Literature Review on Multi-pitch Estimation

Nguyen Tien Dung, Luong Chi Vu, Feng Shuyu
HT055497N, U027353N, HT050618J
{g0505497, u0201065, g0500618}@nus.edu.sg

I. Introduction

Music transcription is to transform an acoustic signal into a symbolic representation. For monophonic music, where only one note sounds at a specific time period, pitch (fundamental frequency, F0) of a note can be calculated by using spectral-location type F0 estimator, spectral-interval type F0 estimator, or periodicity of time-domain amplitude envelope [1]. However, for polyphonic music, where multiple notes are played simultaneously, approaches for single pitch estimation can not directly applied. During the last 10 years, there are many researches on automatic transcription of polyphonic music. Long-term research projects have been undertaken at Stanford University, University of Michigan, University of Tokyo, Massachusetts Institute of Technology, University of London, Cambridge University, and at Tampere University of Technology. In this paper, we will take a literature review of major methods in polyphonic music transcription.

In the following, we categorize existing methods according to [1], and add some recent papers (in red color).

II. Related Work

A. Auditory-scene-analysis based approach

Multiple-F0 estimation is closely related to sound separation. An algorithm that is able to estimate the F0s of several concurrent sounds is, in effect, also organizing the respective spectral components to their sound source ([2], p.240). Bregman's theory is primarily concerned with the **psychology** of auditory perception.

Auditory scene analysis (ASA) is a cognitive function used by human auditory system, which is very effective in perceiving and recognizing individual sound sources in mixture signals. Computational ASA (CASA) is usually viewed as a two-stage process: it first decomposes an incoming signal into its elementary time-frequency components, i.e. *sinusoidal tracks* which are time-varying sinusoidal components, and then organized them to their respective sound source. Some acoustic "cues" can promote the grouping of time-frequency components to the same sound source in human listeners. For example, proximity in time-frequency, harmonic frequency relationships, synchronous changes in the frequency or amplitude of the components, and spatial proximity (i.e., the same direction of arrival).

[3], [4] introduce methods to extract sinusoidal tracks from music signals. The basic algorithms are ([3]): peak detection and peak continuation. For each frame, peaks of spectrum are detected. A "peak" is defined as a local maximum in the magnitude spectrum, and the only practical constraints to be made in the peak search are to have a

frequency range and a magnitude threshold. As there are interactions between different components, the shapes of the spectral peaks cannot be detected without tolerating some mismatch. A practical solution is to detect as many peaks as possible and delay the decision of what is a "well behaved" partial to the next step in the analysis: peak continuation algorithm. Then the peak continuation algorithm adds the spectral peaks of several consecutive frames to peak trajectories.

[5], [6], [7], [8] all use auditory-scene-analysis approach based on sinusoid tracks.

[5] introduces a system, OPTIMA, for musical auditory scene analysis. The sinusoid tracks are integrated into a hierarchy of higher-level constructs: notes, chords, and rhythms by using some knowledge sources. Some of the knowledge sources are based upon statistical analysis of several tonal songs. These analysis provides information such as probabilities of notes being played for a given chord and probabilities of chord transitions encoded into trigram models (Markov chain). Additional knowledge includes tone memory (spectral characteristics of five instruments at varying amplitude levels and pitch), timbre models (an 11-dimensional feature space for recognizing instruments), and perceptual rules for sound separation (only harmonicity and onset timing rules are implemented). For computation, the knowledge sources and sinusoid tracks are integrated by a Bayesian probability network (hypothesis network), in which probabilities from the knowledge sources are used to identify which one of several hypotheses best explains component, note, and chord information at a given time. [6] refines the previous approach by allowing note-transition links to be included in the hypothesis network, which incorporate transition probabilities of note intervals. It also uses a simulated annealing technique to search over various configurations of the hypothesis network.

In Sterian's Ph.D. thesis [7], sinusoid tracks are partitioned into groups representing note events and a list of false alarm tracks not associated with any note event. To choose the most likely partitioning, two independent components are proposed: a likelihood function describing the probability of any given hypothesized partition, and a search algorithm that use this function to identify the most likely partition. While an exhaustive search over all possible partitions is not possible, a multiple-hypothesis tracking strategy is used to find a suboptimal solution.

Godsmark and Brown [8] used a computational model of the peripheral auditory system to extract synchrony strands" (dominant time-frequency components in different bands) for which the grouping cues were extracted. The blackboard" architecture applied in their system was particularly designed to facilitate the integration of, and competition between, the perceptual grouping principles. The strands were organized to sound events and these were further grouped to their respective sources (event streams") by computing pitch and timbre proximities between successive sound events. The model was evaluated by showing that it could segregate melodic lines from polyphonic music. Transcription accuracy as such was not the main goal.

B. Auditory periphery modeling based approach

Meddis et al. proposes "unitary" pitch model [9], [10], which addresses the more peripheral (largely **physiological**) parts of hearing. Simulation experiments show that this model is capable of reproducing a wide range of phenomena in human pitch perception.

The unitary model analyzes the amplitude-envelope periodicity at the outputs of a bank of bandpass filters by the following processing steps([10]):

1) an acoustic input signal is passed though a bank of 40-120 bandpass filters;
2) the signal in each channel is compressed, half-wave rectified, and lowpass filtered;
3) periodicity estimation within channels is carried out by calculating short-time autocorrelation function (ACF) estimates (Equation 1 calculates autocorrelation function $r(\tau)$);

   $$r(\tau) = IDFT(|DFT(x(n))^2|) \tag{1}$$

   which can be expressed in terms of the Fourier spectrum $X(k)$ of a real-valued input signal (Equation 2)

   $$r(\tau) = \frac{1}{K}\sum_{k=0}^{K-1}[\cos(\frac{2\pi\tau k}{K})|X(k)|^2] \tag{2}$$

   where K is the length of the transform frame. The function $cos(2\pi\tau k/K)$ assigns unity weights to spectral components at integer multiples of a F0 candidate τ/f_s, where f_s is the sampling rate.
4) the ACF estimates are summed across channels to obtain a *summary autocorrelation* function (Equation 3):

   $$s(\tau) = \sum_c r_c(\tau) \tag{3}$$

   where $r_c(\tau)$ is the autocorrelation function at subband c. The maximum value of $s(\tau)$ is then used to indicate the perceived pitch period.

[11] extends the unitary model to multipitch estimation in mixture signals. They propose a system where pitch estimation is followed by cancellation of the detected sound, and the estimation is then repeated for the residual signal iteratively. The cancellation is performed either by subband selection or by performing within-band cancellation filtering. Evaluation results are reported only for synthetic and perfectly periodic signals.

[12] developed a computationally efficient version of the unitary pitch model and applied it to the multiple-F0 estimation of musical sounds. Only two subbands were used instead of the 40-120 bands in the original model, yet the main characteristics of the model were preserved. Practical robustness was addressed by flattening the spectrum of an incoming sound by inverse warped-linear-prediction filtering, and by using the generalized ACF ([12]) for periodicity estimation. Extension to multiple-F0 estimation was achieved by canceling subharmonics in the output of the model. From the resulting enhanced summary autocorrelation function, all F0s were picked without iterative estimation and cancellation. The method is relatively accurate and has been statistically evaluated [13].

[14] uses a blackboard architecture to integrate knowledge source about physical sound production, rules governing tonal music, and garbage collection heuristics. Support for different F0s was raised on a frame-by-frame basis and then combined with longer-term power-envelope information to create note hypotheses. Musical rules favoured F0s in certain intervallic relations.

[15], [16] proposed certain modifications to the unitary pitch model in order to obtain a reliable multiple-F0 estimation tool for use in music signals. The first two steps of the unitary model (bandpass filtering, compression, and rectification) were retained but the ACF calculations were replaced by a technique called harmonic selection

and a more complex subband-weighting was applied when combining the results across bands. Computational efficiency was achieved by approximating the HWR operation in the frequency domain according to Equation 4. The method was evaluated by calculating error rates for random mixtures of recorded musical instrument samples. Good accuracy was achieved using analysis frame sizes of 46 ms or longer.

$$\hat{Y}(k) = \frac{\sigma_x}{\sqrt{8\pi}}\delta(k) + \frac{1}{2}X(k) + \frac{1}{\sigma_x\sqrt{8\pi}}\sum_{j=-K/2+k}^{K/2-k} X(j)X(k-j) \tag{4}$$

where $\delta(k)$ is the unit impulse function and σ_x is the standard deviation of $x(n)$. On the right-hand side of this equation, the first term is a dc-component, the second term represents the spectrum of the input signal, and the last term represents the beating components of the amplitude-envelope spectrum.

C. Signal-model based probabilistic inference

As musical signals are highly structure, it is possible to state the whole multiple-F0 estimation problem in terms of a signal model, where parameters should be estimated.

[17] elaborated the signal model (Equation 5) to accommodate time-varying amplitudes, non-ideal harmonicity, and non-white residual noise.

$$y(t) = \sum_{n=1}^{N}\sum_{m=1}^{M_n}[a_{n,m}\cos(m\omega_n t) + b_{n,m}\sin(m\omega_n t)] + e(t) \tag{5}$$

where N is the number of simultaneous sounds, M_n is the number of partials in sound n, ω_n is the fundamental frequency of sound n, and $a_{n,m}$, $b_{n,m}$ together encode the amplitude and phase of individual partials. $e(t)$ is a residual noise component.

A likelihood function for observing $y(t)$ given model parameters was defined. 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 the 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 variable-dimension 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. Although computationally inefficient, the system was reported to work quite robustly for polyphonies up to three simultaneous sounds.

[18] has proposed a method which models the short-time spectrum of a music signal as a weighted mixture of tone models. Each tone model consists of a fixed number of harmonic components which are modeled as Gaussian distributions centered on integer multiples of the F0 in the spectrum. Goto derived a computationally feasible expectation- maximization (EM) algorithm which iteratively updates the tone models and their weights, leading to maximum a posteriori parameter estimates. **Temporal continuity was considered by tracking framewise F0 weights within a multiple-agent architecture**. Goto used the algorithm successfully to track the melody and the bass lines in real-time on CD recordings. Although the overall system of Goto is relatively complex, the core EM

algorithm can be easily implement based on the reference. The algorithm estimates the weights of all F0s, but typically only one (predominant) F0 was found in our simulations, exactly as claimed by Goto.

In [19], the salience $s(\tau)$ of a period candidate τ is modeled as

$$s(\tau) = \sum_{m=1}^{M} g(\tau, m)|Y(f_{\tau,m})| \tag{6}$$

where $f_{\tau,m} = mf_s/\tau$ is the frequency of the mth harmonic partial of a F0 candidate f_s/τ, f_s is the sampling rate, and function $g(\tau, m)$ defines the weight of partial m of period τ in sum. The weight function $g(\tau, m)$ is got through training the sample sets. It is reported this method outperforms two previous methods of the same author [13], [16].

D. Data-adaptive approach

The source signals are estimated from the data. Typically, it is not even assumed that the source (which here refer to individual notes) have harmonic spectra. However, for real-world signal, the performance of these methods are poor. By placing certain restrictions on the sources, these methods become applicable in realistic cases. Such restrictions are, e.g., **independence of the sources and *sparseness* which means that the sources are assumed to be inactive most of the time**.

[20] added temporal continuity constraint to the sparse coding paradigm. He used the signal model 7, where the power spectrogram of the input, $S(t, f)$ is represented as a linear sum of N static source spectra $S_n(f)$ with time-varying gains at n. The term $S_e(t, f)$ represents the error spectrogram. Virtanen proposed an iterative optimization algorithm which estimates non-negative at 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 in real-world music signals.

$$S(t, f) = \sum_{n=1}^{N} a_{t,n} S_n(f) + S_e(t, f) \tag{7}$$

[21], [22] are unsupervised learning techniques, and [23] is a classification approach.

[22] applied sparse coding for the analysis of music signals. Input data was represented as magnitude spectrograms, and sources as magnitude spectra, leading to a source mixing model which is essentially the same as in Equation (9). The authors proposed an algorithm where sources were obtained using gradient-ascent inference and the time-varying gains with maximum-likelihood learning. Their results were promising, although shown only for one example case, a synthesized Bach piece with two to three simultaneous sounds.

E. Other approach

[13] proposes a frequency-domain separation methods for multi-F0 estimation. It is a progmatic approach which decomposes the problem into smaller subproblems, which are noise suppression, predominant-F0 estimation, spectral smoothing and detected sound removing, and solve them one-by one. It is a "complete" multiple-F0 estimation system in the sense that it includes mechanisms for suppressing additive noise and for estimating the number of

concurrent sounds in an input signal ([15]), both of which are not solved in [16]. However, it suffers two major weakness. First, as the iterative cancellation of the detected sound is performed by separating the spectra of the sounds in the frequency domain, the residual signal maybe corrupts, resulting estimation and separation of the individual higher-order harmonic partials not reliable. Second, the successfulness of frequency-domain separation depends highly on the resolution of the spectrum, which is longer analysis frame is preferred. An advantage of [16] is that it constitutes an "explicit" reference implementation of many basic mechanisms that are needed for successful multiple-F0 estimation. Such an implementation is quite instructive in understanding the acoustic and musical constraints of the problem, not only the auditory point of view.

There are batch of other approaches, which can refer to [1].

References

[1] A.P. Klapuri. Automatic music transcription as we know it today. In *Journal of New Music Research*, pages 269–282, 2004.

[2] A.S. Bregman. *Auditory Scene Analysis*. MIT Press., Cambridge, MA, 1990.

[3] X. Serra. Musical sound modeling with sinusoids plus noise. In *Roads, C., Pope, S., Picialli, A., De Poli, G. (eds.), Musical signal processing*, 1997.

[4] S.N. Levine. *Audio representation for data compression and compressed domain processing*. PhD thesis, University of Stanford, 1998.

[5] K. Kashino and H. Tanaka. A sound source separation system with the ability of automatic tone modeling. In *Proc. International Computer Music Conference*, pages 248–255, Tokyo, 1993.

[6] K. Kashino and N. Hagita. A music scene analysis system with the mrf-based information integration scheme. In *Proc. of International Conference on Pattern Recognition (ICPR)*, volume 2, pages 725–729, 1996.

[7] Andrew Sterian. *Model-Based Segmentation of Time-Frequency Images for Musical Transcription*. PhD thesis, University of Michigan, MI, USA, 1999.

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

[9] R. Meddis and M. J. Hewitt. Virtual pitch and phase sensitivity of a computer model of the auditory periphery. i: Pitch identification. In *Journal of the Acoustical Society of America*, pages 2866–2882, 1991.

[10] R. Meddis and L. O'Mard. A unitary model of pitch perception. In *Journal of the Acoustical Society of America*, pages 1811–1820, 1997.

[11] A. Cheveigné and H. Kawahara. Multiple period estimation and pitch perception model. *Speech Communication*, 27:175–185, 1999.

[12] T. Tolonen and M. Karjalainen. A computationally efficient multipitch analysis model. *IEEE Trans. Speech Audio Processing*, 8:708–716, 2000.

[13] A.P. Klapuri. Multiple fundamental frequency estimation by harmonicity and spectral smoothness. *IEEE Trans. Speech and Audio Processing*, 11:804–816, 2003.

[14] K. D. Martin. *Sound-Source Recognition: A Theory and Computational Model*. PhD thesis, Department of Electrical Engineering and Computer Science, MIT, June.

[15] A.P. Klapuri. *Signal processing methods for the automatic transcription of music*. PhD thesis, Tampere University of Technology, Finland, 2004.

[16] A.P. Klapuri. A perceptually motivated multiple-f0 estimation method. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, Oct. 2005.

[17] M. Davy and S.J. Godsill. Bayesian harmonic models for musical signal analysis. *J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith (Eds.), Bayesian Statistics VII*, 2003.

[18] M. Goto. A predominant-f0 estimation method for realworld 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, 2001.

[19] A.P. Klapuri. Multiple fundamental frequency estimation by summing harmonic amplitudes. In *7th International Conference on Music Information Retrieval (ICMIR)*, Victoria, Canada, Oct. 2006.

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

[21] P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 2003.

[22] S. A. Abdallah and M. D. Plumbley. Unsupervised analysis of polyphonic music by sparse coding. *IEEE Transactions on Neural Networks*, 17:179–196, Jan. 2006.

[23] G. E. Poliner and D. P. W. Ellis. A classification approach to melody transcription. In *6th International Conference on Music Information Retrieval*, London, UK, 2005.