AUTOMATIC TRANSCRIPTION OF MUSIC

Anssi P. Klapuri

Institute of Signal Processing,
Tampere University of Technology, P.O.Box 553, FIN-33101 Tampere, Finland
klap@cs.tut.fi

ABSTRACT

The aim of this tutorial paper is to introduce and discuss different approaches to the automatic music transcription problem. The task is here understood as a transformation from an acoustic signal into a MIDI-like symbolic representation. Algorithms are discussed that concern three subproblems. (i) Estimation of the temporal structure of acoustic musical signals, the musical meter. (ii) Estimation of the fundamental frequencies of concurrent musical sounds. (iii) Higher-level musicological modeling to resolve otherwise ambiguous situations. The emphasis is laid on multiple-F0 estimation. Validation experiments are performed using both synthesized and real-world music signals. Demonstration signals are available at http://www.cs.tut.fi/~klap/iiro/smac/.

1. INTRODUCTION

Written music is traditionally presented as a *musical notation* (score) which comprises the times, durations, and pitches of the sounds that constitute a piece. The aim of music transcription is to discover such a musical "recipe" in an acoustic signal, so that a musician or a synthesizer program can reproduce and modify the original performance. Sometimes a more coarse symbolic representation suffices, for example writing down the chords only.

Automatic transcription of music (AToM) is difficult and includes several subproblems to be discussed in the coming sections. Figure 1 gives an idea of the parts an automatic transcriber, although a fixed structure for such systems does not exist. *Multiple-F0 estimation* refers to a core part which estimates the fundamental frequencies (F0s) of several concurrent sounds. Usually, internal models are needed along with the acoustic signal to perform the analysis. Comparing with speech recognition, *musicological models* play the role of a "language model" in music.

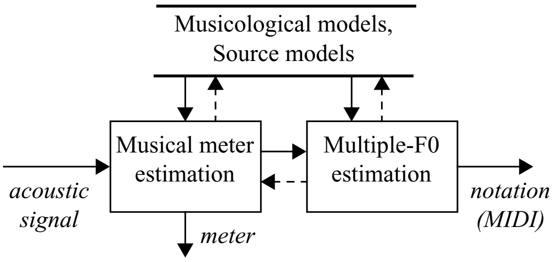

Figure 1. *Relationships of the parts of a transcription system.*

Musical meter characterizes the temporal regularity of a music signal. Musical meter is a hierarchical structure which consists of pulse sensations (periodicities) at different levels [1]. Figure 2 shows the meter at three relatively well-defined levels: *beat* (foot tapping rate), *tatum* (time quantum), and musical *measure* pulse levels. Meter estimation alone has several applications in the synchronization,editing,and analysis of music signals.

In principle, it would be advantageous to perform meter estimation and multiple-F0 estimation in parallel. However, meter estimation can be done relatively robustly and allows the positioning and sizing of the analysis frames in further analysis. For these reasons, it is usually computationally more efficient to perform the two stages in a cascade. Feedback from multiple-F0 stage is needed especially to analyze musical measure pulse which is related to harmonic change rate.

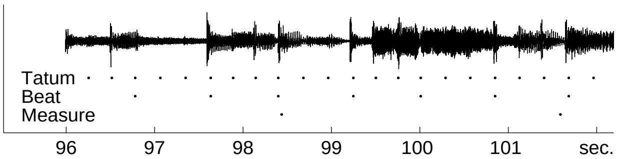

Figure 2. *A musical signal with three metrical levels illustrated.*

2. MUSICAL METER ESTIMATION

Table 1 summarizes some state-of-the-art systems for musical meter estimation. Earlier work has concentrated almost solely on beat tracking, with the few exceptions in [2,3,7]. Most earlier systems have attempted to discover periodicities in symbolic input (MIDI) [2,3,4, see 4 for a review]. Only a few systems have been proposed that process acoustic input signals [4,5,6,7].

Our meter estimation system has been described in [8]. It analyzes meter at the beat, tatum, and measure pulse levels and is able to process acoustic musical signals from all main musical genres. The method consists of three parts. (i) Time-frequency analysis part computes *registral accent signals* which measure the degree of accentuation (emphasis) at four frequency ranges of the input signal. This part is a generalization based on [5] and [6] and has a great impact on performance. (ii) A bank of comb filter resonators is used at each frequency channel to analyze periodicity. The resonators are quite similar to those used by Scheirer in [5]. However, the exact periodicity analysis method is not critical. Among four different periodicity analysis techniques, three performed equally well, and the simplest (comb filters) was chosen. (iii) A musically informed probabilistic model was used to represent dependencies between the different levels of meter and between temporally successive meter estimates. This stage improved the robustness and temporal stability of the system.

3. APPROACHES TO MULTIPLE-F0 ESTIMATION

In this section, we introduce a number of different approaches that have been taken to solve the multiple-F0 estimation problem.

Table 1: *Characteristics of meter estimation systems*

System	Input data	Aim	Target material
Large, Kolen	MIDI	meter	not specified
Goto, Muraoka	audio	meter	pop music, 4/4 time
Scheirer*	audio	beat	mainly ‘strong beat’
Dixon*	MIDI & audio	beat	MIDI: all music types
Temperley*	MIDI	meter	all music types
Klapuri	audio	meter	all music types

* Source codes available for download.

Subheadings are provided to improve readability, but it should be noted that any of the cited papers really cannot be put under a single label. Attempts toward multiple-F0 estimation date back to the 1970s. For a brief historical overview, see [21]. In following, an attempt has been made to list the definitive attributes of a representative set of more recent papers.

3.1 Perceptual grouping of frequency partials

Any algorithm that finds the F0s of multiple concurrent sounds is, in effect, also organizing spectral componets to sound sources [9,p.240]. This organization task is called *auditory scene analysis* (ASA) and human auditory system is very good at it. An important step forward in ASA was taken when Bregman pointed out distinct perceptual cues for grouping time-frequency components into sources. The cues (component features) were briefly: proximity in time-frequency, harmonic frequency relationships, synchronous changes, and spatial proximity [9].

Kashino *et al*. brought Bregman’s ideas to music scene analysis and also proposed several other new ideas for AToM [10]. The front-end of their system used a “pinching plane method” to extract continuous frequency components from the input data. These were clustered into note hypotheses by applying the above mentioned perceptual rules. *Timbre models* were used to identify the source of each note and pre-stored *tone memories* were used to resolve overlapping frequency components. Chordal analysis was performed based on the probabilities of notes to occur under a given chord. Chord transitions probabilities were encoded into trigram models (Markov chains). For computations, a Bayesian probability network was used to integrate the knowledge and to do simultaneously bottom-up analysis, temporal tying, and top-down processing (chords predict notes and notes predict components). Evaluation material comprised five different instruments and polyphonies of about three simultaneous sounds. The work still stands among the most elegant and complete AToM systems.

A more recent example of the perceptual grouping approach is the PhD work of Sterian [11]. He used Kalman filtering to extract continuous sinusoidal partials and represented the grouping rules as a set of likelihood functions which evaluated the likelihood of observed partials given a hypothesized grouping.

3.2 Auditory-model based approach

The “unitary pitch model” of Meddis and Hewitt has had a strong influence on F0 estimation research [12]. While Bregman’s theory focused on what happens in the brain, Meddis and Hewitt modeled the more peripheral (largely physiological) parts of hearing. Although multipitch estimation in sound mixtures was not addressed, research to this direction was inspired, too.

Cheveigne and Kawahara extended the unitary pitch model to the multiple-F0 case. They proposed a system where pitch estimation was followed by the cancellation of the detected sound, and the estimation was repeated for the residual signal [13]. Also a computationally exhaustive joint estimator was proposed. Although evaluation results were shown for a rather artificial data, the proposed iterative scheme was really a successful one. Temporal continuity between frames was not considered.

Tolonen and Karjalainen developed a computationally efficient version of the unitary pitch model and applied it to the multiple-F0 estimation of musical sounds [14]. The spectrum of an incoming sound was flattened using inverse warped-linear-prediction filtering. In pitch computations, only two frequency channels were used instead of the 40–120 channels in the original model, yet the characteristics of the model were mainly preserved. Practical robustness of the system was improved by introducing a FFT-based “generalized autocorrelation” method. Extension to multipitch estimation was achieved by cancelling subharmonics in the output of the model. From the resulting *enhanced summary autocorrelation function*, all F0s can be picked without iterative estimation and cancellation. The method is relatively accurate and can be implemented based on [14].

3.3 Blackboard architecture with auditory front-end

Blackboard architectures were developed to facilitate the integration of different types of knowledge for signal analysis. The *blackboard* is hierarchy of data representations at different analysis (abstraction) levels. The data is common to a set of autonomous *knowledge sources* which operate when requested.

Martin proposed a system for transcribing piano performances of four-voice Bach chorales [15]. In his system, an auditory model was used as a front-end to a blackboard, which employed knowledge about physical sound production, rules governing tonal music, and “garbage collection” heuristics. Support for F0s was raised on a frame-by-frame basis and then combined with the longer-term power envelope information to create note hypotheses. Musical rules favoured F0s in certain intervallic relations.

A more recent model of Godsmark and Brown was particularly designed to facilitate the integration of different auditory organization principles and competition between them [16]. The applied auditory front-end produced “synchrony strands” each of which represented a dominant time-frequency component. These were fused to sound events by extracting features from each strand and by applying Bregman’s primitive organization principles. Sound events were further grouped to their respective sources (event ”streams”) by computing the pitch and timbre proximity between sounds. Musical meter information was used to predict when events will occur and melodic pattern induction to predict recurrent patterns. The model was evaluated by showing that it could segregate melodic lines from polyphonic music and to resolve interleaved melodies. Transcription accuracy as such was not the main goal.

3.4 Signal-model based probabilistic inference

It is possible to state the whole multiple-F0 estimation problem in terms of a signal model, the parameters of which should be estimated. Consider e.g. the model [17]:

$$y_t = \left\{ \sum_{k=1}^{K} \sum_{m=1}^{M_k} a_{k,m} \cos[m\omega_k t] + b_{k,m} \sin[m\omega_k t] \right\} + v_t$$

where K is the number of simultaneous notes, M_k is the number of partials in note k, ω_k is the fundamental frequency of note k, and $a_{k,m}$, $b_{k,m}$ together encode the amplitude and phase of individual partials. The term v_t is a residual noise component.

In principle, *all* the parameters on the right-hand side of the above equation should be estimated, based on the observation y_t and possible prior knowledge about the parameter distributions. As pointed out by Davy *et al.* in [17], the problem is Bayesian since there is a lot of prior knowledge concerning music signals.

Davy and Godsill elaborated the above signal model to accommodate time-varying amplitudes, non-ideal harmonicity, and non-white residual noise [17]. Prior distributions for the parameters were carefully selected. An input signal was first segmented into excerpts where no note transitions occur. Then the parameters of the signal model were estimated in *time domain*, separately for each segment. The main problem of this approach is in the actual computations. For any sufficiently realistic signal model, the parameter space is huge and the posterior distribution is highly multimodal and strongly peaked. Davy and Godsill used Markov chain Monte Carlo sampling of the posterior, reporting that much of the innovative work was spent on finding heuristics for the fast exploration of the parameter space [17]. Although computatinally inefficient, the system was reported to work robustly for polyphonies up to three simultaneous sounds.

Goto has proposed a method which models the *short-time spectrum* of a music signal as a weighted mixture of tone models [18]. Each tone model consists of a fixed number of harmonic components which are modeled as a Gaussian distributions centered at integer multiples of the F0 in the spectrum. Goto derived an expectation-maximization (EM) algorithm which iteratively updates the weights of each tone model and their relative harmonic amplitudes, leading to a maximum *a posteriori* estimate. Temporal continuity was considered by tracking framewise F0 weights in a multiple-agent architecture. Goto used the algorithm successfully to detect melody and bass lines in CD recordings.

3.5 Data-adaptive techniques

In data-adaptive systems, there is no parametric model or other knowledge of the sources. Instead, the source signals are estimated from the data. Typically, it is not even assumed that the sources (notes) have harmonic spectra! For real-world signals, the performance of e.g. independent component analysis alone is poor. However, by placing restrictions for the sources, the data-adaptive techniques become applicable in realistic cases. Such restrictions are e.g. independence of sources and *sparseness* which means that the sources are inactive most of the time.

Virtanen added *temporal continuity* constraint to the sparse coding paradigm [20]. He used the signal model

$$X_t(f) = \sum_{n=1}^{N} a_{t,n} S_n(f) + E_t(f), \qquad (1)$$

which represents the zero-phase power spectrogram $X_t(f)$ of the input as a linear sum of N static source spectra $S_n(f)$ with time-varying gains $a_{t,n}$. The term $E_t(f)$ is error spectrum. An iterative optimization algorithm is proposed which estimates non-negative $a_{t,n}$ and $S_n(f)$ based on the minimization of a cost function which takes into account reconstruction error, sparseness, and temporal continuity. The algorithm was used to separate pitched and drum instruments from real-world music signals.

Another recent example of applying sparse coding to music is that of Abdallah and Plumbley [19].

3.6 Our system

The multiple-F0 estimator proposed by us in [21] is closest to the auditory-model based approach (Sec. 3.2), since psychoacoustics was used as a base of the analysis principles. However, the method is better characterized as problem-solving oriented, focusing on the AToM application, not on auditory modeling.

There is a number of identifiable problems which a multiple-F0 estimator must address in order to resolve real-world music signals. (i) The likelihoods of different F0s must be calculated robustly in the presence of other, co-occurring sounds. (ii) The effect of a true F0 must be cancelled from its harmonics and subharmonics which usually appear as the next-most-likely F0s. (iii) Number of concurrent sounds has to be estimated. (iv) Real-world sounds are often not perfectly harmonic. (v) Generality in regard to different instruments. (vi) Robustness in noise.

Our multiple-F0 method is based on the iterative estimation and cancellation approach (see Sec. 3.2 and [21]). At the estimation stage, problem (i) is addressed by estimating weights (likelihoods) of F0s independently at separate frequency bands, and by letting only a set of selected frequency samples to contribute to the weight, not the overall spectrum. Problem (iv) is addressed by letting the series of partials at subbands be shifted in frequency. Problem (ii) is addressed by the iterative scheme where the spectrum of a detected sound is estimated and subtracted from the mixture, after which estimation is repeated for the residual. Statistical distribution of the features of detected sounds is used to control the stopping of the iteration and thus to estimate polyphony. Points (v) and (vi) are addressed in a preprocessing step which flattens sound spectra and suppresses noise. All the computations are done for a single frame in the frequency domain.

4. MUSICOLOGICAL MODELS

Linguistic information allows speech recognition systems to interpret obscured and ambiguous signals. Musical information is equally important for music transcription. Some systems use such internal models [10,15,16,18], as already described in Sec. 4.

Temperley has presented a very comprehensive rule-based system which models the cognition of basic musical structures [3]. Such a system could be readily used as a post-processor for a musically agnostic transcription system. However, to really benefit of the musical knowledge, it should be utilized already *during* the analysis. Also, while a rule-based model reveals a lot about the human cognition, *probabilistic* models are advantageous in that they evaluate the likelihoods of several candidate analyses.

Consider the following experiment. We represent c*hord unigrams* as 12-bit numbers where each bit signifies the presence/absence of one of the 12 pitch classes (octave equivalence). There are 4096 such unigrams. 359 MIDI pieces were collected and cut into segments where note onsets or offsets do not occur. The harmonic content of each segment was then represented with the appropriate unigram. The probability of occurrence for each uni-

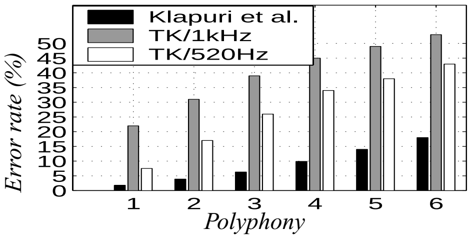

Figure 3. *Error rates of the system of Klapuri and the reference system of Tolonen et al. (with tighter upper limits for F0s).*

gram was computed within the pieces and averaged over all pieces. The results were interesting: among the 30 most probable unigrams were the 12 single notes, seven different major triad chords, five minor triads, and three minor-seventh chords.

The described kind of "brute force" statistical approach has several advantages. (i) The estimated probabilities can be used to rate the likelihoods of several competing F0 mixtures in an AToM system. (ii) The above procedure has *no* heuristic parameters. (iii) No musical expertise was employed, yet the system knows about major and minor triads, the building blocks of Western harmony. (iv) New kind of musical material can be focused simply by retraining the system using the target material.

5. SIMULATION EXPERIMENTS

For meter estimation, we have earlier presented a statistical evaluation of different beat tracking systems using a database of 478 musical pieces and implementations of the original authors. This will not be repeated here. An interested reader is referred to [8].

Reliable comparison of multiple-F0 estimation algorithms is difficult because the systems are typically very complex and source codes are not available. However, a few methods can be replicated. The core EM-algorithm of Goto is implementable based on [18] (but not front-end and post-processing). The EM-algorithm estimates the weights of all F0s, but usually only the predominant F0 was found in our simulations, exactly as claimed by Goto. Also, Tolonen and Karjalainen describe their algorithm and its parametes to sufficient detail to be exactly implementable [14]. Error rates for this reference system (*TK*) and our system are given in Figure 3. As reported by the authors, the method cannot handle "spectral pitch", i.e., F0s above 1 kHz. It was further found out here that the method is best at detecting F0s in the three-octave range between 65 Hz and 520 Hz. Thus, in the simulations the mixtures given to the *TK* method were restricted to contain F0s only below either 520 Hz or 1 kHz, as specified in the results. Given the compactness of the *TK* algorithm, the results are very good.The number of sounds was known to both methods.

Transcription demonstrations for real-world music and synthesized MIDI are at http://www.cs.tut.fi/~klap/iiro/smac/.

6. REFERENCES

[1] Lerdahl, F., Jackendoff, R., *A Generative Theory of Tonal Music*. MIT Press, Cambridge, MA, 1983.

[2] Large, E. W., Kolen, J. F., "Resonance and the perception of musical meter". Connection science, 6(1), 1994,pp.177-208.

[3] Temperley, D. *Cognition of Basic Musical Structures*. MIT Press, Cambridge, MA, 2001.

[4] Dixon, S., "Automatic Extraction of Tempo and Beat from Expressive Performances," *J. New Music Research* 30 (1), 2001, pp. 39-58.

[5] Scheirer, E. D., "Tempo and beat analysis of acoustic musical signals," *J. Acoust. Soc. Am.* 103 (1), 1998, pp. 588-601.

[6] Goto, M., Muraoka, Y., "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.

[7] Goto, M.,Muraoka,Y., "Issues in Evaluating Beat Tracking Systems," *IJCAI-1997 Workshop on Issues in AI and Music*.

[8] Klapuri, A. P., "Musical meter estimation and music transcription," In Proc. Cambridge Music Processing Colloquium, Cambridge University, UK, March 2003.

[9] Bregman, A.S.,"Auditory Scene Analysis," MIT Press,1990.

[10] Kashino, K., Nakadai, K., Kinoshita, T., and Tanaka, H., "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, 1995.

[11] Sterian, A. D., "Model-based segmentation of time-frequency images for musical transcription," Ph.D.thesis, University of Michigan, 1999.

[12] Meddis, R. and Hewitt, M. J., "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification," J.Acoust. Soc. Am. 89(6),p.2866–2882, 1991.

[13] de Cheveigné, A. and Kawahara, H., "Multiple period estimation and pitch perception model," Speech Communication 27, pp. 175–185, 1999.

[14] Tolonen, T. and Karjalainen, M., "A computationally efficient multipitch analysis model," IEEE Trans. Speech Audio Processing, Vol. 8, No. 6, pp. 708-716, Nov. 2000.

[15] Martin, K. D., "Automatic transcription of simple polyphonic music: robust front end processing," Massachusetts Institute of Technology Media Laboratory Perceptual Computing Section Technical Report No. 399, 1996.

[16] Godsmark, D. and Brown, G. J., "A blackboard architecture for computational auditory scene analysis," Speech Communication 27, pp. 351–366, 1999.

[17] Davy, M. and Godsill, S. J., "Bayesian harmonic models for musical signal analysis, " In J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, editors, *Bayesian Statistics VII*, Oxford University Press, 2003.

[18] Goto, M., "A predominant-F0 estimation method for real-world musical audio signals: MAP estimation for incorporating prior knowledge about F0s and tone models," in Proc. Workshop on Consistent and reliable acoustic cues for sound analysis, Aalborg, Denmark, Sep. 2001.

[19] Abdallah, S. A. and Plumbley, M. D., "Sparse coding of music signals," Submitted for Publication.

[20] Virtanen, T., "Sound source separation using sparse coding with temporal continuity objective," International Computer Music Conference, Singapore, 2003.

[21] Klapuri, A., Virtanen, T., Holm, J.–M., "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.