PERIODIC SIGNAL EXTRACTION
WITH GLOBAL AMPLITUDE AND PHASE MODULATION
FOR MUSIC SIGNAL DECOMPOSITION

Mahdi Triki, Dirk T.M. Slock[*]

Eurecom Institute
2229 route des Crêtes, B.P. 193, 06904 Sophia Antipolis Cedex, FRANCE
Email: {triki,slock}@eurecom.fr

ABSTRACT

A key building block in music transcription and indexing operations is the decomposition of the music signal into notes. We model a note signal as a periodic signal with (slow) global variation of amplitude (reflecting attack, sustain, decay) and frequency (limited time warping). The bandlimited variation of global amplitude and frequency gets expressed through a subsampled representation and parameterization of the corresponding signals. Assuming additive white Gaussian noise, a Maximum Likelihood approach is proposed for the estimation of the model parameters and the optimization is performed in an iterative (cyclic) fashion that leads to a sequence of simple least-squares problems. Particular attention is paid to the estimation of the basic periodic signal, which can have a non-integer period, and the estimation of the amplitude signal with guaranteed positivity.

1. INTRODUCTION

Sinusoidal model based music analysis/synthesis has received considerable interest in the computer music community [4, 5, 6]. The sinusoidal transform, originally developed by Quatieri and McAulay [3], represents a signal as a sum of discrete time-varying sinusoids or partials:

$$s(t) = \sum_{k=0}^{P} A_k(t) \cos\left(\theta_k(t)\right) \quad . \tag{1}$$

The estimation of the model parameters is typically carried out using a short-time Fourier transform (STFT) with a fixed analysis frame size and a fixed stride between frames. The sinusoids are extracted by peak-picking in the STFT magnitude spectrum. Intermediate values are obtained by interpolation. A fundamental problem faced by the traditional sinusoidal-model based techniques, and which arises due to the STFT, is smearing of the frequency response [8, 7]. In fact, over the period of a single analysis frame, the algorithm estimates the amplitude, frequency and phase of any sinusoids it believes to be present. Because of the near logarithmic scale of pitch perception, we need very long windows in order to accurately estimate the pitch of low frequency partials. On the other hand, the time resolution of these parameters is only as fine as the window length, itself. And, since the music signal is strongly non-stationary , it is not always possible to find a good tradeoff between time and frequency resolution. Also, determining the sinusoid parameters from the STFT peak amplitude and phase only works well for high frequency resolution, high SNR and in the absence of modulation.

Another drawback of these techniques is that they ignore the harmonic structure of the music signal. In fact, they consider the signal as a mixture of a finite number of arbitrary sinusoids, and not as a periodic signal. For treating periodic signals, the state of the art is limited to the estimation of pure periodic signals with period equal to an integer number of samples [1, 2]. In these references, the authors propose a Maximum Likelihood approach to analyze pure periodic signals. They show that the resulting procedure can be interpreted as a signal projection onto suitable subspaces.

This paper extends the results of those references, and tries to merge the modulated sinusoidal modeling and the periodic signal analysis techniques, by considering periodic signals with non-integer period and global amplitude variation and time warping. The use of this model gives a compromise between reality and a parsimonious parameterization. Indeed, global amplitude variation reflects mostly attack, sustain, and decay of the whole note signal. Whereas, the global time warping allow the capture of vibrato and sliding notes. With an eye on future extensions to polyphonic sounds, the method should be able to work in fairly low SNR. Hence it is important to have parsimonious parameterizations in order to limit the estimation noise. The motivation for the proposed model is to provide a good compromise between approximation noise and estimation noise.
In music, the nominal frequency of a note is known. So we assume an analysis exploring the hypothesis of the presence of a note at any possible nominal note frequency. However, we do not treat the harmonics of a note signal separately as a simple filter bank approach would do (this is basically the state of the art in music signal analysis). Rather, the energy in all harmonics is exploited jointly through the treatment of the complete periodic signal, in order to robustify the detection of the note signal and the estimation of its modulation characteristics. The Global Modulation (GM) assumption helps the separation of note signals that have harmonics in common.

This paper is organized as follows. In section (2), the global modulation model is presented. The extraction procedure will then be derived in section (3). Performance of the algorithm is evalu-

[*] Eurécom's research is partially supported by its industrial partners: Hasler Foundation, Swisscom, Thales Communications, ST Microelectronics, CEGETEL, France Télécom, Bouygues Telecom, Hitachi Europe Ltd. and Texas Instruments. The work reported herein was also partially supported by the SIEPIA project of the French RIAM network (Recherche et Innovation en Audiovisuel et Multimédia).

ated in Section (4), and finally a discussion and concluding remarks are provided in section (5).

2. SIGNAL MODEL

In the sinusoidal modeling, the signal is modeled as a sum of evolving sinusoids as in (1), where $\theta_k(t)$ represents the instantaneous phase of the k^{th} partial. As the music signal is quasi-periodic, $\theta_k(t)$ can be de composed into

$$\theta_k(t) = 2\pi k t f_0 \;+\; 2\pi\varphi_k(t) \tag{2}$$

where $\varphi_k(t)$ characterizes the evolution of the instantaneous phases around the k^{th} harmonic; and can be assumed to be low-frequency. The Global Modulation assumption implies that all harmonic amplitudes evolve proportionally in time; and that the instantaneous frequency of each harmonic is proportional to the harmonic index:

$$\left\{ \begin{array}{l} A_k(n) = A_k\; A(n) \\ 2\pi\varphi_k(n) = 2\pi k\; \varphi(n) \;+\; \Phi_k \end{array} \right. \quad . \tag{3}$$

In summary, we model an audio signal as the superposition of harmonic components with a global amplitude modulation and time warping (that can be interpreted in terms of phase variations):

$$\begin{array}{l} y(n) \;=\; s(n) \;+\; v(n) \\ = \;\sum_k A_k(n)\; \cos\left(2\pi k n f_0 + 2\pi\varphi_k(n)\right) + v(n) \\ = \;A(n)\sum_k A_k \cos\left(2\pi k f_0 \left(n + \frac{\varphi(n)}{f_0}\right) + \Phi_k\right) + v(n) \end{array}$$

where - v_n is an additive white Gaussian noise.
- $A(n)$ represents the amplitude modulating signal
- $\varphi(n)$ denotes the phase modulating signal (that can interpreted in terms of time warping).

2.1. Global Phase/Frequency Modulation

Consider, first, the case of a pure periodic signal:

$$s(n) = \sum_k A_k \cos\left(2\pi k n f_0 + \Phi_k\right) \quad . \tag{4}$$

If the fundamental period $T = \frac{1}{f_0}$ is an integer, then $\theta = [\theta(1)\cdots\theta(T)]^T$, the signal over one period, is sufficient to describe the totality of the periodic signal $S = [s(1)\cdots s(N)]^T$ [1, 2]:

$$S = \begin{bmatrix} I_T \\ I_T \\ \vdots \end{bmatrix} \theta = F\theta$$

where the column space of F corresponds to the signal subspace for a periodic signal of period T. When T is not integer, we shall take the vector θ of size $\lceil T \rceil$ (and not longer, to minimize identifiability problems). Hence θ contains a set of sufficient statistics to describe the whole periodic signal. To generate the periodic signal vector S exactly, the interpolation matrix F would become quite complex. If we want to obtain approximate interpolation and work with FIR interpolation filters, we could extend θ in length by the length of the FIR filter to be used and introduce an F of similar structure as in (5), but with the identity matrices replaced by banded blocks corresponding to the (time-varying) FIR filter. For simplicity, we shall assume that a certain degree of oversampling has been performed so that simple linear interpolation (corresponding to a triangular interpolation filter) produces good results. In that case, the matrix F is given by

$$F \;=\; \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \beta & 1\!-\!\beta & \cdots & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & & \beta\lfloor T\rfloor & 1\!-\!\beta\lfloor T\rfloor \\ 1\!-\!\beta\lceil T\rceil & & \cdots & \beta\lceil T\rceil \\ & \vdots & & \end{bmatrix} \tag{5}$$

where $\beta = 1 - \frac{T}{\lceil T \rceil}$. This is a banded matrix with in every row only two consecutive (in a modulo sense) elements being non-zero, providing a convex combination of two available samples to approximate an intermediately positioned sample.

The same approach can be used to take into account a given time warping by considering $f(t) = f_0 + \psi(t)$, being a piecewise constant function of time (see figure 1). As a result, the phase becomes a piecewise linear function of time. The time period T_1 over which the instantaneous frequency is supposed to be constant is chosen such that $\frac{1}{T_1}$ exceeds (well) the (assumed) bandwidth of variation of the instantaneous frequency. As a result, the frequency and hence phase variation gets parameterized by the subsample values at rate $\frac{1}{T_1}$. The way the figure 1 should be interpreted is that the line indicates for each row of the matrix the point for which an interpolation value has to be provided. In the case of simple linear interpolation the two matrix elements on that row surrounding the intersection of the line with the row will correspond to an appropriate convex combination.

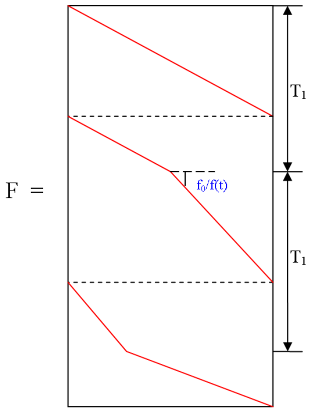

Fig. 1. Interpolation matrix structure

2.2. Global Amplitude Modulation

While time warping focuses on the time evolution of the instantaneous frequency, and allows the modeling of several musical phenomena (vibrato, glissando ...), the global amplitude-modulating signal allows an evolution of the note power, reflecting attack, sustain, and decay. The amplitude signal is assumed to be a non-negative low-pass signal. Hence it can also be represented as an interpolated version of a subsampled signal (see further). So the audio signal can be written as:

$$Y = \underbrace{A\,F\theta}_{=S} + V \tag{6}$$

where :
- $Y = [y(1)\cdots y(N)]^T$, represents the observation vector
- $S = [s(1)\cdots s(N)]^T$, represents the signal of interest
- $V = [v(1)\cdots v(N)]^T$, denotes the noise vector
- $\theta = [\theta(1)\cdots\theta(\lceil T\rceil)]$, characterizes the harmonic signature over essentially one period
- $A = diag[A(1)\cdots A(N)]$, represents the global amplitude modulation signal
- F is an $N \times \lceil T\rceil$ interpolation matrix characterizing the time warping.

3. PERIODIC SIGNAL EXTRACTION PROCEDURE

The previous model is linear in θ, A, or F (separately), F being parameterized nonlinearly. Trying to estimate all factors jointly is a difficult nonlinear problem. Indeed, as the noise is assumed to be a white Gaussian signal, the ML approach leads to the following least-squares problem:

$$\min_{A,F,\theta} \|Y - A\,F\,\theta\|^2 \tag{7}$$

where A and F are parameterized in terms of subsamples. The estimation can easily be performed iteratively though.

3.1. Periodic Signature Estimation

If we assume that the matrices $\widehat{A}$, $\widehat{F}$ are given, the periodic signature θ can be isolated as

$$Y \; = \widehat{A}\,\widehat{F}\,\theta + V = \widehat{H}\,\theta + V \tag{8}$$

Then minimizing (7) w.r.t. θ leads to

$$\widehat{\theta} = \left(\widehat{H}^T\widehat{H}\right)^{-1}\widehat{H}^T Y \quad . \tag{9}$$

Hence the periodic signature gets estimated by using the data over the whole note duration.

3.2. Instantaneous Amplitude Estimation

The amplitude signal could similarly be estimated from (7) by isolating $\widehat{F}$ and $\widehat{\theta}$. However, such an estimation procedure would not guarantee positive values for the estimated amplitude signal. Alternatively, consider performing the square of the note signal:

$$\begin{aligned} s^2(n) \;&= A^2(n)\;\left(\textstyle\sum_k A_k\cos(2\pi k n f_0 + 2\pi k\varphi(n) + \Phi_k)\right)^2 \\ &= A^2(n)\;\overline{\theta^2}\; + \;(\text{high freq. terms}) \end{aligned} \tag{10}$$

where $\overline{\theta^2} = \dfrac{1}{2}\sum_k A_k^2 = \dfrac{1}{\lceil T\rceil}\sum_{k=1}^{\lceil T\rceil}\widehat{\theta}^2(k) = \dfrac{1}{\lceil T\rceil}\|\widehat{\theta}\|^2$ denotes the power of the periodic signal. Taking into account an additive noise leads to:

$$y^2(n) - v^2(n) = \underbrace{A^2(n)\;\overline{\theta^2}}_{\text{signal}} + \underbrace{2s(n)v(n) + (\text{high freq. t.})}_{\text{noise}}$$

from which we shall estimate $A^2(n)$ via least-squares. We propose to express the low-pass character of $A(n)$ by taking $A(n)$ to be piecewise constant (so that $A^2(n)$ is also piecewise constant) over time frames that are a multiple of T (so that the high frequency terms get suppressed well). The length of these time frames (which can differ from T_1, which is typically longer since the frequency varies more slowly than the amplitude) can be time-varying, to accomodate for the time-varying speed of variation of the amplitude (attack versus decay). Thus, $A(n)$ gets estimated using:

$$\widehat{A}(n) = \sqrt{\frac{1}{\overline{\theta^2}}\left\langle y^2(n) - (y(n) - \widehat{s}(n))^2\right\rangle_n} \tag{11}$$

where $\langle\,.\,\rangle_n$ denotes temporal averaging over the piecewise interval containing n; $\widehat{S} = \widehat{A}\widehat{F}\widehat{\theta}$ denotes the latest estimate of the signal of interest.

3.3. Instantaneous Frequency Estimation

As for the instantaneous amplitude, the instantaneous frequency gets estimated on a frame-by-frame basis. In each frame, the instantaneous frequency is optimized using (7):

$$\begin{cases} \min\limits_{f} \left\|Y - \widehat{A}\widehat{F}(f)\widehat{\theta}\right\| \\ \frac{\Delta f}{f_0} \leq \alpha_{max} \end{cases} \tag{12}$$

where Δf denotes the maximum relative frequency variation in the current frame compared to the previous frame, reflecting an assumed limited frequency variation rate. The optimal instantaneous frequency value for the current frame gets determined from a finite set of discrete values within the thus limited range.

4. EXPERIMENTAL RESULTS

Using the proposed approach, we have experimented with a real music signal. The proposed signal represents a single note (pitch = 84 Hz) played by an acoustic guitar. The record has a duration of 1s and is sampled at 22.050 Khz (see figure 2). The SNR of the input signal is 26 dB.

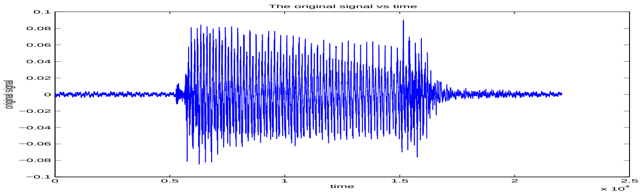

Fig. 2. Original guitar signal.

In order to improve the convergence, the algorithm gets initialized by starting with the periodic signature estimation over a limited portion of the signal, somewhere in the middle, over which no amplitude or frequency modulation is assumed (initially). Then the indicated iterative scheme takes over with amplitude estimation over the whole signal duration, frequency estimation over the whole signal, periodic signature reestimation over the whole signal etc. Iterations are performed untill the relative change in the periodic signal signature gets below 10^{-3}. This corresponds to about ten iterations.

In figure 3, we plot the different outputs of the algorithm: original signal (Y), synthesized signal ($\widehat{S}$) according to the global amplitude and frequency modulation model, signal error ($\widehat{V}$) (difference between the previous two), instantaneous amplitude signal $\widehat{A}(n)$ and instantaneous frequency signal $\widehat{f}(n)/f_0$ (relative to the nominal fundamental frequency).

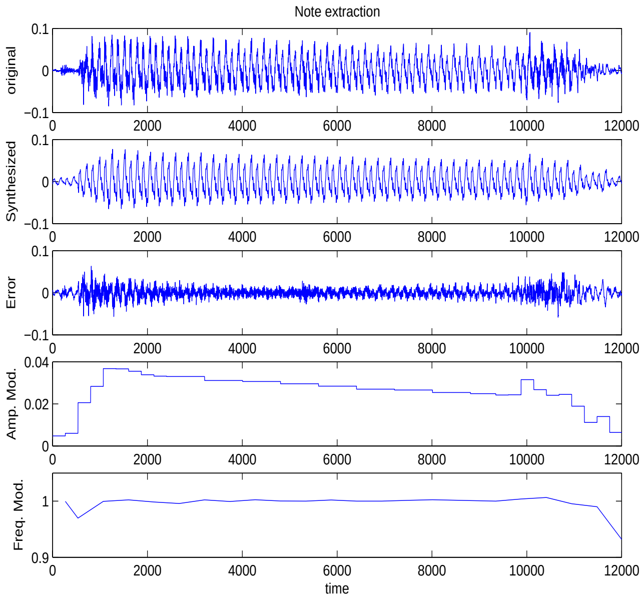

Fig. 3. Amplitude and Frequency modulation extraction.

In order to analyze the extraction quality of our algorithm, we plotted the Fast Fourier transform of the original signal (Y) and the residual error signal ($\widehat{V}$) (figure 4).

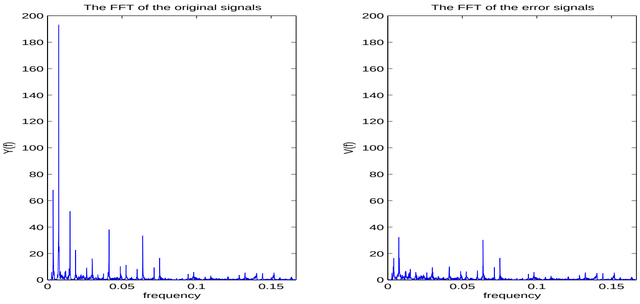

Fig. 4. The FFT of the original and residual error signals.

We notice that the extraction quality depends mainly on the harmonic index. First of all, one may notice that the evolution with harmonic index of the strength A_k of the various harmonics is not at all smooth. Next, modeling degradation appears to start at harmonic 17. Due to the global frequency modulation assumption, modeling error indeed increases with the harmonic index for the instantaneous phase. Another possible explanation might be a deviation from the global amplitude modulation assumption. Indeed, it is often indicated that the amplitude signals of the different harmonics are not proportional for a number of musical instruments. Nevertheless, the global amplitude modulation assumption appears to hold for quite an impressive number of harmonics in this example. We also calculate the signal to (measurement plus approximation) noise ratio for the estimated model:

$$SNR = \frac{\sigma_y^2}{\sigma_{\hat{v}}^2} = 7.9\ dB$$

which is quite far from the SNR of 26 dB. The exact sources for the residual error are currently being investigated.

5. CONCLUSION

In this paper, we have started investigating the decomposition of a music signal into note signals. We have considered the periodic signal model with a slow global amplitude and phase variation. Assuming additive white Gaussian noise, and small time warping variation, a Maximum Likelihood approach is proposed for the estimation of the model parameters and the optimization is performed in an iterative (cyclic) fashion that leads to a sequence of simple least-squares problems. Simulations show that the extraction technique is suitable for the analysis of musical notes, and produces good auditive synthetic results. Possible further sophistications of the model currently being envisaged (depending on the error explanations to be found) are a splitting of the global modulation assumption into subbands and/or allowing a slow variation of the periodic signature waveform over the signal duration.

6. REFERENCES

[1] D.D. Muresan, and T.W. Parks. “Orthogonal, Exactly Periodic Supspace Decomposition,” *IEEE Transactions on Signal Processing*, Vol. 51, No. 9, September 2003.

[2] J.D. Wise, J.R. Caprio and T.W. Parks. “Maximum Likelihood Pitch Estimation,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 51, Pages: 418-421, May 1976.

[3] R. McAulay, T. Quatieri. “Speech Analysis/Synthesis Based on a Sinusoidal Representation,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 34, Pages: 744-754, August 1986.

[4] M. Goodwin, M. Vetterli. “Time-Trequency Signal Models for Music Analysis, Transformation, and Synthesis,” *In Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, Pages: 133-136, June 1996.

[5] M. Goodwin, M. Vetterli. “Atomic Decompositions of Audio Signals,” *In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 1997.

[6] Yinong Ding; Xiaoshu Qian;. “Estimating Sinusoidal Parameters of Musical Tones Based on Global Waveform Fitting,” *In Proceedings of the IEEE Workshop on Multimedia Signal Processing*, Pages:95 - 100, June 1997.

[7] P. Prandom, M. Goodwin, M. Vetterli. “Optimal Time Segmentation for Signal Modeling and Compression,” *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol.3, Pages: 2029-2032, April 1997.

[8] S.N. Levine, T.S. Verma, J.O. Smith. “Multiresolution Sinusoidal Modeling for Wideband Audio with Modifications,” *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol.6, Pages: 3585-3588, May 1998.