PERCEPTUALLY MOTIVATED QUASI-PERIODIC SIGNAL SELECTION FOR POLYPHONIC MUSIC TRANSCRIPTION

Mahdi Triki[†] *and Dirk T.M. Slock* [*]

[†] Philips Research Laboratories, Eindhoven, The Netherlands
[*] Eurecom, Sophia Antipolis, France
Email: mahdi.triki@philips.com, dirk.slock@eurecom.fr

ABSTRACT

A multiple fundamental frequency estimator is a key building block in music transcription and indexing operations. However, systems trying to perform this task tend to be very complex [1]. Indeed, music transcription requires an analysis accounting for both physical and psycho-acoustical matters. In this work, we propose a physically-motivated audio signal analysis followed by an auditory-based selection. The audio signal model allows for a better time/frequency resolution tradeoff, while the auditory distance discards the redundant/non-relevant information. No prior information on the musical instrument, musical genre, and/or maximum polyphony are needed. Simulations show that the proposed technique achieves good transcription results for a variety of string and wind instruments. The proposed scheme is also shown to be robust in the presence of noise, percussive sounds and in unbalanced Signal-to-Interference Ratio (SIR) situations.

Index Terms— music transcription, pitch recognition, frequency-selective, amplitude modulation, perceptual model

1. INTRODUCTION

Transcription of music refers to the process of converting audio signals (of performed music) into symbolic representation of music scores. Conventionally, music transcriptions are written by well trained experts (most probably experienced musicians), which is an expensive and time-consuming procedure. In addition to the straight application itself, automatic transcription has a wide range of potential applications including automatic music analysis, music manipulation (e.g. changing the timbre) and music information retrieval (both in building music databases and in transcribing the query input).

The automatic transcription of real-world music is an extremely challenging task. Indeed, the transcription operation requires an analysis accounting for both physical and psycho-acoustical issues: the process has to consider the relationship between the sound as physical phenomena and the sound perception of the human ear.
On one hand, the hearing system performs very well in complex sound mixtures. Humans are able to hear the pitches of several co-occurring sounds and human musicians are the best music transcribers for the time being. This fact inspires auditory motivated approaches which are built upon computational models of human pitch perception [2, 3]. The majority of these approaches consist of a model of the peripheral auditory system followed by 'some' pitch retrival scheme. Typically, the auditory block (front-end) is composed of a cascade of an auditory filterbank (modeling the movement of the basilar membrane of the internal human ear) and a memory-less transform (that models the hair cell transduction).
On the other hand, from a signal-processing point of view, accessing the high-level information contained in audio signals is complex and requires sophisticated tools. Many previous studies have pointed to three major features that summarize the spectral information contained in an audio signal at a given time: the pitch, the dynamics and the timbre. The *pitch* is related to the perception of the fundamental frequency of the sound and indicates how 'high' or 'low' a note sounds. The *dynamics* refers to the amplitude (and the energy) of the wave and indicates how 'loud' or 'soft' a note is. The *timbre* corresponds to the harmonic series in the frequency domain and characterizes the resonance in the body of the instrument. Each of these features is important for the note detection and recognition task. Moreover, as music transcription aims both to *detect the 'position'* and to *recognize the 'content'* of the musical event (musical notes and effects such as vibrato, glissando, etc.), the processing needs both good temporal and frequency resolutions.

Typically, a pitch (F0) retrieval system contains two building blocks: salience evaluation and pitch selection. These blocks could be organized in a successive, joint, or cyclic manner [1].
The first stage of pitch retrieval is the evaluation of the salience, or strength, function at the different candidate periods. Classically, the salience is inferred as a weighted sum of the harmonic partials of a given pitch candidate, i.e.,

$$Sl(\tau) = \sum_{p=1}^{P} g(\tau, p) Y(f_{\tau,p}) \quad (1)$$

where $f_{\tau,p} = pf_s/\tau$ is the frequency of the p^{th} harmonic of the pitch candidate τ (f_s is the sampling frequency). $Y(f)$ may represent the power [4], amplitude [1] or wavelet [5] spectrum of the input audio signal $y(n)$. P denotes the number of considered harmonics. The function $g(\tau, p)$ defines the weight of the p^{th} partial of the period τ in the sum. Several approaches are proposed to set these key parameters using prior informatioon about musical instrument [1, 6], spectral smoothing considerations [7], or based on psychoacoustic theory [3]. In the present paper, we propose the Quasi-Periodic Signal Extraction (QPSE) technique [9] to evaluate the salience function. The QPSE can be interpreted as a sum of scaled, translated and modulated harmonic atoms. However, contrary to the classic atomic decomposition approaches (STFT, WT), the dictionary is not fixed: the atoms are adapted taking into consideration the structure of the

[*] Eurecom research is partially supported by its industrial members: BMW, Bouygues Télécom, Cisco Systems, France Télécom, Hitachi Europe, SFR, Sharp, STMicroelectronics, Swisscom, Thales.

The herein disclosed information is secret until April 19^{th} 2009; eyes only for the appointed reviewers of this conference.

received signal [8]. The proposed technique is shown to be suitable for the analysis of several string and wind instruments and leads to good monophonic transcription accuracy [10].
The second stage in a pitch retrieval system is note selection. This can be performed by peak-picking in the salience domain [1]. More sophisticated engine such as genetic search [11] and token-passing [12] algorithms were proposed to look for the most probable note combination. In this paper, we propose using a psychoacoustic distance (introduced by S. Van de Par et al. in [13]) to rank and select the musical notes present in the mixture. Information on the maximal note number and/or allowed octaves could help increase the transcription accuracy, but are not necessary.

Figure 1 shows the block diagram of the proposed scheme. The two building block are presented respectively in Sections 2 and 3. Simulation results are shown in Section 4. Finally, a discussion and concluding remarks are provided in Section 5.

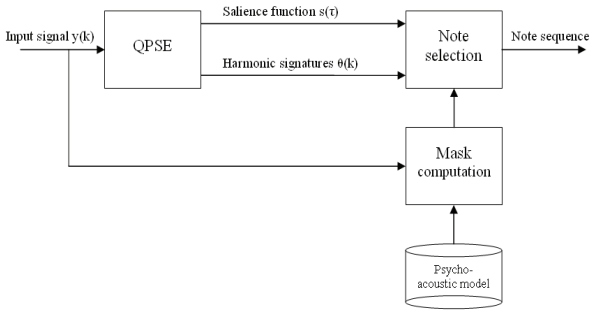

Fig. 1. The block diagram of the transcription scheme

2. QUASI-PERIODIC SIGNAL EXTRACTION

Due to space limitation, this section provides only a brief overview of the quasi-periodic signal model and the related extraction algorithm. The interested reader is referred to [9, Ch.2] and [10] for an exhaustive description and better coverage.
To understand the proposed model, let us first consider the sinusoidal model. This model represents the signal as a sum of discrete time-varying sinusoids or partials:

$$s(n) = \sum_{p=0}^{P} a_p(n) \cos\left(2\pi p n f_0 + 2\pi\varphi_p(n)\right) \quad , \tag{2}$$

where $\varphi_p(n)$ characterizes the evolution of the instantaneous phases around the p^{th} harmonic, and can be assumed to be slowly time varying. We assume that all harmonic amplitudes evolve proportionately in time, and that the instantaneous frequency of each harmonic is proportional to the harmonic index, i.e.,

$$\left\{ \begin{array}{l} a_p(n) = a_p\, a(n) \\ 2\pi\varphi_p(n) = 2\pi p\, \varphi(n) \;+\; \Phi_p. \end{array} \right. \quad . \tag{3}$$

Under these assumptions, one can show that the audio signal is modeled as the superposition of harmonic components with a global amplitude modulation and global time-warping:

$$\begin{aligned} s(n) &= a(n) \sum_p a_p \cos\left(2\pi p f_0 \left(n + \frac{\varphi(n)}{f_0}\right) + \Phi_p\right) \\ &= a(n)\; \theta\left(n + \frac{\varphi(n)}{f_0}\right) \end{aligned} \tag{4}$$

where $a(n)$, and $\varphi(n)$ represents respectively the amplitude and frequency modulating signals. $\theta(n) = \sum_p a_p \cos\left(2\pi p f_0 n + \Phi_p\right)$ is a periodic signal with a period $T = \frac{1}{f_0}$.
A major limitation of the proposed model is that it allows for no spectral variation throughout the note duration, but only amplitude and (synchronized) frequency modulation. Such a model assumes that at any time instant the instantaneous amplitudes and frequencies of the various harmonics of the periodic waveform are proportional. The problem with such a model though is that, in reality, periodic signals produced by musical instruments (e.g. string instruments) have harmonic components that decay at different speeds. Typically, higher harmonics decay faster than lower harmonics. In [10], we have introduced a frequency-selective attenuation to alleviate this side-effect, and this in a time-varying fashion to reflect the time-varying amplitude, i.e.,

$$s(n) = a_n(q)\; \theta\left(n + \frac{\varphi(n)}{f_0}\right) \tag{5}$$

where $a_n(q) = a_{n,L}\, q^L + \cdots + a_{n,0} + \cdots + a_{n,L}\, q^{-L}$ is a symmetric zero-phase FIR filter, $2L+1$ is the amplitude modulating filter length, and q^{-1} is the time delay operator. The two extreme filter lengths correspond to the flat modulation model as in (4) (for $L = 0$), and the bayesian harmonic model in [15](for $L = \infty$).

The assumptions of global amplitude and frequency modulation were introduced to have a parsimonious signal representation. Indeed, the higher the number of parameters per second describing the signal, the noisier the parameter estimates, and consequently the reconstructed signal. Introducing an amplitude modulating signal per harmonic (as in [15, 17]) would allow significant degrees of freedom in describing the signal, but would lead to a high parameter rate (the average number of parameters that appear in the description of one second of the signal). An intermediate parameter rate can be obtained by filtering the periodic signal with the short FIR filter $a_n(q)$ that can introduce frequency-selective attenuation, and this in a time-varying fashion to reflect the time-varying amplitude.

Audio signal extraction is performed by adjusting the degrees of freedom (in $a_n(q)$, $\varphi(n)$, and $\theta(n)$) such that the assumed model best matches the received signal (in the least-squares sense). The degrees of freedom are estimated in a cyclic fashion [9]. Applying the proposed technique to music signal analysis seems natural. Indeed, the proposed model is related to the physics of how sounds are produced in stringed and wind instruments [9]. And the model parameters are tightly related to the three basic features in music sounds: pitch ($\varphi(n)$), dynamics ($a_n(q)$), and timbre ($\theta(n)$). Simulations show that the proposed scheme is suitable for the analysis of several string and wind instruments, and performs good monophonic transcription accuracy [10].

Music transcription needs both good temporal and frequency resolutions (to both *detect the 'position'* and *recognize the 'content'* of a musical event). Compared to the classic frame-by-frame based approaches, the quasi-periodic signal modeling performs better resolution tradeoff (by exploiting the temporal structure of the musical signal). Indeed, the global amplitude modulation model enables the joint extraction of the different partials, while allowing for slow L decay modes. This fact enhances both note detection and recognition accuracy (intuitively, the QPSE tries to estimate simultaneously the spectral structure in both time and frequency directions). In addition, valuable information could be carried out by analyzing individually the different parameters:

- $a_n(q)$: high temporal resolution transcription.

- $\varphi(n)$: detection of several musical effect, e.g., vibrato, glissando, etc...
- $\theta(n)$: accurate musical note selection (as detailed in the next section).

3. MUSICAL NOTE SELECTION

The first building (front-end) block in music transcription and indexing operations is the decomposition of music signals into harmonically related components. The QPSE tries to simultaneously estimate the spectral structure in both time and frequency directions. This fact leads to a better time/frequency resolution tradeoff, and partially alleviates common partials extraction. Contrary to the frequency domain approaches, no explicit constraint on the number of harmonics is done (implicitly, it is constrained by the ratio between the fundamental and sampling frequencies).
Based on the previous decomposition, the salience function can be evaluated at the different period candidates as

$$Sl(\tau) = \frac{\sum_n \hat{s}_\tau^2(n)}{\sum_n y^2(n)} \tag{6}$$

where $\hat{s}_\tau(n) = \hat{a}_n(q)\ \hat{\theta}(n + \frac{\hat{\varphi}(n)}{f_0})$ is the extracted quasi-periodic signal assuming a basic period τ; and $y(n)$ denotes the input audio signal. Note that the proposed function satisfies $0 \leq Sl(\tau) \leq 1$, and it does not contain a myriad of parameters ($g(\tau, p)$ in (1)) that should be learned [1] or set [6]. Moreover, the QPSE enables joint extraction of the different partials while imposing a kind of spectral smoothness (over the time axis) that has been shown to be valuable to increase transcription accuracy [1, 7].

For monophonic music transcription, the note selection can be performed simply by picking the lowest maximum of the salience function [10]. The choice of the lowest period solves the octave indeterminacy. In a polyphonic context, the QPSE (although it helps) is not able to solve alone the common partials and octave indeterminacy problems. In this paper, we propose first to select the periods corresponding to the salience function local maxima (as potential candidates), then to use the distortion detectability distance (introduced in [13]) to discard the 'ghost notes' and perform accurate note selection. The distortion detectability defines a perceptually relevant norm on the harmonic signal subspace spanned by:

$$\bar{s}_\tau(n) = \|\hat{s}_\tau\|\, \theta\left(n \,\%\, \tau\right), \tag{7}$$

where $(.\,\%\,.)$ denotes the modulo operator, and $\bar{s}_\tau(n)$ is the τ-periodic signal sharing the same energy (salience) and basic-waveform with $\hat{s}_\tau(n)$, but compensated for global amplitude and phase modulations.
The auditory model was designed to predict the masked thresholds for *sinusoidal distortions*. The model accounts for the spectral and temporal integration in auditory masking. As shown in [13], the distortion detectability distance can be expressed as

$$D(y, \bar{s}_\tau) = \sum_f \frac{|\overline{S}_\tau(f)|^2}{\nu_y^2(f)} \tag{8}$$

where $|\overline{S}_\tau(f)|^2$ denotes the power spectrum of the signal $\bar{s}_\tau(n)$, and $\nu_y^2(f)$ represents the frequency dependent masking curve which is computed using the input audio signal $y(n)$. The distance is calibrated such that $D = 1$ represents the threshold of detectability, i.e.,

$$\begin{cases} D(y, \bar{s}_\tau) > 1 & \text{the } \tau^{th} \text{ candidate is detectable} \\ D(y, \bar{s}_\tau) < 1 & \text{the } \tau^{th} \text{ candidate is a ghost note} \end{cases} \tag{9}$$

This calibration leads to a simple decision scheme, with no need to a priori adjusted thresholds (that may depend on unknown parameter such as SNR, instrument class, number of notes, etc). According to (8), even if the individual tonal components of a given signal are masked, their combination may still be detectable, which matches the auditory spectral integration results. In addition to its auditory relevance, the computational load of the proposed distance remains reasonable since the masking threshold only needs to be computed once (per frame). Moreover, as the auditory bands are narrow at low frequencies and wide at high frequencies, the masking model discards further high-order harmonics (simply because many components will fall in one auditory filter band). This leads to improvements in performance of the algorithm, since higher-order harmonics are generally less reliable due to their low energy content (low SNR) and the effect of inharmonicity (in stringed instruments). To enhance this fact, it was found that adding artificial white noise (SNR=10 dB) to the input signal while computing the masking curve, or simply raising the masking curve by a given factor was benifical.

4. EXPERIMENTAL RESULTS

Using the previously described building blocks, a simple transcription scheme was built. First, the QPSE was performed for the different possible notes periods at the 1^{st}, 2^{nd} and 3^{rd} octaves; and the local maxima of the salience function (as defined in (6)) were picked. Then, the note selection was performed based on the detectability distance. A simplified auditory masking model (cf. [14]) was used. This model is shown to be effective for musical key extraction [14]. The implementation of the masking model was graciously provided by Steven van de Par from Philips Research Laboratories. Specifically, this measure is used to:

- solve the octave indeterminacy.
- discard ghost notes.
- recover the notes that do not belong the search set.

The system does not make any assumption on the number of sounds in the mixture, and no-estimation of the number of concurrent sounds is required. No information about the instrument class or timber was assumed.

The scheme has been tested using the piano recordings of the RWC (Real World Computing) musical instrument sound database [16]. For each recording, the database includes a reference MIDI file which contains a manual annotation of the note events in the acoustic recordings. The test data consists of random mixture of individual note recordings. The recordings (initially recorded at 44.100 kHz) are downsampled to 22.050 kHz. The maximum number of iterations (in the QPSE cyclic parameters estimation) was fixed to 3. The order of the amplitude modulating filter was set to $L = 4$. No global phase modulation was considered.

A standard error metric was used for evaluation [11]: a recall measure (percentage of original notes that were transcribed), and a precision measure (percentage of transcribed notes that were present

on the original stream). The rates are defined as:

$$\text{recall} = \frac{\#\text{correct notes}}{\#\text{reference notes}}$$

$$\text{precision} = \frac{\#\text{correct notes}}{\#\text{transcribed notes}}$$

where # denotes the cardinality operator.
We have compared the proposed selection scheme to a decision scheme based on the sum of either the harmonic power or amplitudes. In each case, the detectability thresholds were manually tuned in a way to achieve good recall vs. precision tradeoff. But no automatic learning was performed. Figure 2 illustates the transcription results function of the number of concurrent sounds (#polyphony). The polyphony number is assumed to be unknown.

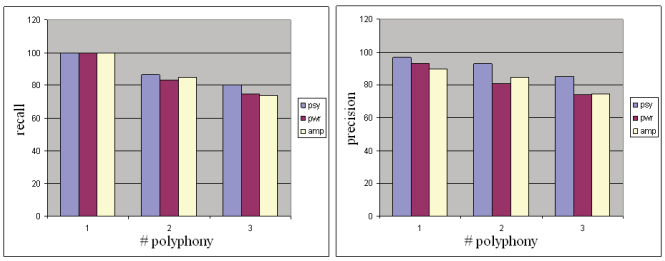

Fig. 2. The recall and precision measures as a function of the number of concurrent sounds (#polyphony) using (respectively from right to left) the psycho-acoustic (psy), the power spectrum (pwr) and the amplitude spectrum (amp) based measures.

The graph shows that the three schemes achieve good transcription accuracy. We remark also that the psycho-acoustic distance outperforms the two other selection schemes. This decision scheme has the additional advantage to be considerably simpler to set and control.
We have also tested the scheme with various musical instruments (guitar, sitar and flute). The data was graciously provided by Antony Schutz from Eurecom. Although the database was not large enough to obtain consistent statistical results, the proposed scheme leads to comparable performance (or even better). The result was expected since these instruments produce less inharmonicity effects (compared to the piano). We have also experienced the robustness of the proposed scheme in the presence of noise, percussive sounds and in unbalanced Signal-to-Interference Ratio (SIR) situations.

5. CONCLUDING REMARKS

This paper describes a method for transcribing realistic polyphonic audio. The analysis accounts for both physical and psycho-acoustical issues. Indeed based on a physically-motivated audio model, the QPSE algorithm estimates the spectral structure of the musical note in both the time and frequency directions; leading to a better time/frequency resolution tradeoff. Based on the extraction SNR, an initial set of note candidates is selected. A perceptually motivated distance is then used to discard the ghost candidates. No prior information on the musical instrument, musical genre, or/and maximum polyphony are needed. Simulations show that the proposed scheme achieves good transcription results for a variety of string and wind instruments. The proposed technique is also shown to be robust in the presence of noise, percussive sounds and in unbalanced Signal-to-Interference Ratio (SIR) situations.

6. REFERENCES

[1] A. Klapuri, "Multiple Fundamental Frequency Estimation by Summing Harmonic Amplitudes," *In Proc. of Int. Symp. on Music Information Retrieval* , Oct. 2006.

[2] A. de Cheveigné, "Separation of Concurrent Harmonic Sounds: Fundamental Frequency Estimation and a Time-Domain Cancellation Model of Auditory Processing," *Journal of the Acoustical Society of America*, Vol.93, 1993.

[3] A. Klapuri, "A Perceptually Motivated Multiple-F0 Estimation Method," *In Proc. of IEEE work. on Applications of Signal Processing to Audio and Acoustics* , Oct. 2005.

[4] R. Meddis and M.J. Hewitt , "Virtual Pitch and Phase Sensitivity of a Computer Model of the Auditory Periphery. I: Pitch identification," *Journal of the Acoustical Society of America*, Jun. 1991.

[5] K. Miyamoto, H. Kameoka, H. Takeda, T. Nishimoto and S. Sagayama, "Probabilistic Approach to Automatic Music Transcription from Audio Signals," *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Apr. 2007.

[6] I. Bruno, S.L. Monni and P. Nesi, "Automatic Music Transcription Supporting Different Instruments," *In Proc. of IEEE Int. Conf. on Web Delivering of Music*, Sept. 2003.

[7] A. Pertusa and J.M. Inesta, "Multiple Fundamental Frequency Estimation Using Gaussian Smoothness," *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Apr. 2008.

[8] M. Triki and D.T.M. Slock, "Music Source Separation via Sparsified Dictionaries vs. Parametric Models," *In Proc. of Int. Sym. on Communications, Control, and Signal Processing*, Mar. 2006.

[9] M. Triki, "Some Contributions to Statistical Signal Processing and Applications to Audio Enhancement and Mobile Localization," *Ph.D. Thesis report*, TelecomParis / Eurecom Institute, Jun. 2007.

[10] M. Triki and D.T.M. Slock, "Periodic Signal Extraction with Frequency-Selective Amplitude Modulation and Global Time-Warping for Music Signal Decomposition," *In Proc. of IEEE Int. Work. on Multimedia Signal Processing*, Oct. 2008.

[11] G. Reis, N. Fonseca and F. Ferndandez, "Genetic Algorithm Approach to Polyphonic Music Transcription," *In Proc. of IEEE Int. Symp. on Intelligent Signal Processing*, Oct. 2007.

[12] M.P. Ryynanen and A. Klapuri, "Polyphonic Music Transcription Using Note Event Modeling," *In Proc. of IEEE work. on Applications of Signal Processing to Audio and Acoustics*, Oct. 2005.

[13] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen and S.H. Jensen, "A Perceptual Model for Sinusoidal Audio Coding Based on Spectral Integration," *EURASIP Journal on Applied Signal Processing*, Issue 9, 2005.

[14] S. van de Par, M.F. McKinney and A. Redert, "Musical Key Extraction from Audio Using Profile Training," *In Proc. of Int. Symp. on Music Information Retrieval*, Oct. 2006.

[15] S.Godsill and M. Davy, "Bayesian Harmonic Models for Musical Pitch Estimation and Analysis," *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, May 2002.

[16] M. Goto, H. Hashiguchi, T. Nishimura and R. Oka, "RWC Music Database: Music Genre Database and Musical Instrument Sound Database," *In Proc. of Int. Symp. on Music Information Retrieval*, Oct. 2003.

[17] E. Vincent and M.D. Plumbley, "Low Bitrate Object Coding of Musical Audio Using Bayesian Harmonic Models," *IEEE Trans. on Acoustics, Speech and Language Processing*, Vol.15, pp.1273-1282, Apr. 2007.