HARMONIZE-DECOMPOSE AUDIO SIGNALS WITH GLOBAL AMPLITUDE AND FREQUENCY MODULATIONS

Mahdi Triki

Philips Research Laboratories, Eindhoven, The Netherlands
mahdi.triki@philips.com

ABSTRACT

A key building block in music transcription and indexing operations is the decomposition of music signals into notes. We model a note signal as a periodic signal with slow (frequency-selective) amplitude modulation and global frequency-warping. Global frequency-warping allows for an inharmonic frequency modulation, while the global amplitude modulation allows the various harmonics of the periodic signal to decay at different speeds. The global frequency-warping is achieved by a Laguerre transform (that has shown to fit stiffed strings inharmonic behavior). Assuming additive noise, the estimation of the model parameters and the optimization is performed in a Harmonize-Extract fashion. Simulations illustrate that the extraction technique oversteps the limitation of the global AM-FM representation and analysis techniques and allows the processing of inharmonic string instruments (e.g. piano).

1. INTRODUCTION

Motivated by the relationship between AM and FM modulation and the processes of sound production and perception, spectral-based techniques for the analysis, transformation and synthesis of audio signals have received considerable interest in the computer music community. The interested reader is referred to [1, 2, 3] for more comprehensive reviews and better coverage.

One of the most successful and ubiquitous is the family of parametric representations based on the sinusoidal modeling paradigm. The sinusoidal transform, originally developed by Quatieri and McAulay [4], represents a signal as a sum of P discrete time-varying sinusoids or partials:

$$s(n) = \sum_{k=0}^{P} a_k(n) \cos\left(\psi_k(n)\right) \quad . \tag{1}$$

A variety of approaches has been proposed in the literature to perform AM-FM signal decomposition. Among them, the family of time-frequency representations has been relatively successful due to their implementation simplicity and their capability of handling noise to some extent. These generally employ frame-based non-parametric spectral analysis techniques to detect peaks corresponding to sinusoidal-like components. Various techniques have been proposed for accurate peak localization based on non-linear interpolation (e.g. [5]), dichotomy (e.g. [6]) and/or high-resolution analysis (e.g. [7, 8]). Subsequently, these peaks are linked across consecutive time frames [9, 10] and/or coherent frequency bands [11, 12]. A second class of approaches address AM-FM decomposition using multiband filtering and demodulation [13, 14]. The basic idea is to first locate local spectral formants. Next, the instantaneous AM and FM signals are individually tracked for the different formants. In [15], we have introduced an alternative approach for AM-FM audio signal decomposition. Instead of addressing individual frames and/or formants, the harmonic structure and temporal consistency are both exploited to identify modulations that are common to all partials of a given sound. We have considered a periodic model with non-integer period and global AM and FM modulation (i.e., global amplitude variation and time-warping). The proposed scheme does not treat the harmonics of an audio signal separately as a simple filter bank approach would. Rather, the energy in all harmonics is exploited jointly through the treatment of the complete periodic signal, in order to robustify the estimation of its modulation characteristics. The Global Modulation (GM) assumptions help the separation of audio signals that have harmonics in common. Furthermore, valuable information could be obtained by individually analyzing the model parameters. Indeed, global amplitude variation reflects mostly attack, sustain, and decay of the whole note signal, whereas global time-warping allows for the detection of musical effects (e.g. vibrato, glissando, etc). In [16] the GM representation was further developed by introducing a frequency-selective global amplitude modulation. The amplitude variations of the various harmonics are modeled using a short FIR filter that introduces a frequency-selective attenuation (allowing for different attack/decay modes), and this in a time-varying fashion to reflect the time-varying amplitude. Simulations show that the proposed scheme is suitable for the analysis of several string and wind instruments [16], and shows good potential for music transcription applications [17].

The GM-based models allow a parsimonious representation of the instantaneous amplitude of the different partials, which leads to a good estimation vs. modeling noise tradeoff and an effective signal decomposition. The proposed representations, however, allow only for (slow) variations of the fundamental frequency and assumes perfect harmonicity [15, 16]. For many musical instruments, the harmonic assumption does not match due to the stiffness of the string and non-rigid terminations. For instance, it is generally agreed that the characteristic timbre of the piano is caused in part by the inharmonicity. Moreover, observing that the frequency partials are not at the expected ‘ideal’ spectral locations is essential for the wide-band pitch estimation and spectrum separation of harmonic sounds [19]. In this respect, some parsimonious models for inharmonicity were introduced based on the physics of musical instruments. The most common model expresses the partials position function of the fundamental frequency and the string stiffness coefficient [19]. An alternative approach mimics the inharmonic behavior of a stiffed string using a Laguerre transform [1]. In both approaches, the accuracy of the estimation of the partials position is crucial (to identify the stiffness coefficient or the Laguerre parameter). On the other hand, this estimation is inevitably contaminated by ambient noise, weak transmission/reproduction artifacts, and quantization errors due to the DFT processing. Therefore, contrary to the state-of-the-art approaches, we propose a time-domain Harmonize-Decompose approach (with no explicit estimation of the partial positions): the audio signal is first harmonized via a Laguerre transform, and then decomposed using the quasi-periodic signal extraction. The Laguerre factor together with the harmonic model parameters are optimized such that the output signal best fits the GM model.

The remainder of this paper is organized as follows. In Section 2, a brief overview of the global AM-FM signal representation and analysis techniques is presented. Next, the Harmonize-Decompose scheme is introduced and experimentally investigated in Sections 3 and 4, respectively. Finally, a discussion and concluding remarks are provided in Section 5.

2. GLOBAL AM-FM SIGNAL REPRESENTATION

In sinusoidal modeling, the signal is expressed by a sum of evolving sinusoids as in (1), where $\psi_k(n)$ represents the instantaneous phase of the k^{th} partial. Since the energy of the audio signal is concentrated around the multiples of the fundamental frequency f_0, $\psi_k(n)$ can be decomposed into

$$\psi_k(n) = 2\pi k n f_0 \; + \; 2\pi \varphi_k(n) \tag{2}$$

where $\varphi_k(n)$ characterizes the evolution of the instantaneous phases around the k^{th} harmonic, and can be assumed to slowly vary over time. In [16], we have assumed that the time variations of the instantaneous amplitudes and frequencies of the different harmonics are correlated, and we have expressed the audio signal as a superposition of harmonic components with global frequency selective amplitude modulation and global time-warping, i.e.,

$$s(n) = a_n(q)\, \theta \left(n + \frac{\varphi(n)}{f_0} \right) \tag{3}$$

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, and $2L{+}1$ denotes the amplitude modulating filter length. The introduction of q, where q^{-1} is the one sample time delay operator: $q^{-1}\theta(n) = \theta(n-1)$, allows us to introduce a compact notation of transfer functions in the time domain.
- $\theta(n) = \sum_k a_k \cos\left(2\pi k f_0 n + \Phi_k\right)$ is a $T = ceil\left\{\frac{1}{f_0}\right\}$ periodic signal (normalized waveshape), having a constant spectrum over the whole signal duration. $\theta(n)$ characterizes the spectral envelope of the audio source, and may be considered as a signature for the source (e.g., musical instrument) identification and recognition applications.
- $\varphi(n)$ denotes the global phase modulating signal that can be interpreted in term of global time-warping. The global phase modulation allows an accurate modeling and tracking of the fundamental frequency variations, but does not account for the inharmonic effects that may be present in the signal.

Audio enhancement and/or separation is performed by adjusting the degrees of freedom (in $a(q)$, φ, and θ) such that the received signal matches the best with the assumed model (in the least-squares sense). The degrees of freedom are estimated in a cyclic fashion. The proposed technique was shown to be effective for musical signal enhancement and separation [16]. Furthermore, the different parameters are related to the three basic features in music sounds: pitch (φ), intensity ($a(q)$), and timbre (θ). The proposed enhancement technique can also be interpreted as a sum of a scaled, translated and modulated harmonic atom (θ). However, contrary to the classic atomic decomposition approaches, the dictionary is not fixed: the atoms are adapted taking into consideration the structure of the received signal.

3. INHARMONIC GLOBAL AM-FM SIGNAL REPRESENTATION

Frequency-selective global modulation leads to a parsimonious representation that efficiently models the different modes of instantaneous amplitude variation with a limited parameter rate (the average number of parameters that appear in the description of one second of the signal). This fact leads to a good estimation vs. modeling noise tradeoff and an effective signal decomposition. The proposed representation, however, allows only for slow variations of the fundamental frequency and assumes perfect harmonicity. For many musical instruments (e.g. piano lower tones), the harmonic assumption is not matched and the proposed model fails to model and properly extract the musical notes [16]. Unfortunately, estimating the inharmonicity is a complex problem [11] and requires prior decision on what is harmonic and what is not. Furthermore, the fundamental frequency of the harmonic series must be known with high accuracy in order to avoid incorrect rendering of the higher order partials (since frequency errors generally increase with partial order). Last but not least, the signal re-synthesis from a rigid model results in an artificially and dull sounding audio due to the static relationship between partials [18].

Generally, a frequency-warped signal could be implemented by processing the input signal with the time-varying filter:

$$\lambda_n(q) = \lambda_{n-1}(q) * A(q) \tag{4}$$

where $A(q)$ is an all-pass filter that characterizes the frequency-warping. The warping scheme could be effectively expressed and implemented using a dispersive delay lines structure [1, ch.11]. Remark that the global time-warping (introduced in the QPSE structure) could also be interpreted as a linear frequency-warping and expressed as in (4), where $A(q) = q^{-Tf_0}$ accounts for the non-integer periodicity of the input signal. Another well investigated choice of $A(q)$ is the first-order all-pass filter

$$A(q) = \frac{q^{-1} - b}{1 - bq^{-1}} \tag{5}$$

By varying the real parameter $-1 < b < 1$, one obtains the family of Laguerre curves shown in Figure 1. It has been reported that

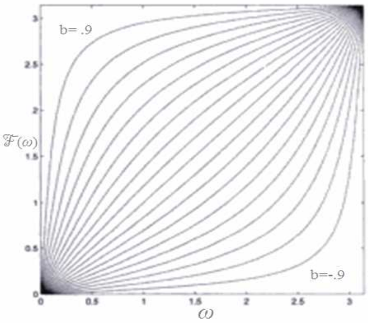

Figure 1: The family of Laguerre warping maps.

(with an appropriate choice of the Laguerre parameter b), a Laguerre curve closely matches the distribution of the partials of the low-pitch piano tones, and is consistent with the eigenfrequencies derived from the physical model of stiff strings [20]. The optimal Laguerre parameter was derived as a function of the fundamental and the shifted frequencies [1]. In this paper, we propose an alternative approach that does not rely on an explicit estimation of the partials positions. The basic observation is that the global modulation representation and its extraction accuracy depend on the uniform spacing of the frequency peaks. The extraction error is minimized if an appropriate warping curve compensates for inharmonicity.

Taking into consideration the inharmonicity in the global AM-FM representation, the audio signal could be expressed as:

$$s(n) = a_n(q)\mathcal{F}_b\left\{\theta\left(n+\frac{\varphi(n)}{f_0}\right)\right\} \tag{6}$$

where $\mathcal{F}_b\{.\}$ denotes the Laguerre transform operator.

Given that the amplitude modulation filter varies slowly over time and that the Laguerre parameter is generally small (limited frequency warping), the audio signal could be approximated as:

$$s(n) \approx \mathcal{F}_b\left\{a_n(q)\theta\left(n+\frac{\varphi(n)}{f_0}\right)\right\} \tag{7}$$

Therefore, we propose a time-domain Harmonize-Decompose (HD) approach that jointly optimizes the Laguerre transformation and AM-FM decomposition:

$$\min_b\left\{\min_{a(q),\theta,\varphi}\left\|y_b(n) - a_n(q)\theta\left(n+\frac{\varphi(n)}{f_0}\right)\right\|^2\right\} \tag{8}$$

where $y_b(n) = \mathcal{F}_b^{-1}\{y(n)\} = \mathcal{F}_{-b}\{y(n)\}$ is the harmonized received signal (the link between the Laguerre transform at its inverse is a consequence of the mirror symmetry of the Laguerre warping curves (Figure 1)). By plotting the curve of normalized decomposition error (defined in (9)) function of the Laguerre parameter (see Figure 2), we have noticed that the decomposition error is generally a convex function of the Laguerre parameter. Thus, we propose a fast golden-section scheme to search/identify the optimal parameter b. The (fast) quasi-periodic signal extraction algorithm is used for the AM-FM decomposition of the harmonized audio signal [16].

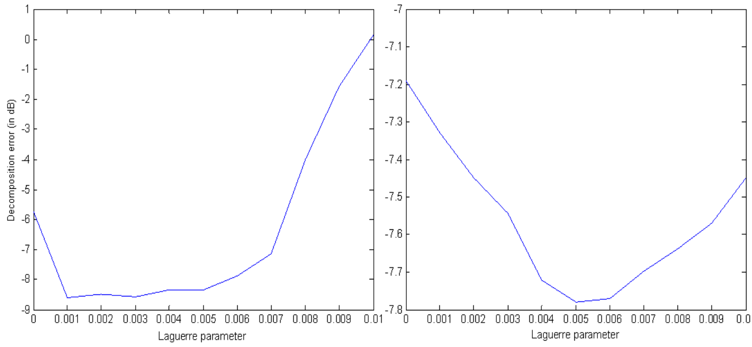

Figure 2: Decomposition error vs. Laguerre parameter b for the piano tones $A2$ (left) and $G3$ (right).

4. EXPERIMENTAL RESULTS

We validate the proposed extraction approach using real piano signals. The audio signals were recorded at 44.100 kHz, then downsampled to 11.025 kHz. We have computed the normalized decomposition error:

$$Err = \frac{\sum_n\left(\mathcal{F}_{\widehat{b}}^{-1}\{y(n)\} - \widehat{s}(n)\right)^2}{\sum_n y(n)^2} \tag{9}$$

where $\mathcal{F}_{\widehat{b}}^{-1}\{y(n)\}$ is the harmonized received signal and $\widehat{s}(n) = \widehat{a}_n(q)\widehat{\theta}\left(n+\frac{\widehat{\varphi}(n)}{f_0}\right)$ is the reconstructed harmonic signal with global AM-FM modulation. Note that $\sum_n \mathcal{F}_{\widehat{b}}^{-1}\{y(n)\}^2 = \sum_n y^2(n)$ as the Laguerre transform is unitary (energy preserving) [1, Ch11]. Remark also that due to the energy preservation property, the normalized decomposition error is also a good measure of the enhancement accuracy of the overall scheme (after reconstruction).

We have compared the harmonic global AM-FM decomposition (computed with Quasi-periodic Signal Extraction (QPSE) algorithm [16]), and the proposed Harmonize-Decompose approach (that we refer to as HD-QPSE). The smoothing AM and FM modulation factors were set to $T_a = T_f = T_\varphi = 3T$ ($T = ceil(1/f_0)$ is the period of the harmonic component, assumed known). The Laguerre transform was computed using the short-time Laguerre transform (segmented with a 512 length Hamming window with 50 % overlap).

Fig. 3 and 4 plot the normalized decomposition errors for two piano tones $A2$ and $G3$. As a reference, we have plotted the decomposition accuracy of a time-frequency based representation. The desired signals were retrieved using an ABSOLA analysis/synthesis algorithm (with peaks interpolation and tracking) [1, Ch10]. In the time-frequency processing, the block size, zero-padding factor, and maximum number of sinusoids were set to 512, 2, and 32 respectively (the signals were segmented using a Hamming window with 50% overlap).

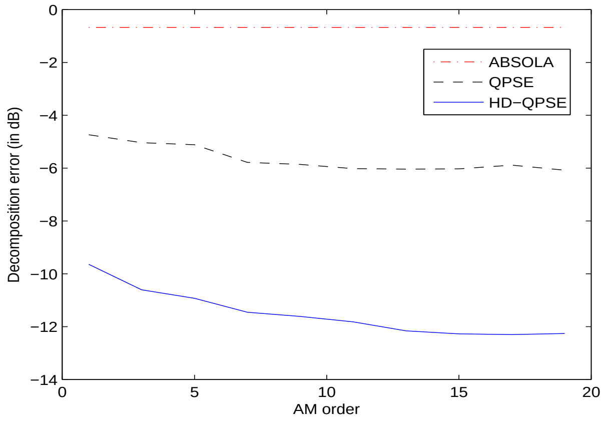

Figure 3: Normalized decomposition error function of the amplitude modulating order L (piano tone $A2$).

Curves show that the Laguerre transformation effectively compensates the piano inharmonic effects and considerably improves the extraction accuracy. The relative enhancement does not depend on the AM-FM modeling (the order of the amplitude modulating filter), and it is more critical for lower tones (as the inharmonicity artifacts are more severe).

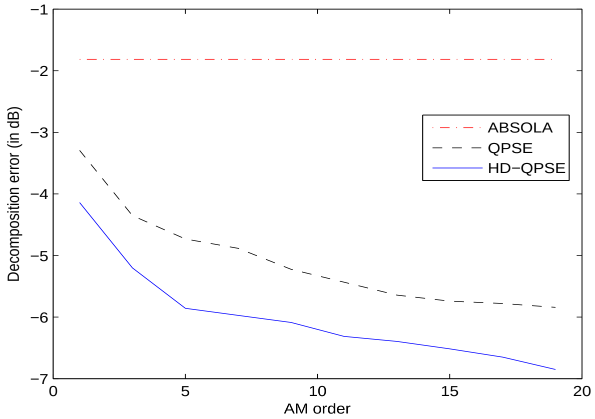

Figure 4: Normalized decomposition error function of the amplitude modulating order L (piano tone $G3$).

Applying the HD-QPSE scheme on guitar signals shows a limited improvement over the harmonic QPSE analysis. The improvement vanishes if we relax the zero-phase constraint of the amplitude modulating filter (e.g. by considering causal filters). In such a case, the AM filter tracks the variations of both amplitude (expressing the various decay modes) and phase (accounting for the guitar (slight) inharmonicity). Indeed, because the guitar string is only slightly inharmonic ($b \ll 1$), the Laguerre update filter could be approximated by:

$$A(q) = \frac{q^{-1} - b}{1 - bq^{-1}} \approx -b + (1 - b^2)q^{-1} + bq^{-2} \qquad (10)$$

Thus, one could express the Laguerre transform as a linear smoothing of the time-varying amplitude modulating filter (similar to is proposed in [16]).

5. CONCLUDING REMARKS

In this paper, we have investigated signal enhancement techniques exploiting the structure of the audio signal and accounting for inharmonicity artifacts. Contrary to the state of the art approaches, our scheme does not require an explicit estimation of the partials' positions. We propose a Harmonize-Decompose approach where the audio signal is first harmonized via a Laguerre transform, then decomposed using the quasi-periodic signal extraction. The Laguerre factor and global harmonic AM-FM parameters are jointly optimized such that the output signal best fits the global modulation model. Simulations show that the HD extraction technique oversteps the limitation of the global AM-FM representation and analysis techniques and allows the processing of inharmonic string instruments (e.g. piano). We have also observed that a slight inharmonicity (e.g. in a guitar) could be considered by relaxing the zero-phase constraint on the amplitude modulating filter.

6. REFERENCES

[1] U. Zölzer (Ed.), "DAFX - Digital Audio Effects," *John Wiley & Sons*, 2002.

[2] J. Beauchamp, "Analysis and Synthesis of Musical Instrument Sounds," *in Analysis, Synthesis, and Perception of Musical Sounds*, J. Beauchamp (Ed.), Springer, 2007.

[3] M.G. Christensen and A. Jakobsson, "Multi-Pitch Estimation," *Morgan & Claypool Publishers*, 2009.

[4] R. McAulay and T. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol.34, Aug. 1986.

[5] F. Keiler and S. Marchand,"Survey On Extraction of Sinusoids in Stationary Sounds," *In Proc. of DAFX*, Sept. 2002.

[6] E. Aboutanios,"A Modified Dichotomous Search Frequency Estimator," *Signal Processing Letters*, Feb. 2004.

[7] M.G. Christensen, A. Jakobsson, and S.H. Jensen, "Joint High-Resolution Fundamental Frequency and Order Estimation," *IEEE Trans. on Audio, Speech, and Language Processing*, Jul. 2007.

[8] R. Badeau, B. David, and G. Richard, "High-Resolution Spectral Analysis of Mixtures of Complex Exponentials Modulated by Polynomials," *IEEE Trans. on Signal Processing*, Apr. 2006.

[9] X. Serra and J. Smith, "Spectral Modeling Synthesis:A Sound Analysis/Synthesis Based on a Deterministic plus Stochastic Decomposition," *Computer Music Journal*, 1990.

[10] T. Virtanen and A. Klapuri,"Separation of Harmonic Sound Sources Using Sinusoidal Modeling," *In Proc. of ICASSP*, 2001.

[11] X. Wen, "Harmonic Sinusoid Modeling of Tonal Music Events," *PhD Thesis*, University of London, 2007.

[12] R. Gribonval and E. Bacry, "Harmonic Decomposition of Audio Signals with Matching Pursuit," *IEEE Trans. on Signal Processing*, Jan. 2003.

[13] A. Potamianos and P. Maragos, "Speech Analysis and Synthesis Using an AM-FM Modulations Model," *Speech Communication*, Jul. 1999.

[14] R.B. Sussman and M. Kahrs,"Analysis and Resynthesis of Musical Instrument Sounds Using Energy Separation," *In Proc. of ICASSP*, May. 1996.

[15] M. Triki and D.T.M. Slock, "Periodic Signal Extraction with Global Amplitude and Phase Modulation for Music Signal Decomposition," *In Proc. of ICASSP*, March 2005.

[16] 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 MMSP*, Oct. 2008.

[17] M. Triki and D.T.M. Slock, "Perceptually Motivated Quasi-Periodic Signal Selection for Polyphonic Music Transcription," *In Proc. of ICASSP*, Apr. 2009.

[18] T.H. Andersen and K. Jensen, "Importance of Phase in the Sinusoidal Model," *J. Audio Eng. Soc.*, Nov. 2004.

[19] A. Klapuri, "Wide-band Pitch Estimation for Natural Sound Sources with Inharmonicities," *AES Convention*, May 1999.

[20] G. Evangelista and S.Cavaliere, "Discrete Frequency Warped Wavelets: Theory and Applications," *IEEE Trans. on Signal Processing*, Apr. 1998.