IMPROVED ESTIMATION OF THE AMPLITUDE ENVELOPE OF TIME-DOMAIN SIGNALS USING TRUE ENVELOPE CEPSTRAL SMOOTHING

Marcelo Caetano, Xavier Rodet

Analysis/synthesis Team, IRCAM

ABSTRACT

The amplitude modulations of musical instrument sounds and speech are important perceptual cues. Accurate estimation of the amplitude, or equivalently energy, envelope of a time-domain signal (waveform) is not a trivial task, though. Ideally, the amplitude envelope should outline the waveform connecting the main peaks and avoiding over fitting. In this work we propose a method to obtain a smooth function that approximately matches the main peaks of the waveform using true envelope estimation, dubbed true amplitude envelope. True envelope is a cepstral smoothing technique that has been shown to outperform traditional envelope estimation techniques both in accuracy of estimation and ease of order selection. True amplitude envelope gives a reliable estimation that follows closely sudden variations in amplitude and avoids ripples in more stable regions with near optimal order selection depending on the fundamental frequency of the signal.

Index Terms— Amplitude envelope estimation, true envelope, cepstral smoothing, musical instrument sound, speech

1. INTRODUCTION

The estimation of the amplitude envelope of a time-domain signal (or waveform) is a classical engineering problem that arises in amplitude demodulation [1], onset detection [2], and attack time estimation [3], temporal modeling of sounds such as automatic transcription [4] and temporal evolution of musical instrument sounds [5]. The amplitude envelope is a perceptually important feature of musical instrument sounds and speech. It has been shown to be correlated to the percussiveness of musical instruments sounds [6], to speech intelligibility [7] and even to affect pitch perception [8]. The classical approach to the estimation of the amplitude envelope of a time-domain signal is the technique known as envelope follower [1], [4], [6], which consists basically of rectifying the waveform and then low-pass filtering it. This can be easily implemented in either the analog or digital domains. Another very popular solution is to calculate the instantaneous root mean square (RMS) value of the waveform through a sliding window with finite support [9]. More recently, some authors have proposed other techniques to obtain a more reliable estimation of the amplitude envelope of waveforms. An early attempt [4] consisted of a piece-wise linear approximation of the waveform. The amplitude envelope is created by finding and connecting the peaks of the waveform in a window that moves through the data. Jensen [10] proposed a method that fits curve shape approximations to model the amplitude envelope of the partials of an additive model of instrument sounds and later Skowronek [6] applied it to approximate the global amplitude envelope. However, advances in spectral envelope estimation techniques such as linear prediction [11] and cepstral smoothing [12] are promising candidates to obtain an accurate estimation of the amplitude envelope when applied in the time-domain [5]. In this work we propose to apply the true envelope estimation technique in the time domain to accurately estimate the amplitude envelope of waveforms such as isolated musical instrument sounds or speech. We will show that the true amplitude envelope renders estimations that respond well to sudden changes in amplitude while remaining smooth during more stable regions of the waveform.

In the next section we present the classical amplitude envelope estimation techniques, followed by the recently proposed use of linear prediction. Next, we introduce the true envelope cepstral smoothing algorithm and we explain how we apply it in the time domain. Finally, we present and evaluate the results, comparing our proposed approach with the others.

2. CLASSICAL AMPLITUDE ENVELOPE ESTIMATION

The classical amplitude envelope estimation techniques explained in this section are low-pass filtering (LPF), root-mean square (RMS), and analytic signal. We will also review a recent proposal to the use of linear prediction [13], dubbed frequency-domain linear prediction (FDLP).

2.1. Low-Pass Filtering (LPF)

Low-pass filtering is the most straightforward way of obtaining a smooth signal that follows the amplitude evolution of the original waveform. It is based on a classical amplitude demodulation envelope follower technique [2], that low-pass filters a half-wave (*hwr*) or full-wave rectified (*fwr*) version of an amplitude modulated (AM) signal. The principle of amplitude modulation (AM) is that the amplitude changes of the signal carry the information we seek. There are many possible filter designs with different characteristics and the choice affects the quality of the final envelope. For instance, Jensen [10] proposes to convolve the waveform with a Gaussian window function, resulting in a suboptimal estimation. Also, the cut-off frequency of the filter has a major impact on the result. High cut-off frequencies will likely produce an amplitude envelope with ripples and very low cut-off frequencies are less responsive to sudden amplitude changes.

2.2. Root-Mean Square (RMS) Energy

The RMS value is perhaps the most popular [9] method for estimating the temporal evolution of the signal energy because it can be easily adapted to obtain an estimate of the amplitude envelope by simply applying it with a sliding window, as shown in equation (1)

$$RMS(t) = \sqrt{\frac{1}{T}\sum_{i=1}^{T} w_i(t) x_i^2(t)} \tag{1}$$

where $x_i(t)$ is the i^{th} sample of the signal centered around t as seen through the window $w_i(t)$, t is the number of samples the analysis window moves, and *T* is the window length. Usually a rectangular window is used, but other choices are also possible [10]. The RMS calculation is a special case of the generalized mean with exponent *p=2* and as such, also functions as a sort of moving average, low-pass filter that smoothes out the signal. The analysis step *t* imposes a trade-off between the temporal sampling rate of the envelope and how much information it represents. Small values of *t* react sooner to sudden changes in amplitude, while presenting ripple in more steady regions and larger values smooth out the ripples but tend to lag behind abrupt energy changes.

2.3. Analytic Signal

The Hilbert transform is part of a signal processing technique for amplitude demodulation [1]. The Hilbert transform of a signal *x*(*t*) is defined as

$$\hat{x}(t) = x(t) * \frac{1}{\pi t} \qquad (2)$$

where * stands for convolution. Using equation (2), we can define the analytic signal *z*(*t*) as

$$z(t) = x(t) + j\hat{x}(t) = r(t)\exp[j\theta(t)] \qquad (3).$$

The analytic signal is useful for envelope detection since its modulus *r*(*t*) and time derivative of the phase $\theta(t)$ can serve as estimates for the amplitude envelope and instantaneous frequency of *x*(*t*) under certain conditions. Notably, if the Hilbert transform of *x*(*t*) is equal to its quadrature signal [1], then the estimates are equal to the actual information signals [1]. Synthetic (i.e., AM) signals can be constructed to have this property, but there is no reason to expect that acoustic musical instrument sounds or speech also present it. A more realistic condition is verified when we are dealing with narrowband signals [14], which is rarely the case for musical instrument sounds and speech. The analytic signal can be effectively used to extract the amplitude envelope of individual partials if applied to each frequency bin of the STFT, but when applied to the whole signal it is equivalent to trying to demodulate several AM signals at the same time, so we use it as half-wave rectifier in this work.

2.4. Frequency-Domain Linear Prediction (FDLP)

Traditional linear prediction [11] estimates the spectral envelope from the time-domain signal. The idea behind FDLP [13] is to exploit time-frequency duality to extract the temporal amplitude envelope by applying linear prediction to a spectral representation. In particular, the used spectral representation is the discrete cosine transform (DCT) given by equation (4), since it is real-valued.

$$\hat{X}(k) = \sum_{n=0}^{N-1} x(n)\cos\left(\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right) \qquad (4)$$

The envelope peaks, whose number and width are determined by the model order, will now be their frequency domain counterparts, the rectified waveform peaks. Thus, the model order has to be adjusted with respect to the temporal structure of the signal, and not to the formant structure of the spectrum.

3. TRUE ENVELOPE ESTIMATION

The true envelope estimator [12] has been shown to outperform linear prediction [11] or cepstral methods such as discrete cepstrum [15] both in terms of accuracy and ease of model order selection. Recently the iterative procedure has been significantly improved such that the computational costs are in the similar to the costs of the Levinson recursion such that real time processing can be achieved [12]. True envelope estimation is based on cepstral smoothing of the log amplitude spectrum and the resulting estimation can be interpreted as the best band limited interpolation of the major spectral peaks in such a way that the peak matching is maximized and inter-peak valleys are avoided.

3.1. Cepstral Smoothing

The real cepstrum is usually defined as the inverse Fourier transform of the log magnitude spectrum [12], as shown in equation (5)

$$\hat{x}(n) = \sum_{k=0}^{K-1} \log\left|X(k)\right| \exp\left(\frac{j2\pi kn}{N}\right) \qquad (5).$$

Regarding the log magnitude spectrum as a signal, we can interpret each cepstral coefficient as a measure of the energy present in discrete frequency bands of that signal. Low-pass filtering the cepstrum (also called liftering) would result in a smoother version of the log magnitude spectrum, given by equation (6)

$$C(k) = \sum_{n=0}^{N-1} w_n \hat{x}(k) \exp\left(\frac{-j2\pi kn}{N}\right) \qquad (6)$$

where *C*(*k*) is the smoothed spectrum (corresponding to the spectral envelope estimation) and w_n is a low-pass window in the cepstral domain usually defined as

$$w_n = \begin{cases} 1, |n| < n_c \\ 0.5, |n| = n_c \\ 0, |n| > n_c \end{cases} \qquad (7)$$

where n_c is the cutoff quefrency. If we only want to represent the spectral envelope, discarding information about the partials we should set the cutoff quefrency below the period of the signal. One major drawback of this operation is that we discard spectral energy when setting cepstral coefficients to zero. The result is a smooth curve that is always below the peaks of the log magnitude spectrum. The true envelope estimator uses cepstral smoothing in an iterative procedure that aims at connecting the peaks of the log magnitude spectrum as explained below.

3.2. True Envelope

Let *X*(*k*) be the *K*-point DFT of the signal frame *x*(*n*) and $C_i(k)$ the smoothed spectrum at iteration *i*. The algorithm then iteratively updates the resulting spectral envelope $A_i(k)$ with the maximum of the original spectrum and the current spectral envelope $C_{i-1}(k)$

$$A_i(k) = \max\left(\log\left|X(k)\right|, C_{i-1}(k)\right) \qquad (8)$$

and applies cepstral smoothing to $A_i(k)$ to obtain $C_i(k)$. The procedure is initialized setting A_0(k) = log|*X*(*k*)| and starting the cepstral smoothing to obtain $C_0(k)$.

3.4. Optimal Order Selection

The order of the cepstral representation of the spectral envelope is the number of cepstral coefficients we keep in the cepstral smoothing procedure, and as such is proportional to the fundamental frequency of the original signal. The optimal order should give a spectral envelope that follows the overall shape of the filter without representing the harmonic structure of the spectrum. In order to estimate the optimal order, we use the source-filter model and think of the spectrum as the result of the interaction of two components, represented by the source, an input signal that contains information about the frequencies of the partials and the filter that shapes the source spectrum. According to this model, the spectral envelope represents the filter that has been excited by the source. For near harmonic sources, the resulting spectrum will be quasi-harmonic. In terms of the interaction between source and filter, we can think of the resulting harmonics sampling the filter with a sampling rate that depends on the fundamental frequency of the source spectrum. According to the sampling theorem, we must sample the filter with at least twice the maximum frequency present in that signal. If we assume that the spectral envelope should not contain information about the harmonic structure of the spectrum, the maximum frequency present in that signal is the fundamental frequency F_0, such that the related Nyquist frequency (assuming a sampling rate of Fs) is $Fs = (2F_0)$. This formula provides a simple way of selecting the cepstral order because higher sampling frequencies would reveal (maybe partially) information about the harmonic structure of the spectrum and lower sampling frequencies would smooth out the spectral envelope, not revealing information about the (formant) peaks. We can therefore postulate the near optimal order of the cestrum given only that the maximum frequency difference between two spectral peaks that carry envelope information is known. If the difference between those peaks is Δ_F then the cepstral order should be

$$\hat{O} = \frac{F_S}{2\Delta_F} = \alpha \frac{F_S}{F_0}, \alpha = 0.5 \qquad (9).$$

While the optimal order, that is the order that provides an envelope estimate with minimum error, depends on the specific properties of the envelope spectrum, the order selection according to equation (9) is reasonable for a wide range of situations and the resulting error is generally rather close to the one obtained with the optimal order.

4. TRUE AMPLITUDE ENVELOPE (TAE)

Ideally, the amplitude envelope should be a curve that outlines the waveform, following its general shape without representing information about the harmonic structure. One of the most challenging aspects of this problem is that we are looking for a curve that is smooth during rather stable regions of the waveform, while being able to react to sudden changes (such as percussive onsets) when they occur. Here, we propose to use a dual of true envelope in the time domain. The time domain signal is subjected to the algorithm instead of the Fourier spectrum. In this way, the amplitude envelope is expected to match the amplitude peaks corresponding to the period of the waveform more closely than the previously introduced methods. The idea behind TAE is to mimic the structure of the spectrum with the time-domain signal to be able to apply the true envelope method directly. The basic steps to estimate the TAE are as follows. First we obtain a rectified version of the waveform (so that that are no negative amplitudes), next we zero-pad the rectified waveform to nearest power of two (thus mimicking the DFT), and then we finally add a time-reversed version of the zero-padded rectified waveform to represent the negative frequencies. Before estimation, we still need to exponentiate the amplitudes because true envelope supposes that we are fitting a smooth curve to the log magnitude spectrum. The result is illustrated in Fig. 1. The last step is the application of the true envelope estimation technique to obtain the true amplitude envelope (TAE), represented as a solid line outlining the rectified waveform.

It is important to notice that the peaks of the waveform do not carry the same information as the spectral peaks. Each peak of the spectrum corresponds to a partial, such that for quasi-harmonic spectra the separation between spectral peaks is F_0. On the other hand, in only one period, the peaks of the half-wave rectified waveform generally contain information about all the frequencies contained in that signal (depending on their phases). Therefore, the time-domain counterpart of the near optimal order selection must take into account only the period of the waveform, instead of all rectified peaks. The optimal order is now directly proportional to the fundamental frequency of the waveform, instead of inversely proportional when using true envelope in the spectral domain because the separation of the spectral peaks Δ_F is now represented by Δ_T and given by the period of the signal T_0 for a half-wave rectified waveform (*hwr*) as equation (10) shows

$$\hat{O} = \frac{F_S}{2\Delta_T} = \alpha \frac{F_S}{T_0}, \alpha = 0.5 \qquad (10).$$

A full-wave rectified (*fwr*) version would present twice as many main peaks, requiring half T_0, or α=1.

5. EVALUATION

This section compares the amplitude envelopes obtained with LPF, RMS, FDLP and TAE for a sustained and a percussive musical instrument sounds and a speech utterance. Figure 2 shows a half-wave rectified (*hwr*) version of a bass clarinet, a clavinet and speech utterance waveforms and the amplitude envelope estimates. We are looking for the estimation that best fits the model waveforms, following the amplitude evolution by matching the main peaks while avoiding ripples in more stable parts. It should be noted that FDLP was normalized and scaled to the maximum of the *hwr* waveform and that RMS was also low-pass filtered to eliminate the ripples during mostly the more stable parts. Upon close inspection, Figure 2 shows that TAE renders the best fit, closely following the peaks without ripples. TAE is also very responsive to sudden changes in amplitude, as can be seen for the clavinet sound. We should notice that the fit depends largely on the order. For isolated notes, the optimal order is expected to be the same throughout, but for speech the situation is not so simple because the F_0 changes dynamically. For speech utterances we can use the mean F_0 (supposing it does not change a great deal).

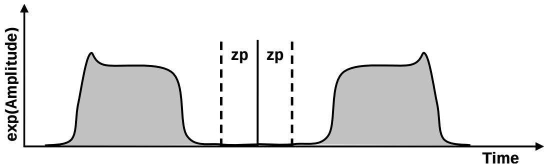

Fig. 1. True amplitude envelope estimation. The figure shows the half-wave rectified and zero padded (zp) version of the waveform with its time-reversed counterpart used in the true amplitude envelope estimation method.

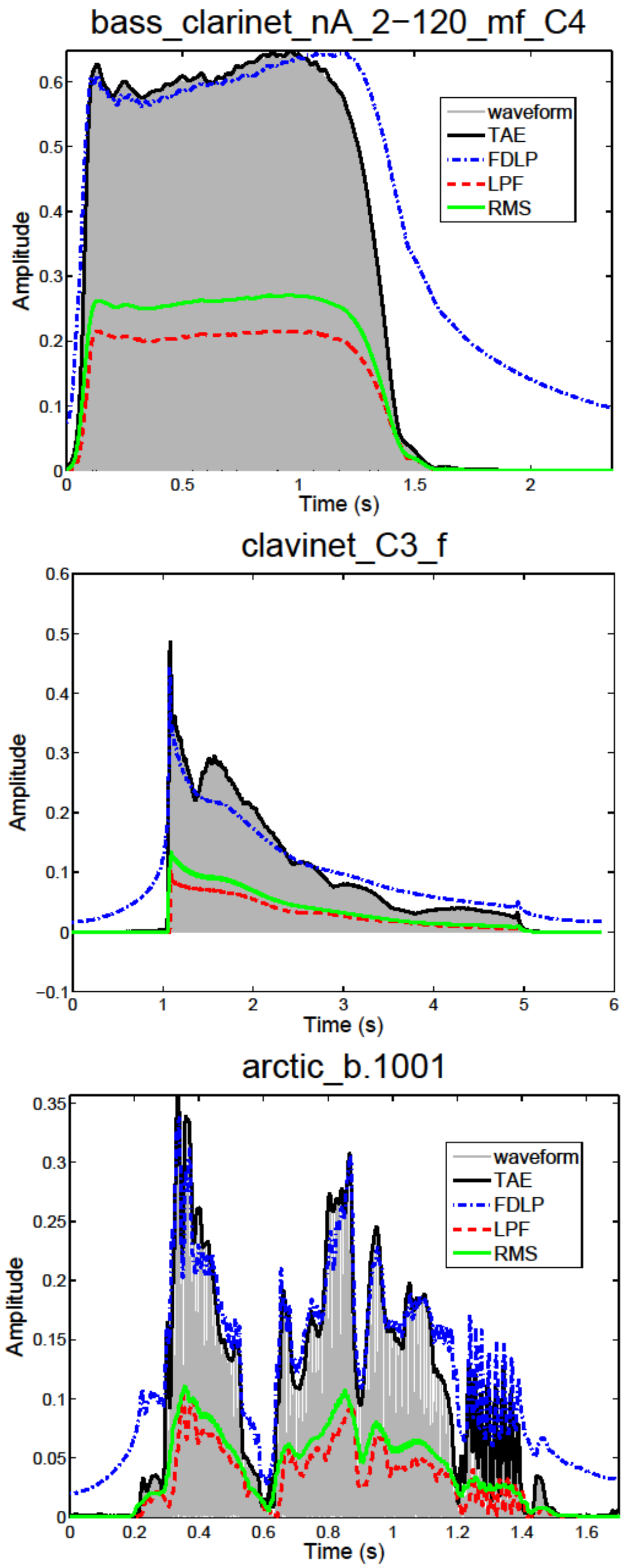

Fig. 2. Half-wave rectified (*hwr*) waveform and amplitude envelope estimation methods; true amplitude envelope (TAE), frequency-domain linear prediction (FDLP), root-mean square (RMS) and low-pass filtering (LPF).

6. CONCLUSIONS AND FUTURE PERSPECTIVES

We proposed the true amplitude envelope (TAE) estimation technique as an improvement to the classical methods found in the literature. We have shown that TAE outperforms them by providing a smooth function that approximately matches the main peaks of the waveform avoiding under or over estimation due to the near optimal order selection based on the period of the signal. One important feature of TAE due to its cepstral nature is the ability to represent well sudden changes in the waveform while avoiding ripples during more stable parts. Future perspectives of this work could include evaluating the perceptual impact of the resulting envelopes with listening tests, developing an objective measure to evaluate the results and order estimation for signals with time varying F_0.

7. ACKNOWLEDGEMENTS

This work is supported by the Brazilian Governmental Research Agency CAPES (process 4082-05-2).

8. REFERENCES

[1] P. Maragos A. Potamianos, "A Comparison of the Energy Operator and the Hilbert Transform Approach to Signal and Speech Demodulation," vol. 17, no. 1, 1994.

[2] L. Daudet, S. Abdallah, C. Duxbury, M. Davies, M.B. Sandler J.P. Bello, "A Tutorial on Onset Detection in Music Signals," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 1035-1047, 2005.

[3] D. Luce and M. Clark, "Durations of Attack Transients of Nonpercussive Orchestral Instruments," *J. Audio Eng. Soc.*, vol. 13, no. 3, pp. 194-199, 1965.

[4] A. Schloss, "On the Automatic Transcription of Percussive Music - From Acoustic Signal to High-Level Analysis," Stanford University, Ph.D. Thesis 1985.

[5] J.J. Burred, X. Rodet M. Caetano, "Automatic Segmentation of the Temporal Evolution of Isolated Acoustic Musical Instrument Sounds Using Spectro-Temporal Cues," in *Proc. DAFx*, 2010.

[6] M. McKinney, J. Skowronek, *Features for Audio Classification: Percussiveness of Sounds.*: Springer, 2006.

[7] R. Drullman, "Temporal Envelope and Fine Structure Cues for Speech Intelligibility," *J. Acoust. Soc. Am.*, vol. 97, no. 1, pp. 585-591, 1995.

[8] W.M. Hartmann, "“The Effect of Amplitude Envelope on the Pitch of Sine Wave Tones,”," *J. Acoust. Soc. Am.*, vol. 63, pp. 1105-1113, 1978.

[9] J. Hajda, "“A New Model for Segmenting the Envelope of Musical Signals: The relative Salience of Steady State versus Attack, Revisited,”," *J. AES*, Nov. 1996.

[10] K. Jensen, "“Envelope Model of Isolated Musical Sounds,”," in *Proc. DAFx*, 1999.

[11] J. Makhoul, "“Linear prediction: A Tutorial Review”," *Proc. IEEE*, vol. 63, pp. 561-580, Apr. 1975.

[12] F.Villavicencio, X. Rodet, A. Röbel, ""On Cepstral and All-Pole Based Spectral Envelope Modeling with Unknown Model Order,"" *Pattern Recognition Letters*, vol. 28, pp. 1343-1350, 2007.

[13] D.P.W. Ellis, M. Athineos, "“Frequency-Domain Linear Prediction for Temporal Features.”," in *Proc. IEEE ASRU Workshop*, 2003.

[14] W.M. Hartmann, *Signals, Sound, and Sensation.*: Springer-Verlag-AIP Press, 1997.

[15] X. Rodet, T. Galas, "“An Improved Cepstral Method for Deconvolution of Source-Filter Systems with Discrete Spectra: Application to Musical Sound Signals,”," in *Proc. ICMC*, 1990.