstral coefficients, the fundamental frequency, the spectral centroid, and features describing the modulation properties of notes were used to form 17-dimensional feature vectors, which were then k-means clustered.

Based on the observations in simulation experiments, stream formation according to sources is possible provided that the timbres of concurrently active sound sources are different enough, and that the distinctive characteristics do not get lost in the separation process. A successful separation and streaming process enables musically meaningful manipulation and remixing of polyphonic and multitimbral music.

7. SIMULATION RESULTS

7.1 Multipitch estimation

The first simulation experiment is analogous to the task usually given to a freshman music student: the system is presented with isolated musical chords to be transcribed. Acoustic material consisted of a database of sung vowels plus 26 musical instruments comprising plucked and bowed string instruments, flutes, and brass and reed instruments. These introduce several different sound production mechanisms, and a variety of spectra. Semirandom sound mixtures were generated by first allotting an instrument, and then a random note from its whole playing range, however, restricting the pitch over five octaves between 65 Hz and 2100 Hz. A desired number of simultaneous sounds was allotted, and then mixed with equal mean square levels. Acoustic input was fed to the multipitch algorithm that estimated the pitches in a single time frame.

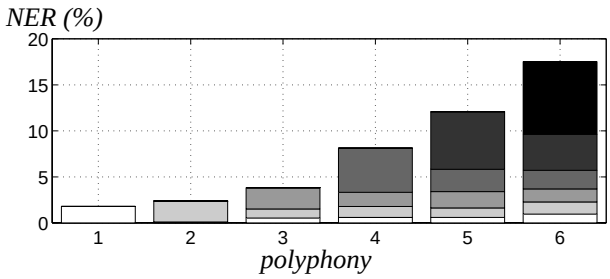


Fig. 5. Note error rates for multipitch estimation in different polyphonies. Bars represent the overall NERs, and the different shades of gray the error cumulation in iteration.

Table 1: Note error rates (%) in the presence of drum sounds.

Analysis frame size	Polyphony					
	1	2	3	4	5	6
190 ms	**6.9**	**11**	**14**	**20**	**29**	**39**
93 ms	**14**	**20**	**29**	**41**	**51**	**61**

Note error rate (NER) metric was taken into use to measure the pitch estimation accuracy. A correct pitch is defined to deviate less than half a semitone (±3 %) from the correct value, making it "round" to a correct note in a western musical scale. NER is defined as the sum of the pitches in error divided by the number of pitches in the reference transcription.

Results for multipitch estimation in different polyphonies are shown in Fig. 5. Random mixtures of one to six sounds were generated, five hundred instances of each. The estimator was then requested to find N pitches in a single 190 ms time frame 100 ms after the onset of the sounds. Here the number of sounds to extract, i.e., the number of iterations to run, was given along with the acoustic mixture signal. In Figure 5, the bars represent the overall NERs as a function of the polyphony, where e.g. the NER for random four-voice polyphonies is 8.1 % on average. The different shades of grey in each bar indicate the error cumulation in the iteration, errors occurred in the first iteration at the bottom, and errors of the last iteration at the top.

As a general impression, the system works quite reliably and exhibits graceful degradation in increasing polyphony, with no abrupt breakdown in any point. Analysis of the error cumulation reveals that the errors occurred in the last iteration account for approximately half of the errors in all polyphonies, and the probability of error increases rapidly in the course of iteration. Besides indicating that the iterative analysis does not work perfectly, conducted listening tests suggest that this is a feature of the problem itself, rather than only a symptom of the algorithms used. In most mixtures, there is a sound or two that are very difficult to hear out because their spectrum is virtually buried under the other sounds.

Table 1 shows the statistical error rate of the overall multipitch estimation system after the noise suppression and polyphony estimation parts were integrated to it. The results have been averaged over three different signal-to-noise ratios: 23 dB, 13 dB, and 3 dB. The test cases were randomly generated as described above, but now interfering drum sounds were randomized and mixed

from Roland R-8 mk II drum machine. Also, the number of concurrent sounds was not told to the system.

The error rates in Table 1 have been calculated by summing together inserted, deleted (missing), or erroneously transcribed notes, and dividing the sum by the number of notes in reference. Among the errors, about two thirds were deletions, which is the least disturbing error type. The amount of inserted notes stays around 1 % in all cases. The rest are erroneous notes. The bias towards underestimating the number of concurrent voices was deliberately implemented, since insertion errors are far more disturbing than deletion errors. Noise suppression enabled reliable pitch estimation still in 3 dB SNRs.

7.2 Comparison to human performance

Listening tests were conducted to measure the human pitch identification ability, particularly the ability of trained musicians to transcribe polyphonic sound mixtures.

Test stimuli consisted of computer generated mixtures of simultaneously onsetting sounds that were reproduced using sampled Steinway grand piano sounds from McGill University Master Samples collection. The number of co-occurring sounds varied from two to five. The gap between the highest and the lowest pitch in each individual mixture was never wider than 16 semitones in order to make the task feasible for those subjects that did not have absolute pitch, i.e., the rare ability to name the pitch of a sound without a reference tone. Mixtures were generated from six partly overlapping pitch ranges. Here results are reported for three different ranges. The low register extended from 33 Hz to 130 Hz, the middle register from 130 Hz to 520 Hz, and the high register from 520 Hz to 2100 Hz. In total, the test comprised 200 stimuli from 20 different categories.

The task was to write down the musical intervals, i.e., pitch relations, of the presented sound mixtures. Absolute pitch values were not asked, and the number of sounds in each mixture was told in beforehand. Thus the test resembles the musical interval and chord identification tests that are part of the basic musical training in western countries.

A total of ten subjects participated the test. All of them were trained musicians in the sense of having taken several years of musical ear training. Seven subjects were students of musicology at a university level. Two were more advanced musicians, possessing absolute pitch and distinguished pitch identification abilities. One subject was an amateur musician of similar musical ability as the seven students.

Figure 6 shows the results of the listening test. Chord error rates (CER) are plotted for different stimulus categories. CER is the percentage of sound mixtures where one or more pitch identification error occurred. The labels of the categories consist of a number which signifies the polyphony, and of a letter which tells the pitch register used. Letter “m” refers to the middle, “h” to the high, and “l” for the low register. Performance curves are averaged over three different groups. The lowest curve represents the two most skilled subjects, the middle curve the average of all subjects, and the highest curve two clearly weakest subjects.

For the sake of comparison, the stimuli and performance criteria used in the listening test were used to re-evaluate the proposed computational model. Five hundred instances were generated from each category included in Fig. 6, using exactly the same software code that produced samples to the listening test. These were fed to the described multipitch system without tailoring its code or parameters. The CER metric was used as a performance measure.

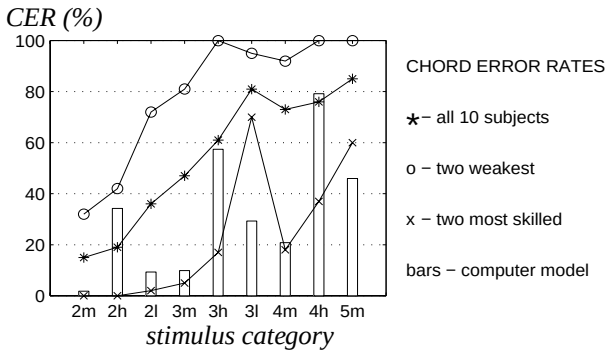


Fig. 6. Chord error rates of the human listeners and of the computational model for different stimulus categories.

The results are illustrated with bars in Fig. 6. As a general impression, only the two most skilled subjects perform better than the computational model. However, performance differences in high and low registers are quite revealing. The devised algorithm is able to resolve combinations of low sounds that are beyond chance for human listeners. This seems to be due to the good frequency resolution applied. On the other hand, human listeners perform relatively well in the high register. This is likely to be due to an efficient use of the temporal features, onset asynchrony and different decay rates, of high piano tones. These were not available in the single time frame given to the multipitch estimator.

7.3 Continuous musical signals

Combining the multipitch estimation part with the temporal processing modules, the system is applicable to the transcription of continuous musical recordings. Since exact musical scores were not available for musical recordings, no statistics on the performance are provided. Instead, excerpts from the original signals and synthesized transcriptions for them are available for listening at http://www.cs.tut.fi/~klap/.

Accurate and realistic evaluation of a transcription system is best achieved by transcribing synthesized MIDI-songs. These have the advantage that the exact reference score is available in the MIDI-data. Also, high-quality MIDI-songs are available that are complex enough to simulated real performances. A simulation environment was created which allows reading MIDI-files into the Matlab environment and synchronizing them with an acoustic signal synthesized from the MIDI.

Figure 7 illustrates the results for two MIDI-songs, a jazz piano performance, and Mozart’s clarinet quintet. These pieces do not have drums and thus represent relatively easy transcription tasks. In these examples, the results are presented in the form of a “fundamental-frequency gram” (F0-gram, lending the term from spectrogram). Original score is plotted with circles and the multipitch estimation results in successive time frames with black dots. Ideally, the train of dots should cover the time span of each note. However, this is not the case even in these relatively easy cases: some notes remain undetected, most of them are detected only part of their lifetime, and some extraneous notes appear.

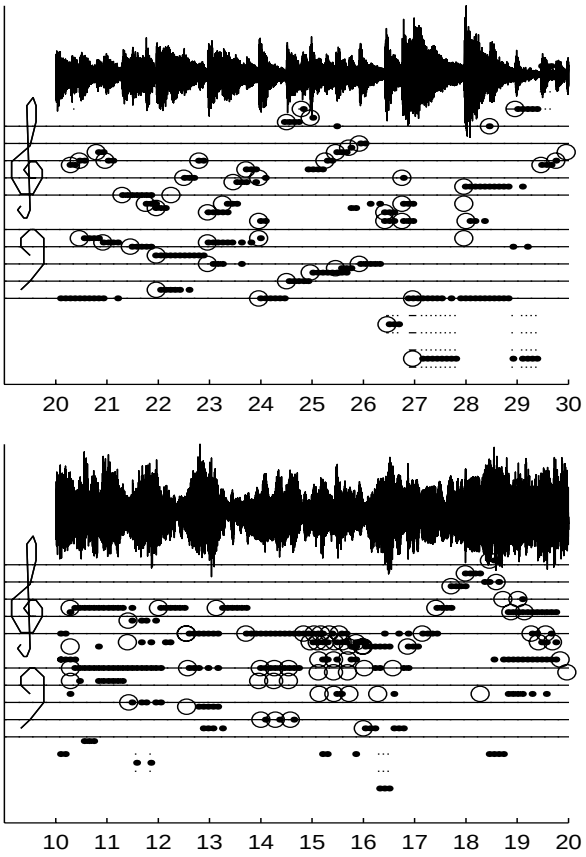


Fig. 7. F0-grams calculated for synthesized MIDI-songs: a jazz piano performance (top) and Mozart's clarinet quintet (bottom). Circles denote the original score and black dots the transcription. Time axes are in seconds. See text for details.

Figure 8 shows the transcription result for a MIDI-version of Abba's song "Waterloo". Here temporal processing has been included by applying multpitch estimation only at the beginnings of each detected event. As a result, the transcription is temporally discrete, although duration values have not been included. A large portion of notes remain undetected, a phenomenon typical to musical pieces with percussive instruments together with other, relatively soft instruments. In many pieces, even half of the notes remain undetected. Another typical defect is that onset/offset determination fails, resulting in multiple detections of a single long-duration note at successive metrical points. Also, there were several pieces which the system hardly made sense at all. However, when listening to pieces that were synthesized from the transcribed versions, the harmonic progression and musical key often remained comprehensible even after such messy transcriptions.

8. CONCLUSIONS

The performance of the presented system was shown to be comparable to that of trained musicians in chord identification tasks. As a striking contrast to that, the system performs essentially worse in the case of real-world musical recordings. This leads to the most obvious conclusion of this paper: the system has an applicable musical ear, but it does not understand anything about

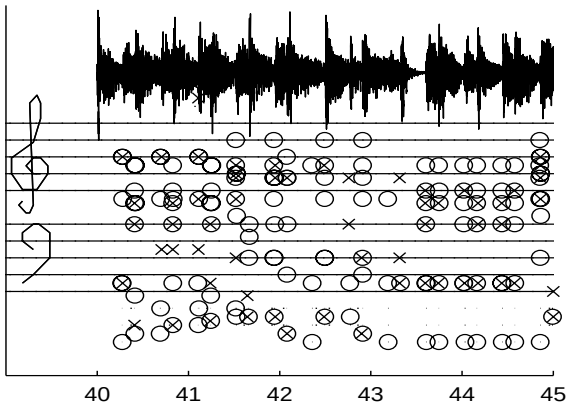


Fig. 8. Transcription of a synthesized MIDI-song, "Waterloo". Circles denote the original score and crosses the transcription. The piece has regular rock drums, not shown in the score.

music. Indeed, musical knowledge or rules were not utilized at all. The system does not even utilize the context, it simply looks at each segment of the input signal at a time and finds the musical notes in it. This is not likely to correspond to the experience of a human listener, but instead, resembles the long-ago abandoned "phonetic typewriter" approach to speech recognition without language models.

Among different types of musical input, acoustic music turned out to be best transcribed. In the absense of drums, even rich polyphonies yield harmonically satisfactory transcriptions. This may be a side-product of the inability to use musical models: the system is at its best when using its musical ear, without having to resort to musical predictions which it is not able to make. The current system as such is already applicable to the chord-level transcription of acoustic music.

An essential challenge in the area of automatic transcription of music is to formalize musical knowledge and statistics of musical material to models which can be used to co-operate with, and direct the attention of, bottom-up signal analysis. Such models could include, for example, the induction and matching of musical patterns, the predictions of which help in otherwise ambiguous situations. A detected musical scale or harmonic state should affect the *a priori* probabilities of different note events. The progression of such harmonic states, in turn, could be statistically modeled in the form of chord *N*-grams, provided that the musical training data is available in a suitable format. Also, the closer the real applications get, the more pragmatic and heuristic rules are likely to appear in order to adapt to users' special cases.

REFERENCES

[1] Moorer, J. A. (1977). "On the Transcription of Musical Sound by Computer," Computer Music Journal, November 1977, 32–38.

[2] Chafe, C., and Jaffe, D. (1986). "Source Separation and Note Identification in Polyphonic Music," Proc. IEEE International Conf. on Acoust., Speech, and Signal Processing, Tokyo, 1289–1292.

[3] Maher, R. C (1990). "Evaluation of a Method for Separating Digitized Duet Signals," J. Audio Engin. Soc., Vol. 38, No.

12, 956–979.

[4] Hawley, M. (1993). “Structure out of sound,” Ph.D. thesis, Massachusetts Institute of Technology.

[5] Rossi, L., Girolami, G., and Leca, M. (1997). “Identification of polyphonic piano signals,” ACUSTICA • acta acustica Vol. 83, 1077–1084.

[6] Katayose, H., and Inokuchi, S. (1989). “The Kansei Music System,” Computer Music Journal, Vol. 13, No. 4, Winter 1989, 72–77.

[7] Kashino, K., Nakadai, K., Kinoshita, T., and Tanaka, H. (1995). “Organization of Hierarchical Perceptual Sounds: Music Scene Analysis with Autonomous Processing Modules and a Quantitative Information Integration Mechanism,” Proc. International Joint Conf. on Artificial Intelligence, Montréal.

[8] Martin, K. D. (1996a). “A Blackboard System for Automatic Transcription of Simple Polyphonic Music,” Massachusetts Institute of Technology Media Laboratory Perceptual Computing Section Technical Report No. 385.

[9] Brown, G. J., and Cooke, M. P. (1994). “Perceptual grouping of musical sounds: A computational model,” J. of New Music Research 23, 107–132.

[10] Godsmark, D., and Brown, G. J. (1999). “A blackboard architecture for computational auditory scene analysis,” Speech Communication 27, 351–366.

[11] Goto, M. (2000). “A robust predominant-F0 estimation method for real-time detection of melody and bass lines in CD recordings,” Proc. IEEE International Conf. on Acoust., Speech, and Signal Processing, Istanbul, Turkey.

[12] Moelants D., Rampazzo C. (1997). “A Computer System for the Automatic Detection of Perceptual Onsets in a Musical Signal”. In Camurri, Antonio (Ed.). “*KANSEI, The Technology of Emotion*”, pp. 140–146. Genova, 1997.

[13] Bilmes J. (1993). “Timing is of the Essence: Perceptual and Computational Techniques for Representing, Learning, and Reproducing Expressive Timing in Percussive Rhythm”. MSc thesis, Massachusetts Instute of Technology, 1993.

[14] Schloss A. (1985). ”On the Automatic Transcription of Percussive Music — From Acoustic Signal to High-Level Analysis”. Ph.D. thesis, Stanford University, 1985. Report STAN-M-27.

[15] Large, Kolen. (1994). “Resonance and the perception of musical meter”. Connection science, 1994, Vol. 6 Issue 2/3.

[16] Scheirer E. “Tempo and Beat Analysis of Acoustic Musical Signals”. Machine Listening Group, MIT Media Laboratory, 1996.

[17] Goto, M., Muraoka, Y. (1996). “Beat Tracking based on Multiple-agent Architecture — A Real-time Beat Tracking System for Audio Signals,” In Proc. Second International Conference on Multiagent Systems, pp.103–110, 1996.

[18] Dixon, S. (1999). “A beat tracking system for audio signals,” in Proc. Conf. Comput. and Mathem. Methods in Music, Vienna, Austria, Dec. 1999.

[19] Klapuri. (1999). “Sound Onset Detection by Applying PsychoacousticKnowledge," In Proc IEEE International Conference on Acoustics, Speech and Signal Processing, 1999.

[20] Lerdahl, F., Jackendoff, R. (1983). “A Generative Theory of Tonal Music”. MIT Press, Cambridge, MA, 1983.

[21] Seppänen, J. (2001). “Tatum grid analysis of musical signals,” In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2001.

[22] Bregman A. S. (1990). *Auditory scene analysis: the perceptual organization of sound.* MIT Press, Cambridge, Massachusetts.

[23] Hess. (1983). “Algorithms and Devices for Pitch Determination of Musical Sound Signals”. Springer–Verlag, Berlin.

[24] Klapuri, Virtanen, Holm. (2000). “Robust multipitch estimation for the analysis and manipulation of polyphonic musical signals,” In Proc. COST-G6 Conference on Digital Audio Effects, DAFx-00, Verona, Italy, 2000.

[25] Klapuri, A. (2001). “Multipitch estimation and sound separation by the spectral smoothness principle,” In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2001.

[26] Klapuri, Virtanen, Eronen, Seppänen. (2001). “Automatic transcription of musical recordings,” In Proc. Consistent & Reliable Acoustic Cues Workshop, CRAC-01, Aalborg, Denmark, September 2001.

[27] Huron, D. (1989). “Voice Denumerability in Polyphonic Music of Homogeneous Timbres,” Music Perception, Summer 1989, Vol. 6, No. 4, 361–382.

[28] Hermansky, H., Morgan, N., Hirsch, H.-G. (1993). “Recognition of speech in additive and convolutiove noise based on RASTA spectral processing,” IEEE International conference on Acoustics, Speech, and Signal Processing, Minneapolis, Minnesota, 1993.

[29] Virtanen, Klapuri. “Separation of Harmonic Sounds Using Multipitch Analysis and Iterative Parameter Estimation,” In Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2001.

[30] Eronen A. "Comparison of features for musical instrument recognition," In Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2001.

[31] Scheirer, Eric D. (1996). “Bregman’s Chimerae: Music Perception as Auditory Scene Analysis”. Presented at the 4th International Conference on Music Perception and Cognition, Montreal, August 1996