WHISPER: a robust histogram-based spectral envelope extraction algorithm with pitch detection potential

Elena Novaretti
Music composer, DSP coder
elena@elenadomain.it - Rapallo (GE) ITALY

[KEYWORDS: spectral envelope estimation, spectral envelope extraction, period histogram]

May, 2021

ABSTRACT

In this paper a novel method for realtime spectral envelope extraction, especially suitable for human voice but also for generic audio signals, is presented (codename WHISPER - Weighted HIstogram SPectral Envelope extRaction). The short-time Fourier magnitude spectrum is progressively smoothed in frequency domain by means of a modified bi-directional 1-pole low-pass filter with increasing cutoff period and lifted towards prominent peaks. At every step a cumulative periodicity histogram is populated with the size of the gaps forming between local maxima left in the smoothing curve, weighted by their pivotal magnitude. Under the right operating conditions, a highest peak always tends to grow at the histogram position corresponding to the fundamental frequency (F0) in case of periodic signals, even corrupted by noise, or to a suitably low-end position in case of non-periodic or noise signals, in both cases representing the optimal spectral envelope sampling period *P.* The cutoff period is continuously updated as the histogram's weighted average and the process iterated until no trace of periodicity can be detected in the growing curve any more, according to the periodicity histogram being computed and to a second volatile histogram populated with the gaps forming at every pass. In few iterations a smooth and time-stable envelope curve, never under- or over-fitting, is grown at very little computation cost, without resorting to any peak-picking, spline interpolation or cepstral means. In case of harmonic spectra, the algorithm has also the potential to provide pitch and harmonics-to-noise ratio information, making it suitable for integrated spectral envelope and pitch detection.

1. Introduction

Estimation of the spectral envelope is both important and advantageous in many branches of physics as well as audio engineering and digital signal processing. In music applications in particular, reliable estimation of the evolving spectral envelope of a sound opens up a wide range of advanced sound (re-)synthesis and manipulation techniques hardly achievable otherwise, and which can be implemented within the consolidated Phase Vocoder[1] or similar frameworks. Some examples are inverse-filtering and impulse response recovery, spectral whitening, pitch shifting with formant preservation[2], creative manipulation of a vocal timbre, and surely many others.

Unfortunately at the time of writing no *conclusive* methods for spectral envelope estimation have been produced by the current art, being the very problem plagued by some unavoidable ill-conditioning.

Several methods have been proposed in literature though, some even quite complex and computationally demanding, each one with its own both strong and weak points[3,4,5 and others]

The knowledge of a spectral envelope curve *E(w)* such that

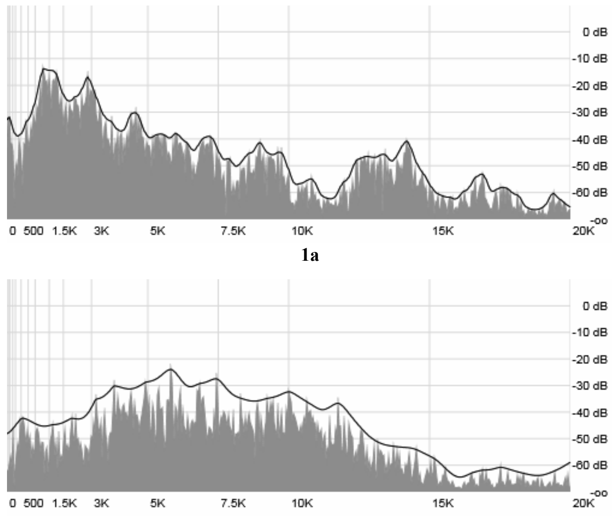
$$S(w) = W(w) \cdot E(w)$$

where *S* is a given magnitude spectrum and *W* is a flattened (“whitened”) version of it which we of course aren't to know, is by definition impossible and no unique solution exists. We cannot, in fact, discover two numbers given only their product; the most we can do is estimate them if we have some clues to do that.

Actually, despite inifitine pairs of *W* and *E* exist which give the starting spectrum *S* when multiplied together, only a limited subset of such pairs is what we aim at: the wanted envelope curve shall 1. pass thru all *prominent* peaks in *S* only and 2. never pass under any peaks in *S.* The big problem being that no mathematically sound definition can be given to «prominent» peaks, despite their naked-eye determination may look trivial. The only vague definition we can attempt for them is the most signifying peaks which represent the spectral energy distribution thru frequency left after an imaginary carving operation.

Only in case of a perfectly harmonic spectrum the prominent peaks coincide unambiguously with the peaks of the harmonic series, while in all other real world scenarios (as in case of harmonic spectra mixed with or constituded entirely by noise or non-harmonic spectra) their discrimination may become problematic. A robust and reliable envelope estimator (EE) is expected nevertheless to return a consistent curve in any situation, even in case of spectra representing aspirated (unvoiced) vowels, unvoiced consonants (as the letters F or T), voiced consonants containing a noise portion (as the letter S in Please or J in the French Je) or even percussions or colored/filtered noise - a fact this often overlooked in literature (see Fig.1)

Missing therefore any local criterion to decide whether a given peak can be labeled as prominent or not, the best we can do is aiming at a global estimate of an optimal frequency-domain grid period *P* to properly downsample the magnitude spectrum in order to obtain an envelope curve which can satisfy the first of the simple requirements listed above, even at the cost (negligible in most real world cases) of losing accuracy in all those cases where such an optimal period is locally varying, as in case of polyphonic or deliberately mixed spectra. We even dare saying that no spectral envelope estimation attempt can be pursuited at all without a preliminary knowledge of *P*.



1a

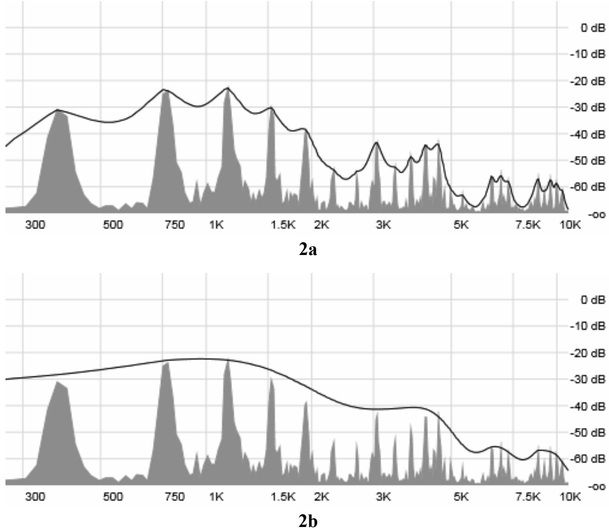
1b

Fig. 1 *Spectral envelopes of non-harmonic sounds: (1a) the aspirated (unvoiced) letter A and (1b) a crash cymbal*

It is evident that in case of harmonic or predominatingly harmonic spectra, *P* corresponds to F0. In case of polyphonic spectra, *P* shall arguibly correspond to the fundamental of the series whose overall magnitude is the largest. In case of noise instead, where prominent peaks are much more dense, *P* shall be a suitably small value in order to proper sample the highly irregular spectral magnitude shape.

The computation of a spectral envelope translates ultimately to a smoothing process (plus a lifting process), no matter which method is adopted. If cepstral smoothing is employed (a very disadvantageous choice in our opinion, even computationally), a cepstral order has to be imposed. In case of LPC, a definite LPC order has to be chosen aswell. If smoothing is carried in frequency domain treating the magnitude spectrum as a time-domain signal, once more a definite cutoff frequency for the kind of low-pass filter used shall be decided. Even when proceeding with peak-picking methods followed by interpolation, a kind of distance factor or the width of a sliding maximum window shall be similarly imposed, to decide which peaks to select and which ones to discard, which still indirectly translates to a smoothing process. It is evident that there can be no magic number we can use as the optimal smoothing factor for all kinds of spectra, because such value is strictly dependent on the short-time spectral structure. A smoothing value producing a tightly fitting curve for, say, a harmonic spectrum with a fundamental frequency of 500 Hz would produce a curve which starts descending thru the prominent peaks with a fundamental of 1 Khz (*under-fitting*) and conversely a too coarse curve over-estimating the true spectral trend (*over-fitting*) with a fundamental of, say, 200 Hz or lower (see Fig.2).

Any method not considering the strict dependency of *P* on the spectral content is doomed to fail producing poor results.
Some authors suggest standard autocorrelation (AC)[6] as a mean to estimate F0 and to use it as smoothing factor. Such a choice unfortunately is valid only in case of perfectly harmonic spectra (as in case of unambiguous voiced segments); in case of non-harmonic spectra like noise, any pitch detection (PD) algorithm like AC or the Spectral Comb[7] or the Harmonic Product Spectrum tend to return randomly fluctuating values often located in the high frequency end, which is actually the opposite to what we need. In cases of ambiguous spectra instead, most PD algorithms would simply return meaningless values which once more would have nothing to do with the optimal grid period *P*.



2a

2b

Fig.2 *effect of inadequate choice of the sampling period P either too small or too large, causing under-fitting (2a) or over-fitting (2b) respectively*

It is therefore evident that a novel method for estimating *P* is required; and since, as we have seen, *P* always corresponds to F0 in case of harmonic spectra, the same algorithm is indirectly expected to provide integrated pitch detection (PD) capability as a byproduct of the spectral envelope estimation, a fact which can of course be profited.

In addition, the computed spectral envelope curves shall be reasonably time-stable, i.e. not flicker between similar spectral frames: the similarity between the magnitude spectra must always match the similarity between the respective envelope curves, otherwise audible artifacts will occur with many final applications. One requirement to grant time-stability is avoiding or at least reducing to a minimum any binary or abrupt decision as those involved with any peak-picking scheme or max/min operations, where even a slight change in some secondary spectral detail could influence the result dramatically. A geometrical and where possible also continuous approach has therefore to be preferred, without making any discrete assumptions about the structure of the spectrum itself.

2. Principle overview

Before describing the devised algorithm in detail, we shall introduce three fundamental concepts on which it is intrinsically based

2.1. Envelope curve growth

If we know the optimal value for *P* (in bin units), by progressively applying a bi-directional 1-pole low-pass filter *F* thru the short-time magnitude spectrum *M(w)* such that

$$M'_0(w)=M(w); \quad M'_t(w) = F(M'_{t-1}(w),P)$$

using *P* as cutoff period, followed at every pass by a curve “warping” stage where the growing curve *M'(w)* is “pulled” towards *M(w)* at every point (peaks) w_n in the latter where it results $M'(w_n) < M(w_n)$, the wanted envelope curve is perfectly and nicely fit in few passes (usually 6-7 or even less).
The ratio between the average of the gaps between local maxima

left in M' at every iteration and P is a quite reliable indicator of when to stop the process, to be left with an envelope curve which is never either under- or over-fitting (this won't be, however, the stop criterion adopted by the present algorithm, since P can't be always estimated unambiguously).

A smooth and nice envelope curve passing thru all prominent peaks (prominent according to P) is thus naturally grown at very little computation cost, without resorting to any peak-picking, spline interpolation or even cepstral methods.

The resulting curve fits the magnitude spectrum even more accurately when operating on a perceptual vs. linear scale, like a log scale (taking infinities into account) or a normalized dB scale with linear "toe" from minus infinity to -60 dB, for example.

The problem of computing a spectral envelope reduces therefore to the problem of estimating the optimal value of P for every short-time spectral frame.

2.2. Estimation of P

As a general rule, by progressively smoothing the magnitude spectrum *M(w)*

$$M'_0(w) = M(w);\quad M'_t(w) = F(M'_{t-1}(w), p_t)$$

with an ever increasing filter factor p_t covering a suitable range, and populating at every pass t a cumulative histogram $H(g)$ with the m gaps $g_n = w_n - w_{n-1}$ forming between the resulting local maxima w_n in the smoothing curve $M'(w)$, weighted by their pivotal magnitude $M(w_n)$,

$$H_t(w_n - w_{n-1}) = H_{t-1}(w_n - w_{n-1}) + M(w_n),\quad n=0...m-1$$

the highest peak always *tends* to form at the position in $H(g)$ corresponding to the most signifying spectral spacing (P). The result is *virtually* immune from gross errors (unless in presence of actual sub- or upper harmonic series), can reveal a fundamental F0 even with very low harmonic-to-noise ratio (see Fig.3), and has the desired tendency to grow in the low end in form of a hyperbolic cluster in case of predominating noise.

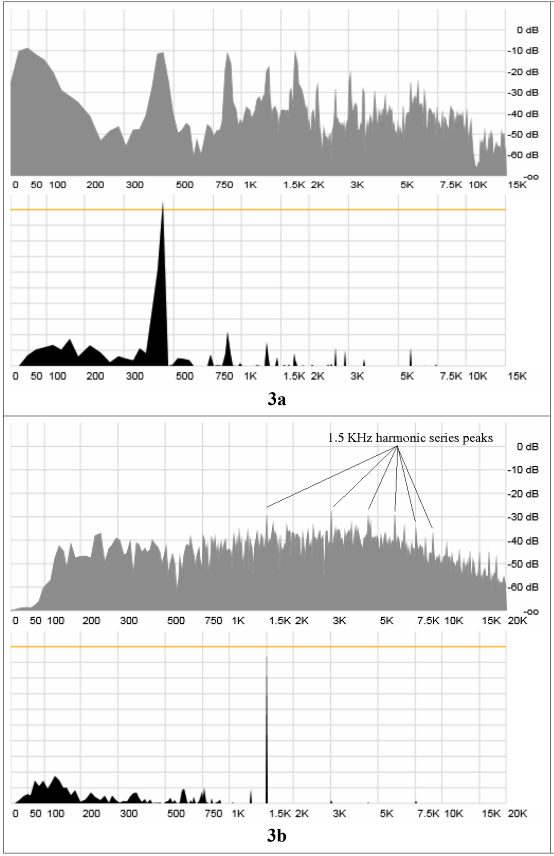


Fig.3 *A couple of spectra with mixed harmonics and noise above the respective plots of H(g). (3a) A breathy letter A (F0 ≈ 400 Hz) with a low-frequency bump (unfiltered) which does not minimally influence the histogram. (3b) A harmonic series with fundamental of 1.5 KHz almost submerged in noise, notice how the periodicity histogram can still resolve a clear peak at 1.5 KHz*

This "magic" happens because at every pass the less prominent peaks get smoothed out and disappear first, while the most prominent ones tend to last for longer.

However, if we don't stop smoothing in time, from some point on "high order" gaps corresponding to local maxima in the very smoothing curve will start building up in the higher histogram end causing confusion.

Since this general principle can really be implemented with infinite variations, its exact bias and robustness against noise and gross errors is highly dependent on the specific implementation details.

For example, many IIR filtering schemes can be adopted to smooth *M(w)* as if it were a time-domain signal: from a simple 1-pole low-pass filter to a maximizing "umbrella" function (either triangular or exponential) to fitting gaussians at every peak covering any peaks below it; what matters is applying the filter bi-directionally to grant horizontal symmetry around peaks in *M'(w)* (symmetrical impulse response). In our case the most effective filter turned out a two-passes "maximizing" 1-pole low-pass filter operating maximization against the original magnitude spectrum in the first pass only:

$$\begin{gathered}\textit{left to right:}\\ M'(w) \rightarrow \max(M(w) + k\,(M'(w-1) - M(w)),\ M(w))\\ w=0...nbins-1\end{gathered}$$

$$\begin{gathered}\textit{right to left:}\\ M'(w) \rightarrow M'(w) + k\,(M'(w+1) - M'(w))\\ w=nbins-1...0\end{gathered}$$

where

$$k = p/(2\pi+p)$$

being p the cutoff period in bin units and *nbins* the number of FFT bins in the half-complex spectrum, that is half the frame size plus one.

This filtering scheme also reduces shifting of local maxima to a minimum extent, which would result in smearing in the histogram otherwise.

2.3. Harmonic considerations on the envelope curve

As a valid criterion to decide whether a magnitude spectrum has been enveloped properly by a fitting curve without leaving room for ambiguity, we can consider the periodicity of the same curve. In general we can expect that a spectral envelope will hardly be periodic; when this is the case, we should rather suspect that any residual periodicity is a clue of an underestimate of P still allowing the harmonic structure to emerge (under-fitting, see Fig. 2a again). This criterion alone *almost* suggests another way of estimating P, i.e. choosing the lowest value of P for smoothing which does not generate any periodicity in the resulting curve any more.

Unfortunately there are actually many real world situations in which an envelope presents instead with a substantially periodic structure, as in case of harmonics with alternating higher and lower magnitudes or even in case of formants accidentally placed on nearly harmonic frequency positions.

Therefore, if P is known, we shall check whether a periodicity of width P is still detectable in the final curve, rather than just a periodicity of *any* width.

To detect periodicity, it will be convenient to resort to a partial histogram (i.e cleared at every iteration) *H'(g)* populated with the per-pass gap counts as described in 2.2 but always incremented by unity, i.e. without scaling by the pivotal gap magnitude, and check the number of entries in proximity of its position *g* corresponding to the supposed *P* candidate. Such number will likely be zero once the curve under investigation does not contain traces of periodicity of width *P* any more (see Fig.4)

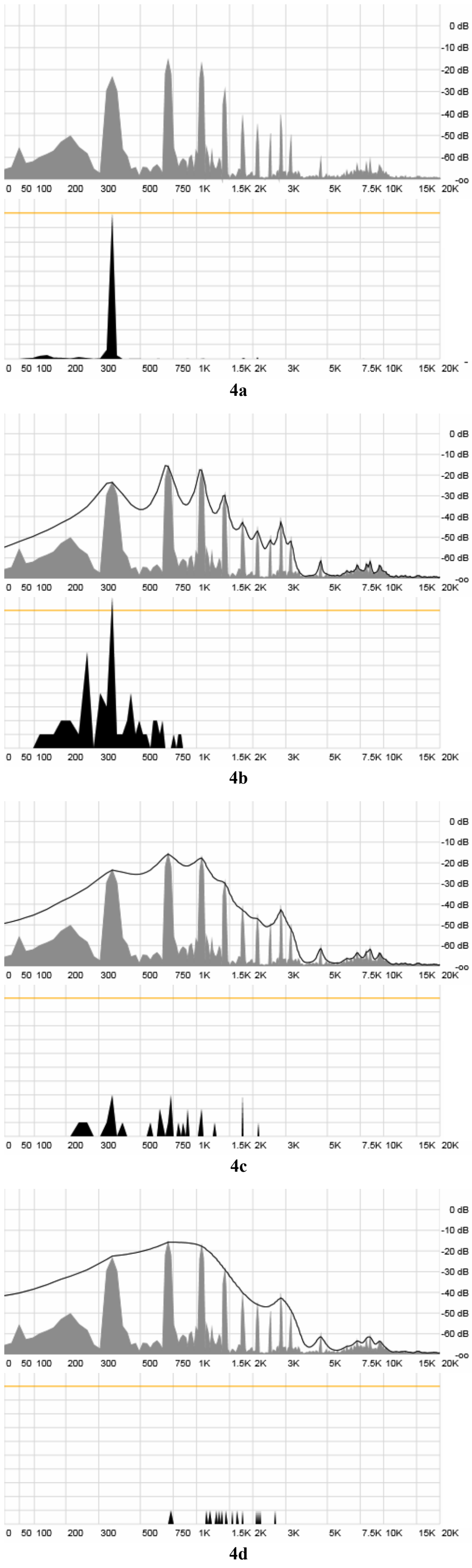


Fig. 4 *Periodicity analysis of the growing envelope curve*

(4a) On top the spectrum of the letter O, F0 ≈350 Hz, sung by a female individual. Below the respective periodicity histogram H(g)

(4b) Growth of the envelope curve at an early pass; below we can examine the plot of H'(g) for the same pass (every grid row represents unity): a high peak (gap count) can be detected at the position of the highest peak in H(g) (P), confirming that there is still a substantial degree of periodicity P in the envelope curve.

(4c) Growth of the same envelope curve at a more advanced stage: the plot of H'(g) below shows that periodicity P is decreased but still present.

(4d) Envelope curve ready. The plot of H'(g) confirms no more periodicity P (i.e around 350 Hz) in the finished curve.

3. The envelope extraction algorithm in detail

Experimentation suggested that there are almost infinite ways to combine the three key principles exposed in 2., but only few combinations produce robust and reliable algorithms which empirically perform well in a wide collection of notable sample cases.

Ideally, the algorithm should be composed of a *P* estimation stage followed by an envelope growth stage. This approach would perhaps result more reliable but at a much higher computation demand, since the magnitude spectrum should be scanned several tens times in total; also by doing so we would be computing a growing curve twice, the first time only as a “probe” curve to be later discarded.

Therefore we decided to adopt a “tandem” approach, where the periodicity histogram is computed while growing the actual envelope curve, and *P* gets determined progressively converging to its supposed value during the process.

Initialization:

-As input a short-time half-complex magnitude spectrum *R(w)* = *|S(w)|* of nominal frame size *z*, 0<=*w*<=*z*/2 is supplied. From the algorithm's perspective, whether the input spectral frames are windowed (by Hann windows for example) or not makes absolutely no difference; care must be taken however to avoid windows producing ringing side-lobes as in case of rectangular zero-padded windows, which could understandably trick the algorithm by introducing false periodic components.

-Two magnitude arrays *M(w)* and *M'(w)* are initialized with the magnitude spectrum *R(w)* converted to a normalized dB scale. In our example we are using a normalized logarithmic scale with a linear “toe” in the initial seventh from -∞ to -60 dB

$$M(w) = M'(w) = \begin{cases} R(w)>=10^{-3} : Log(R(w)) \cdot 2/7 +1 \\ R(w)<10^{-3} : R(w)\cdot 10^{3}/7 \end{cases}$$

-A cumulative, weighted periodicity histogram *H(g)* is cleared

-The filter cutoff period *p* is assigned an initial small value which appears adequate of *z / 512*

Iterations:

-A volatile, un-weighted periodicity histogram *H'(g)* is cleared

-*M'(w)* is progressively smoothed by using a bi-directional 1-pole low-pass filter maximizing against the reference magnitude spectrum *M(w)* in the first pass only:

left to right:

$$M'(w) = max(M'(w) + k\,(M'(w-1)-\ M'(w)),\quad M(w)\,),\quad w=0...z/2$$

right to left:

$$M'(w)= M'(w)+ k\,(M'(w+1)-M'(w)\,),\ \ w=z/2...0$$

where

$$k = p/(2\pi+p)$$

In theory we could avoid the maximization, but by doing so we reduce shifting of local maxima in the progressively smoothed envelope curve, which would cause smearing in the histograms.

It has to be noted that the second pass (from right to left), not performing any maximization against $M(w)$, will naturally cause the growing curve to descend below the level of the prominent peaks (the same would happen if we were using a plain low-pass filter of course). This is something we must allow and which will be compensated in the next operation, in order for the resulting curve to better fit the magnitude spectrum

-$M'(w)$ is "pulled" against $M(w)$ at all points w_m where

$$M(w_m) > M'(w_m),\quad M(w_m) > M(w_m-1),\quad M(w_m) > M(w_m+1)$$

This is achieved by segment-wise multiplication of $M'(w)$ *in* [w_m , $w_{m'}$] by a linearly interpolated factor r:

$$M(w_{m'}) > M'(w_{m'}),\ M(w_{m'})>M(w_{m'}-1),\ M(w_{m'})>M(w_{m'}+1);$$
$$m' > m;$$

$$r(w) = M(w_m)/M'(w_m) +$$
$$(M(w_{m'})/M'(w_{m'}) - M(w_m)/M'(w_m)) \cdot (w-w_m) / (w_{m'}-w_m);$$

$$M'(w) = M'(w) \cdot r(w),\ w= w_m \dots w_{m'}$$

where w_m and $w_{m'}$ are two consecutive local maxima in $M(w)$

-$M'(w)$ is scanned to identify gaps between local maxima. For every gap $g_n= w_n - w_{n-1}$ found, where

$$M'(w_n)>M'(w_n-1),\ M'(w_n)>M'(w_n+1),$$
$$M'(w_{n-1})>M'(w_{n-1}-1),\ M'(w_{n-1})>M'(w_{n-1}+1),$$

$H(g)$ is populated cumulatively, weighting every entry by the gap's pivotal magnitude such that

$$H_t(g_n) = H_{t-1}(g_n) + M(w_n)$$

while $H'(g)$ is populated by just incrementing the respective entries by one unity (plain gap count), such that

$$H'_t(g_n) = H'_t(g_n) +1$$

-The filter cutoff period p is updated as the weighted average of H, which can be thought as its barycenter

$$p = \frac{\sum_{g=0\dots z/2} g \cdot H(g)}{\sum_{g=0\dots z/2} H(g)}$$

This effectively allows p to grow towards the actual optimal unknown value P, and represents a more elastic (continuous) criterion than simply assigning p the position of the highest peak in H. It can easily be seen that in the ideal, unrealistic case of just one peak in H the weighted average corresponds to its position.

-As stated in 2.3, the process must be stopped once no trace of periodicity of order P is detectable in M' any more. P is however something we can't know precisely. We just have a value p growing towards P more or less quickly depending on the spectral structure, but which is meant to be used uniquely as the filter cutoff period, and a periodicity histogram H which is expected, but not guaranteed, to form a substantial peak (when not the highest) at position P. In the ideal case it would be enough to stop the iterations once the value of $H'(f)$ becomes zero or however falls below some safety threshold, where $H(f)$ is the highest peak in H. Here however it is about the mathematics of uncertainty, and any incautious operation could compromise the result. Also, as we have stated previously, similar abrupt decisions must be avoided because potentially detrimental.

A more elastic but equivalent criterion, which was found to work particularly well in pretty all situations, is stopping when the average of H' weighted by H falls below unity. It is evident that in the unrealistic and ideal case where $H(P)$ is the only peak present in H, the criterion reduces to the abrupt decision above.

The process is therefore stopped once

$$\frac{\sum_{g=0\dots z/2} \left(\sum_{i=-n\dots n} H'(g+i)\right) \cdot H(g)}{\sum_{g=0\dots z/2} H(g)} < 1$$

As it can be seen, we are here considering the sum of $2n+1$ values around every value of $H'(g)$ rather than just $H'(g)$, to account for possible histogram smearing caused by rounding errors or peaks shifting during the smoothing process; a value of $n=1$ appears suitable in practice, at least for $1024<=z<=8192$.

When the weighted average above falls below unity (considering possible rounding errors, one may actually want to check if the result is < 0.9999999999), we can judge safely enough that the growing envelope curve in M' does not substantially contain any more traces of periodicity of order P, and the process can be stopped.

The curve in M' can now be converted back to linear scale and returned as the resulting envelope.

4. Pitch detection potential

As we have already mentioned in 1., the capability of the presented algorithm to provide PD functionality has to be considered a nice bonus. Despite in the present work we intend to focus mainly on the envelope estimation aspect, we shall even mention briefly how this intrinsic feature could be exploited and its limitations.

As we have seen, $H(g)$ will contain a highest peak at the position corresponding to F0 (in bin units) in cases of unambiguous harmonic spectra; in case of ambiguous, polyphonic or non-harmonic spectra, the position of the highest peak will however still correspond to an optimal frequency domain period P to downsample the envelope curve. A number of ambiguous cases can show up in the real world though; if we decide to employ $H(g)$ for PD purposes, we can't afford any uncertainty, whereas an imprecise estimate of P (unless completely off) is still acceptable for producing an overall correct envelope curve (within some tolerance). The same of course applies to the standard AC method, where its "raw" (i.e non processed) result intended as the highest peak is consistent with the pitch only in case of non-ambiguous frames.

One can realize immediately that the biggest limitation of the present method is the incapability to detect pure tones, being the algorithm fundamentally based on harmonicity analysis.

After all, from an EE perspective, the algorithm as it is does probably the intended job even in this situation and its behaviour should not be altered: what the envelope of a spectrum constituted substantially by a single peak is expected to look like ? Should it result in a large over-fit of the peak as though it had some invisibly small harmonics, or rather (see Fig. 5i) in a tight fit almost resembling the original magnitude spectrum ? This remains an open question, even if we suspect that the second answer is perhaps the most correct.

In the light of that, the employment of the present algorithm to detect the pitch of pure sinusoids or comparable sounds, including whistles, has to be ruled out from beginning. For processing human voice instead, which is harmonic in its own nature, the chance of estimating both the spectral envelope and the pitch of a spectral frame with a single O(n) algorithm is particularly appealing; one popular example could be pitch correction with formant preservation ("Autotune").

A fundamental requirement for every PD algorithm is the ability to provide, in addition to the pitch estimate, the detection of voiced and un-voiced (VUV) segments, so that in case of segments judged un-voiced the respective pitch value will be simply considered meaningless and ignored (or not computed at all). What to do in such cases is clearly application-specific: for example, a tuning correction software might want to simply return unchanged a segment considered un-voiced, while a voice-driven synthesizer might rather want to interpolate between the pitch values of the segments judged as voiced.

One problem with conventional threshold-based approaches however is that an audio segment may be judged un-voiced not because devoid of any harmonic content but rather because its ambigue structure does not allow to clearly determine its pitch, a superficial choice which could once more lead to audible artifacts with some applications.

In addition, a comprehensive PD algorithm should also provide some clue about the harmonics-to-noise ratio (HNR) of an audio segment: a vocoder aimed at voice reconstruction/re-synthesis, for example, might want to rely on a similar information to decide the ratio of white noise to mix with a pulse train of the detected pitch before applying the formants envelope.

It looks like the content of *H(g)*, if processed correctly, can definitely provide this additional information aswell. For example, suitable post-processing based on the ratio of the energy of the main peak and the backgound peaks, or on the distribution of these, may likely offer clues to a VUV decision together with an HNR index.

We have investigated the potential PD capabilities of the present algorithm only marginally though. From some preliminary tests, the reliability of the raw, un-aided estimate seems comparable to that of the standard AC method, with more robustness against noise even with quite low HNR. Therefore, if we exclude the chance of detecting the pitch of pure tones or alike without resorting to tricks (like adding tiny artificial harmonics), the periodicity histogram can seriously represent a valid substitute.

5. Evaluation

The present work was basically motivated by the need to provide the realtime phase-vocoder-like modular framework *Elena Design's Spectral Modules*[8] for the platform *SynthEdit*[9], developed by the Author of this work, with a robust and reliable envelope extraction plugin. It was inside this same framework where the present algorithm was developed and all the relevant experimentation carried. The final algorithm was therefore coded in highly optimized C++ with GCC 10.2 in shape of an EE module compatible with and tested within said spectral audio processing framework, using various frame sizes, default Hann windowing and an overlap factor of 50%.

Development and tests were carried on a 64-bits digital audio workstation based on Intel i7 8700K directly within SynthEdit environment. A condenser microphone Behringer C3 was used to record the voice samples and without using any soundproof room or bump filters, to better judge the performance of the algorithm on un-optimized, real-world audio material even contaminated by noise.

Since the main goal was achieving robustness across all possible scenarios, we decided not to employ any reference library of audio material for the tests, but rather process un-optimized and often randomly selected segments of speech, singing voices, instruments or even arranged music, including ambiguous spectral frames, noises and secondary details.

CPU load averaged to about 0.2-0.3% with per-core peaks of 2/3% standing to SynthEdit builtin Debug/CPU monitor, using a sample rate of 48 KHz, an overlap factor of 50% and a framesize z=2048; a more precise estimate is not possible though, considering that actual frame processing within the used framework occurs at regular intervals only, every time a new spectral frame is transmitted (that is, about 46 times in a second when using the values above). With longer frame sizes computation increases accordingly, but this compensates with the fact that less frames per second are processed.

The algorithm performed correctly always reproducing a consistent envelope curve for all cases tested and with excellent time stability. Since there is not perhaps any objective criterion to quantify the fidelity of a task like spectral EE other than by naked eye, we have included a comprehensive collection of sample pictures to help judging the results (see Fig. 5a-5o). Only in some cases we could detect some slight overestimate of peaks when they are located in deep "grooves", but this is a limit of the smoothing process involved (only an actual peak-picking scheme would join all prominent peaks perfectly) and it is in our opinion devoid of any serious implications in whatever final application.

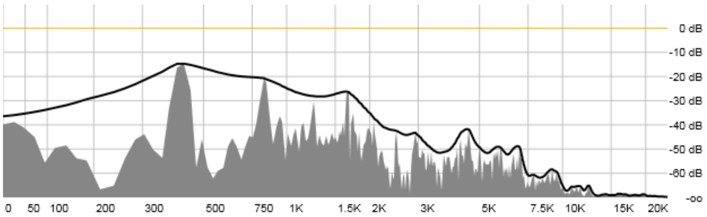


Fig 5a *A breathy vocal "Ah" (z=2048)*

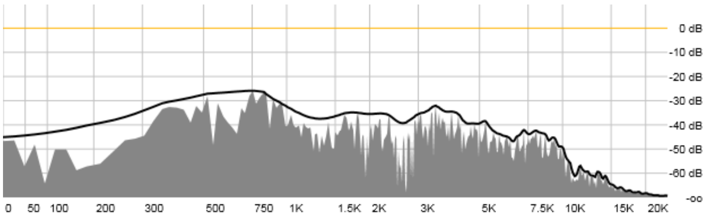


Fig 5b *A hand-clap (z=2048)*

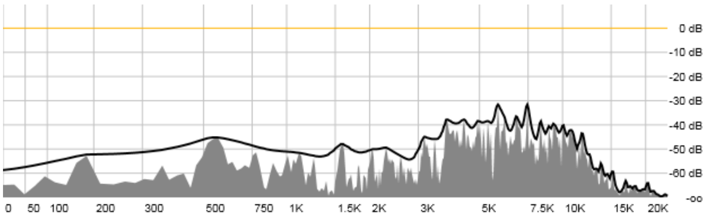


Fig 5c *A crash cymbal (z=2048)*

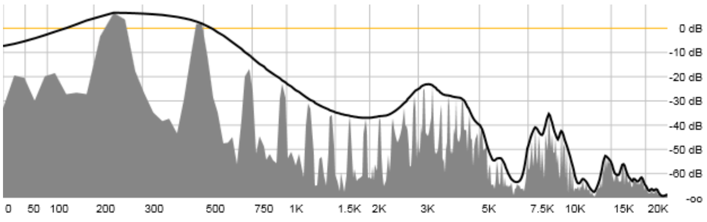


Fig 5d *The vocal I (z=2048)*

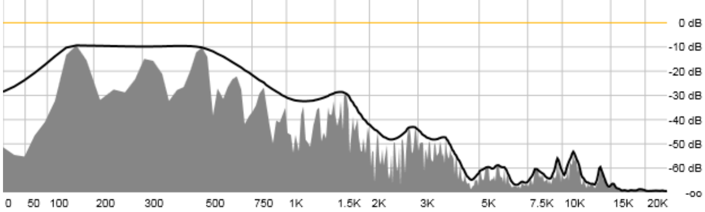


Fig 5e *A random selected, irregular segment of speech (z=2048)*

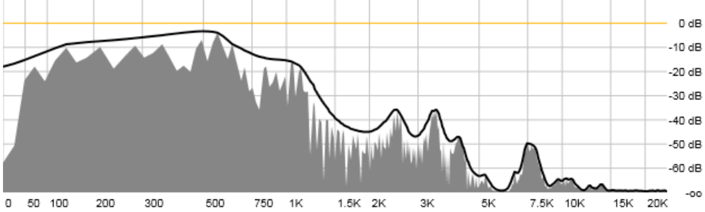


Fig 5f *Another irregular segment of speech (z=2048)*

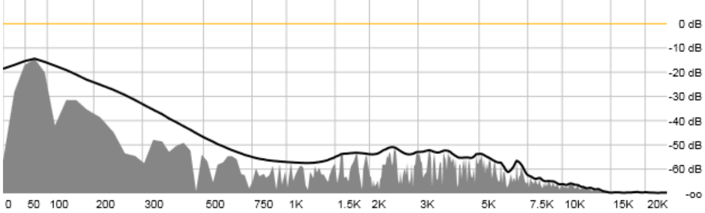


Fig 5g *Attack transient of a kick-drum (z=2048)*

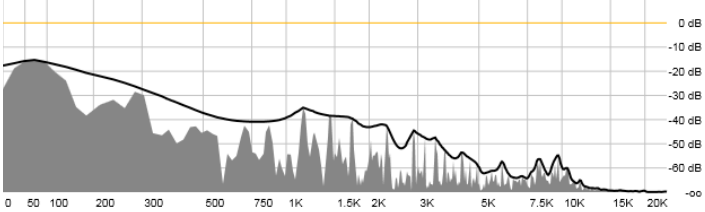


Fig 5h *A randomly picked segment of arranged music (z=2048)*

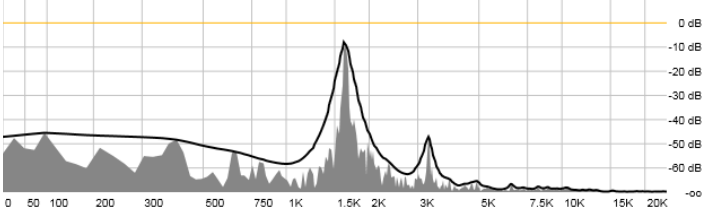


Fig 5i *A human whistle (z=2048)*

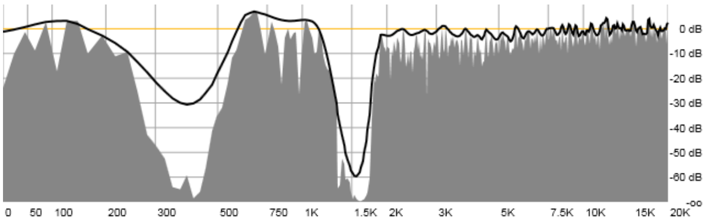


Fig 5j *White noise filtered by a hand-drawn curve (z=2048)*

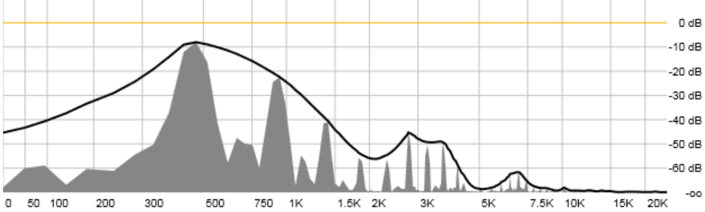


Fig 5k *The letter "U" (z=1024)*

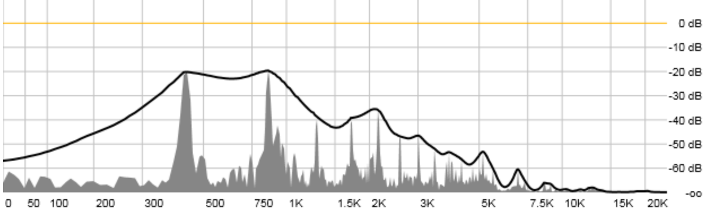


Fig 5l *A breathy "E" (z=4096)*

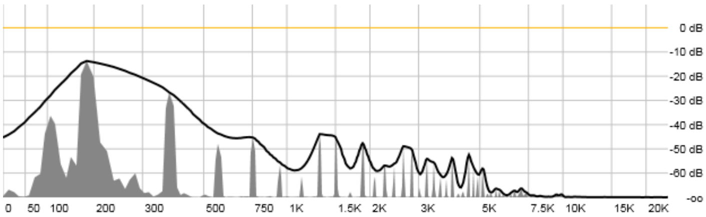


Fig 5m *A guitar note (z=4096)*

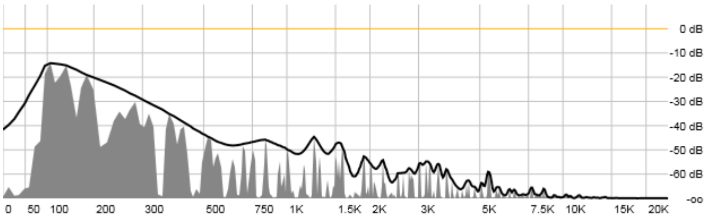


Fig 5n *A guitar chord (z=4096)*

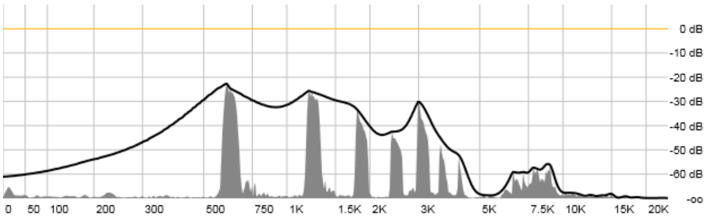


Fig 5o *A high pitched vowel with non-stationary peaks (modulated) (z=8192)*

6. Conclusions

A novel spectral envelope estimation algorithm has just been presented based on some unprecedented ideas, where specifically 1. an optimal spectral sampling period is estimated by means of a periodicity histogram; 2. the envelope curve is grown by simply low-pass filtering the magnitude spectrum as if it were a time-domain signal, constantly pulling the result towards the prominent peaks by a warping stage; 3. the residual periodicity of the growing curve is used as a solid stop criterion.

Its intrinsic adaptiveness makes the algorithm particularly suitable even for cases like inharmonic or noise spectra, where most published alogorithms simply fail to return consistent results, being mostly tailored on harmonic spectra. In particular for formant analysis, the chance to faithfully extract the envelope of unvoiced phonemes is extremely advantageous, being these sampled with a much finer detail than corresponding voiced ones. Also, since the estimated sampling grid is based on an optimal spacing without any reference offset, the algorithm can safely process harmonic spectra shifted in frequency whereas similar approaches based on strict F0 analysis would fail.

The algorithm performed reliably in all real-world tests always returning the expected envelope curves and with perfect time-stability, without ever producing gross over- or under-fitting errors of the magnitude shape, as it often happens instead with many prior art techniques. The pretty negligible CPU demand makes the algorithm perfectly suitable for even complex STFT realtime processing chains.

As we have explained in the introduction, however, no EE algorithm can be perfect given the ambiguous nature of the problem; even in our case we cannot exclude that some limit or deliberately prepared special cases might trick it, despite it proved quite robust empirically.

The three fundamental principles which the algorithm is based on can really be combined in several ways to produce a more or less stable scheme; the one we have chosen and presented as working implementation was just the best combination we could come to, and presently we cannot absolutely exclude that better embodyments may exist.

The pitch detection potential offered as by-product of the harmonicity analysis surely represents an added value, and its full exploitation is matter for future research.

References

[1] Flanagan, J.L and Golden, R.M. (1966) “Phase Vocoder”, Bell System Technical Journal 45 (9): 1493-1509

[2] Lenarczyk, M. “Real time pitch shifting with formant structure preservation using the phase vocoder” - Interspeech 2017: Show & Tell contribution

[3] Hung-Yan Gu, Sung-Feng Tsau (2009) “A Discrete-cepstrum Based Spectrum-envelope Estimation Scheme and Its Example Application of Voice Transformation”, Computational Linguistics and Chinese Language Processing Vol.14 N.4

[4] Roebel, A. and Rodet, X. “Efficient Spectral Envelope Estimation and its Application to pitch shifting and envelope preservation”, International Conference on Digital Audio Effects, Sep 2005, Madrid, pp 30-35; hal-01161334

[5] Villavicencio, F., Roebel, A., Rodet, X. “Improving LPC Spectral Envelope Extraction of Voiced Speech by True Envelope Estimation”, IRCAM

[6] Hong Kook Kim, Hwang Soo Lee, “Use of Spectral Autocorrelation in Spectral Envelope Linear Prediction for Speech recognition”, IEEE Transactions on Speech and Audio Processing, Vol.7 No. 5, Sept. 1999

[7] Liénard, J.S., Barras, C. Signol, F. “Using sets of combs to control pitch estimation errors”, Proceedings of Meetings on Acoustics, Vol.4, 060003 (2008)

[8] Elena Spectral Modules for SynthEdit
https://www.kvraudio.com/product/elena-spectral-modules-for-synthedit-by-elena-design

[9] SynthEdit, visual modular developing environment for virtual sound synthesizers and effects - www.synthedit.com