# A WAVELET-BASED METHOD FOR THE ESTIMATION OF CLARITY OF ATTACK PARAMETERS IN NON-PERCUSSIVE INSTRUMENTS

*Gianpaolo Evangelista*

Independent Researcher
Vienna, Austria
gianpaevan@gmail.com

*Alberto Acquilino* *

Department of Music Research
McGill University, Montreal, Canada
alberto.acquilino@mail.mcgill.ca

## ABSTRACT

From the exploration of databases of instrument sounds to the self-assisted practice of musical instruments, methods for automatically and objectively assessing the quality of musical tones are in high demand. In this paper, we develop a new algorithm for estimating the duration of the attack, with particular attention to wind and bowed string instruments. In fact, for these instruments, the quality of the tones is highly influenced by the attack clarity, for which, together with pitch stability, the attack duration is an indicator often used by teachers by ear. Since the direct estimation of the attack duration from sounds is made difficult by the initial preponderance of the excitation noise, we propose a more robust approach based on the separation of the ensemble of the harmonics from the excitation noise, which is obtained by means of an improved pitch-synchronous wavelet transform. We also define a new parameter, the noise ducking time, which is relevant for detecting the extent of the noise component in the attack. In addition to the exploration of available sound databases, for testing our algorithm, we created an annotated data set in which several problematic sounds are included. Moreover, to check the consistency and robustness of our duration estimates, we applied our algorithm to sets of synthetic sounds with noisy attacks of programmable duration.

## 1. INTRODUCTION

Methods for assessing the tone quality of instrumental sounds are desired in various Music Information Retrieval (MIR) applications. In [1] a model was proposed for the evaluation of the quality of single notes from trumpet, clarinet, and flute by analyzing five sound attributes: dynamic stability, pitch stability, timbre stability, timbre richness, and attack clarity. For several musical instruments, such as wind and bowed strings, the attack phase is a very critical segment of the note that heavily influences the timbral and articulation aspects of the overall produced tone. It is the time interval in which, by exciting the right resonant modes, the noisy excitation gives way to a louder and possibly stable harmonic sound: the transition from pure chaos to ordered chaos.

The characterization of the salient elements of the attack-transient could play a significant role in unassisted practice while learning to play musical instruments. The learners can check the quality of the tones they produce from objective feedback parameters such as attack time and pitch stability. Moreover, accurate descriptors associated with note quality can enhance the search in sound databases for the best or most suitable tones. In this paper, we focus on sound descriptors that are relevant to the automatic assessment of the clarity of the attack.

There is no common consensus on the definition of the attack boundaries. In [2], Luce and Clark defined this as the time from the onset of the note until the sound pressure level reaches 3 dB below the steady-state value. Their empirical approach provided a first framework for measuring attack durations in non-percussive instruments, accounting for variations in pitch, dynamics, and performer's style. However, as also noted in [3], the tones produced by several instruments do not show clear decay and sustain phases. This implies that one cannot rely on the detection of a proper steady-state amplitude of the tones.

Measurements of the duration of the attack reported in [1] made use of methods implemented in the Timbre Toolbox (TT) [3]. In its early releases [4], the duration of the attack is detected by finding the time interval from the onset of the note to the instant in which the maximum amplitude, or a given percentage of it, is attained. To address the limitations of fixed threshold methods, Peeters introduced the *weakest-effort method* [3]. This method, implemented as an option in subsequent releases of the TT, uses highly smoothed versions of the signal envelopes to compute the start and end of the attack based on adaptive thresholds estimated according to the behavior of the signal during the attack phase.

The noise present in the raw envelopes may lead to large errors in positioning the end of the attack phase. However, excessive smoothing of the envelopes results in estimated amplitudes that do not adhere tightly to the signal, which affects the detection of the duration of the attack [5]. In our experiments, we found that both direct envelope thresholding and weakest-effort methods are unreliable for finding the attack durations of sounds with noisy attacks (see Section 4).

Hajda [6] was one of the first researchers to propose a theoretical model that combines spectral and temporal information to better characterize the attack-transient, acknowledging the interplay between these two domains in the perception of musical sounds.

In this paper, we propose a methodology and an algorithm for the accurate measurement of the attack duration in non-percussive instruments based on a peculiar time-scale representation. The idea is to first extract two signals resulting from the separation of the harmonic ensemble, i.e., the signal composed of all the harmonics grouped together, from the blowing or bowing noise, the mixing of which recovers the original signal. The amplitude envelopes of these two signals can be analyzed to detect relevant events.

In order to achieve an accurate noise / harmonic ensemble separation we revisit a method based on the Pitch-Synchronous

* Partial funding for this study was provided by a Doctoral Research Scholarship from the FRQ-NT and by Gary Scavone through the NSERC Discovery Grant (ID: RGPIN-2020-04874)

Wavelet Transform (PSWT) [7, 8] and introduce improvements based on interpolation (upsampling) and period regularization based on local pitch-shifting.

Based on the envelope of the harmonic ensemble, we provide a new definition of the attack time. Furthermore, based on the comparison of the harmonic and transient noise envelopes, we introduce the *noise ducking time* as the time in which the attack noise subsides to the sound of the harmonic ensemble, which is useful for discerning the quality of the attack. Together with the pitch profile, the detected attack and noise ducking times provide sufficient time-frequency cues to determine the clarity of the attack of the produced tones.

Separating the noisy excitation from the resonant component of the sound also has a pedagogical benefit, since the student can be presented with acoustic feedback, which could be crucial to revealing, understanding, and correcting mistakes.

This paper is organized as follows. In Section 2, we analyze possible application scenarios of reliable methods to estimate the duration of the attack. In Section 3, we recall wavelet concepts and multiplexing associated with the PSWT. We also point out improvements in the resolution of the transform based on signal upsampling and period regularization by interpolation. In Section 4, we detail our algorithms for the estimation of the duration of the attack and of the noise ducking time. We also present a consistency and robustness analysis based on synthetic sounds with known characteristics. In Section 5, we describe the acquisition of a data set with quality annotations by an expert, which we used to validate our algorithm with instrumental sounds. In Section 6, we discuss the results obtained from the application of our methods to the attack clarity of the sounds in our and in other available databases. In Section 7, we draw our conclusions.

## 2. CASE STUDY DESCRIPTION AND APPLICATIONS

In this paper, we focus on determining descriptors of the transient attack of isolated monophonic tones produced by wind and bowed string instruments. The significance of this focus lies in the critical role that attack-transient characteristics play in the assessment of the quality of tones [9, 10]. In pedagogical applications, the availability of an automatic objective quality evaluator is bound to help enhance the technical skills and expressive capabilities of the students. In the exploration of musical tones databases, quality features, such as the duration of the attack and the noise ducking time, can improve the search for the best match among the recorded tones.

Wind and bowed instruments exhibit attack transient durations that typically differ from each other [2]. As described in [11], these durations depend on the physics of the resonator itself, which cannot react instantaneously to an excitation; rather, vibrations must gradually build up to reach their full amplitude. This phenomenon is related to the fact that part of the energy provided externally to the resonant system is radiated, while another part is absorbed by the instrument. The attack phase ends when an equilibrium is reached between the input energy and the total of the absorbed and radiated energy, allowing the oscillation to attain its quasi-steady state condition.

Within certain limits, performers can influence the duration of the starting transient. Different types of articulation (e.g., staccato, détaché, martelé) are associated with varying rates of transient development, which musicians can utilize to make stylistic choices in their performances. Learning to control the type and duration of the attack thus becomes an important skill for instrumentalists, enabling them to select the appropriate attack style required by the performance context.

A crucial technical aspect that students should master to express a broad palette of musical ideas is to achieve an attack that is pure, i.e., uncontaminated by noise or unwanted frequency components, and accurate with respect to the desired pitch [12]. It is not uncommon for beginners to make articulation mistakes that produce sounds with excessively long attack transients, generally perceived as unpleasant. For wind instruments, these may include obstacles to the emission of airflow within the oral cavity, such as diction errors, suboptimal tongue positioning, incorrect jaw opening, or an overly constricted throat [13]. For bowed string instruments, errors can involve inadequate bow pressure on strings, irregular bow speed, incorrect bow angle, or uneven bow contact with strings, all of which can disrupt sound production and lead to undesirable articulation [14].

These considerations underline the need for a tool that provides a robust and consistent measure of the attack transient duration, much like how a chromatic tuner is essential for learning to play in tune, together with a measure of the noise extent. Such an educational system would offer teachers greater clarity and objectivity in their instructions and provide learners with objective means to verify their technique during individual practice sessions. For example, an instructor might indicate: "*For the next lesson, try to play the C4 note with an attack duration shorter than 40 ms with piano, mezzo-forte, and forte dynamics*."

Previous studies have attempted to address this need. The results of the ML-based model in [1] seem to hint that the features that are best related to the clarity of the attack of trumpet and clarinet tones were tonal descriptors deriving from the pitch, while a temporal property, the duration of the attack, scored the best for the case of flute tones. However, some mistakes, e.g. pitch instability during the attack, can be included in data sets more often than other ones, as blurred or breathy attacks. Thus, the distribution of various types of playing mistakes in the training data set may well bias the final score. Moreover, the attack duration estimator used in [1] may not be adequate for the analysis of noise-driven sounds, as demonstrated by the example in Fig. 5 in Section 4 and by other examples or use of the software contained in the companion page to this paper [15].

In the next section, we start our journey to discuss a new method for the estimation of attack characteristics based on transient / harmonic ensemble separation.

## 3. NOISE + HARMONIC ENSEMBLE DECOMPOSITION

The excitation noise in wind or bowed instruments is wideband, whereas, when a steady tone is reached, most of the energy concentrates in narrow bands centered on harmonic frequencies. A simple idea to improve the attack duration estimators is to isolate the resonant signal from the noise. Intuitively, this can be realized by designing two comb filters, one peaking on the harmonic frequencies and the other one notching these frequencies. Thus, the output of one of the filters is the signal composed of the harmonic ensemble and the output of the other is the noisy component. Clearly, by filtering with the notch comb, we introduce tiny holes in the spectrum of the noisy component. However, for our purposes, this spectral alteration of the noise is not critical to the listening experience or to the extraction of relevant attack parameters. Moreover, in our algorithm for the estimation of the duration of the attack described in Section 4, we need to detect the time at which a full

resonance develops, which happens at the end of the attack phase. To do so, after the note onset we detect the amplitude envelopes of the separated signals and check when the harmonic component starts overwhelming the noisy component and when it reaches a percentage, e.g. $-3$ dB, of its maximum level.

While comb filters were our basic inspiration, the scheme based on wavelets that we revisit in this section has many advantages. In the first place, being realized with multirate filter banks, it features a very efficient implementation of high-order comb filters. Moreover, the whole separation procedure is structured in a series expansion over a complete and orthogonal set of functions that does not introduce energy bias. As we shall see, the bandwidth of the comb filters is controlled by the number of scales we use in the wavelet transform. Next, we recall basic concepts about wavelets and outline a comb extension of wavelets.

The Wavelet Transform (WT) [16, 17] is a time-scale representation of signals which is equivalent to a time-frequency representation on a logarithmic frequency axis. It is mostly useful for separating transients at several time scales from the average behavior of signals. Properly sampling in the time-scale plane, one can arrive at a class of complete and orthogonal sets of wavelets in $L^2(\mathbb{R})$, which are suitable for the wavelet series expansion of any finite energy signal $s(t)$:

$$s(t) = \sum_{n=1}^{\infty} \sum_{m=-\infty}^{+\infty} a_{n,m} \psi_{n,m}(t), \tag{1}$$

where

$$a_{n,m} = < s, \psi_{n,m} > = \int_{-\infty}^{+\infty} s(t) \psi^*_{n,m}(t) dt \tag{2}$$

are the wavelet expansion coefficients and $<,>$ denotes the scalar product in $L^2(\mathbb{R})$.

In its canonical form, the wavelet decomposition achieves segregation of constant or nearly constant components from fluctuations from the constant behavior. This is realized by means of a generalized sum (average) and differences (innovations) encoding scheme based on a multirate pruned tree of Quadrature Mirror Filters (QMF) leading to band-pass wavelets with band allocation similar to that of a graphic equalizer (e.g., fractional octave bands). For the simplest case of octave band (dyadic) wavelets, one has

$$\psi_{n,m}(t) = 2^{-n/2} \psi(2^{-n} t - m), \; n \in \mathbb{N}, \; m \in \mathbb{Z} \tag{3}$$

where $\psi(t) = \psi_{0,0}(t)$ is the *mother wavelet* and, due to their roles, the indices $n$ and $m$ are respectively called the *scale index* and the *time-shift index*.

In the construction of the wavelet sets one can show the existence of a low-pass function $\phi(t) \in L^2(\mathbb{R})$, called the *scaling function*, which in our context can be useful to express the residue of a scale-truncated wavelet expansion:

$$s(t) = s_f(t) + s_h(t) \tag{4}$$

where

$$s_f(t) = \sum_{n=1}^{N} \sum_{m=-\infty}^{+\infty} a_{n,m} \psi_{n,m}(t) \tag{5}$$

is the scale-truncated wavelet expansion and

$$s_h(t) = \sum_{k=-\infty}^{+\infty} b_{N,k} \phi_{N,k}(t) \tag{6}$$

is the *scaling residue*, where

$$b_{N,k} = < s, \phi_{N,k} > = \int_{-\infty}^{+\infty} s(t) \phi^*_{N,k}(t) dt \tag{7}$$

are the *scaling coefficients*, with

$$\phi_{N,k}(t) = 2^{-N/2} \phi(2^{-N} t - k), \; k \in \mathbb{Z} \tag{8}$$

In the canonical wavelet expansion, the signal $s_h(t)$ in (6) represents the quasi-constant trend – nearly DC level or deep low-pass – while $s_f(t)$ in (5) represents the fluctuations from the quasi-constant behavior up to scale index $N$. However, with unchanged form but different wavelets, we are going to change the rules of the game here.

In fact, for the representation of pitched-tones it is certainly more useful to segregate the periodic or quasi-periodic trend from the fluctuations over the periodic trend. In order to achieve that, we need a modification of the wavelets, which actually results from a different computational scheme. For convenience, in our discussion we switch to discrete-time wavelets and signals, equipped with the scalar product in $\ell^2(\mathbb{Z})$.

If the time period $P$ of the signal is constant, a winning idea is to arrange all periods in the columns of a matrix, as shown in Fig. 1, and then compute a canonical wavelet transform along each of the rows [8]. If the signal were perfectly periodic, then all the columns of the matrix would be identical, so that each row signal would be constant. Thus, the band-pass wavelets would not play a role in this case, leading to zero wavelet expansion coefficients. However, if the signal is not exactly periodic, then the wavelets will represent all deviations from the periodic behavior at several time scales.

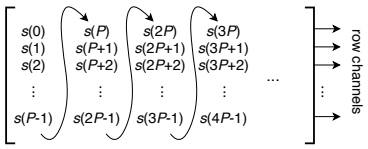



Figure 1: The construction of the matrix of the periods of a discrete-time pseudo-periodic signal $s(n)$ and the forming of the row channels to be represented by means of canonical wavelets.

From the flow point of view, the formation of the matrix is equivalent to demultiplexing the original signal to $P$ channels corresponding to the rows of the matrix. Equivalently, one can define the multiplexed wavelets [8], which already incorporate the multiplexing operations and enjoy the same formal structure as in (1 - 8) but different physical interpretation. In fact, the DTFT $\hat{\Phi}(f)$ of the discrete-time multiplexed scaling function is related to the DTFT $\Phi(f)$ by $P$-fold shrinking:

$$\hat{\Phi}(f) = \Phi(Pf) \tag{9}$$

Since the scaling function of the canonical wavelets is low-pass, due to the periodicity of the DTFT the scaling function for the multiplexed wavelets is a comb covering the harmonics of the signal. Given the sampling rate $f_s$, the bandwidth $BW_N^{tooth}$ of each tooth of the harmonic comb at scale level $N$ is

$$BW_N^{tooth} = \frac{f_s}{2^N P} \tag{10}$$

which can become very narrow as $N$ increases. As shown in Fig. 2, the multiplexed wavelets are also comb-shaped, but their teeth peak on sets of sidebands of the harmonics. These sidebands become narrower and closer to the harmonics as the scale index $n$ grows.

Scale-truncation of the multiplexed-wavelet expansion achieves the separation of the noisy excitation – the signal $s_f(t)$ in (5) – from the resonant part or harmonic trend – the signal $s_h(t)$ in (6), which is required for our attack duration estimator and for the presentation of the acoustic feedback of the excitation for pedagogical purposes.

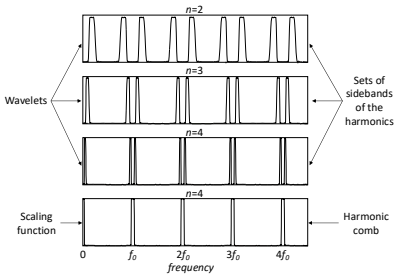



Figure 2: Magnitude Fourier Transforms of scaling function and comb wavelets at several scales.

Things become a little wilder when the local period of the signal is not constant and one truly needs to compute a Pitch-Synchronous Wavelet Transform (PSWT) with variable pitch. Two main modes were presented in [7], in which the shorter periods were either zero- or constant-padded to form a matrix whose columns are the size of a pre-assigned or calculated maximum period. The extra samples are subsequently deleted in the final reconstruction.

### 3.1. Improvement of the PSWT

In this work, we successfully experimented with a new technique, which is essentially based on time warping. Prior to multiplexed wavelet analysis, which can then be carried out with constant pitch, we stretch each period to a maximum period using interpolation based on polyphase anti-aliasing filters [18]. In the synthesis, we decimate the periods back to their original lengths, again with polyphase anti-aliasing filters. Since in the synthesis we separate the wavelet contribution (noisy component) from the scaling residue, it is necessary to perform period decimation separately on these two signals.

Another improvement to the PSWT based method that we carried out in our experimentation is to up-sample the signal prior to pitch detection and wavelet analysis. Since we based pitch detection on a sliding-window autocorrelation method, the estimated period is an integer approximation of the true period. Up-sampling has a mitigating effect on the quantization of the period estimate, which makes the separation of the noisy components of the signal from the resonant part much more accurate. In fact, in the frequency domain, the scaling function forms a comb tuned to the pitch of the tone that is supposed to trap all the harmonics. In case of mistuning, the higher harmonics could fall out of the harmonic comb and end up in the territory of the wavelets, i.e., in sidebands of the harmonics, thus contributing to the fluctuations component, which is an undesired behavior.

In general, the number of scales $N$ at which one truncates the wavelet analysis is also limited by mistuning: at lower $N$ the teeth of the comb are less narrow so they are more keen to cover the harmonics, but this also means that more energy from the noisy component would be covered by the scaling residue and not by the wavelets. Therefore, more accurate tuning achieves deeper analysis and better segregation of the components.

Upsampling the signal by factor 10 adds a decimal point to the resolution of the period estimate. Here again we used polyphase filter interpolation. We were able to achieve great segregation pushing the number of scales to 4-5 for most sounds in our data set and in other public databases.

It must be pointed out that, while vibrato can and must be tolerated, erratic pitch variations as in the sounds typically produced by beginners are considered to be mistakes which, besides being detected by the pitch instability indicator, can be heard in the acoustic presentation of the noisy signal sound. Since a pitch detection and tracking module is embedded in our PSWT-based attack duration estimator, the pitch profile is easily displayed and pitch instability measures, such as the pitch STD, are easily computed, once we remove the outliers [1]. We use these features to complement our attack duration estimate in determining the clarity of attacks.

The pitch estimation we used in conjunction with the PSWT is period synchronous, where a detection frequency range is preset. A window of length equal to 2-3 maximum periods is sliding on the signal by an amount equal to the last detected period. When no pitch is detected, the maximum frequency, corresponding to the minimum period, is outputted as an outlier "pitch" of the current signal segment. Therefore, pure noise samples are arranged into short segments that are then stretched by interpolation to the maximum period length $P$, before ending up in columns of the demultiplexing matrix 1. However, since the samples of adjacent noise segments greatly differ from each other and from subsequent pitched periods of the signal, they are mostly picked up by the row-channel wavelets and do not contribute to the row-channel scaling residue, which is in line with our separation idea.

A block diagram of the complete procedure to extract the noisy and resonant parts of the signal is shown in Fig. 3. Sound examples of noise-resonance segregation in various tones of natural instruments and synthetic sounds can be found in [15].

### 3.2. Complexity

The multiplexed wavelet analysis-synthesis block has linear complexity in terms of the number of samples and, in principle, can be computed in real time using FIR QMF filters. Clearly, up-sampling and period interpolation increase the complexity factor and introduce further latency. In our foreseen applications, either as feedback for the student musician or in database quality indexing, real-time is not a strict requirement. Our off-line interpreted Matlab implementation running on a basic M2 ARM CPU laptop with 8GB RAM, including signal up-sampling/down-sampling factor of 4, autocorrelation-based pitch detection, pitch regularization, and the computation of 5 multiplexed analysis/synthesis wavelet scales rooted on order 9 Daubechies' QMF filters [16], runs slightly faster than real time, within a time factor of 0.875. The system lags behind real-time when the up-sampling/down-sampling factor is increased. For reference, when increasing this factor to 10 the computation in Matlab requires double the time required by real-time operation.

## 4. ATTACK DURATION ESTIMATION ALGORITHM

The accurate estimation of the duration of the attack transients in musical sounds, particularly for noise-driven harmonic instruments such as wind and bowed string instruments, presents a significant

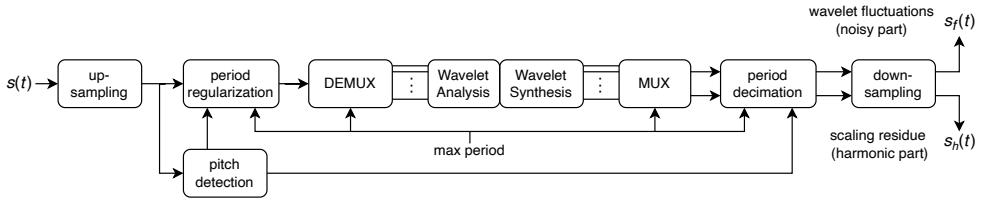



Figure 3: Block diagram of the PSWT-based separation of noisy and resonant parts.

challenge due to the presence of the excitation noise, which, in some cases, overshoots above the steady-state oscillation. For these instruments, the sound can be characterized as pseudoperiodic, exhibiting a clear harmonic structure only once the note is fully developed.

In this section, we propose a novel algorithm that leverages the pseudoperiodic nature of harmonic instruments to reliably estimate the duration of the attack transient. The suggested algorithm is designed for isolated monophonic pseudoharmonic sounds, which are characterized by three distinct segments as follows: initial silence, noisy attack transient, and fully developed sound in steady state. Estimation of the duration of the attack can take great advantage of the separation of the noisy component from the harmonics, which simplifies the detection of the onset of periodic behavior, i.e., the end of the attack phase. The proposed method relies on the separation of signals based on the PSWT and its improvements described in Section 3.

We detect the amplitude envelopes of the two signals, $e_h(t)$ for the harmonic content of the sound and $e_f(t)$ for the noisy fluctuations associated with the attack transient. We experimented with various methods to extract the envelopes and to interpolate them and we found that the classical sliding-window maximum method with linear or spline interpolation gives the best results for its adherence to the signal dynamics, when the window length is tuned to approximately one period of the signal.

We detect the onset time $t_{on}$ of the tone as the instant when the amplitude of the input signal lies for the first time above a threshold $A_{thr}$. The minimum useful threshold level depends on the Signal-to-Noise Ratio (SNR) of the recording and is estimated, with a margin, from the recording of the silence preceding the note. Depending on the instrument and play mode, during the attack transient the amplitude associated with the fluctuations can be significantly higher than that of the harmonic signal.

At the end of the attack phase, the amplitude envelope of the harmonic ensemble attains higher levels. We detect the attack offset $t_{off}$ when the level of the harmonic signal reaches a fraction $\alpha$ of its maximum value. In other words, given the envelope $e_h(t)$ of the harmonic ensemble signal $s_h(t)$ and the input signal $s_{in}(t)$, we have:

$$\begin{aligned} t_{on} &= \min_t\{t : |s_{in}(t)| > A_{thr}\} \\ t_{off} &= \min_t\{t : e_h(t) \geq \alpha r\} \\ dW &= t_{off} - t_{on} \end{aligned} \tag{11}$$

where $r = \max_t e_h(t)$ and $dW$ is the wavelet-based estimation of the attack duration time. A block diagram showing the computation flow for the estimate of the duration of the attack is shown in Fig. 4. In most of our experiments, we let $\alpha = 10^{-3/20} \approx 70.8\%$ which yields a 3 dB attenuation, but, in specific applications, $\alpha$ can be considered as a free calibration parameter.

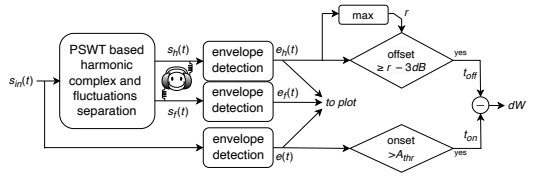



Figure 4: Block diagram of the attack duration estimator $dW$ by means of PSWT based separation.

An example of attack duration estimate that illustrates the advantage of applying our PSWT-based method is shown in Fig. 5. There, a trumpet sound signal is plotted in which the initial excitation noise peak is higher than the steady-state amplitude. The classical method, also included in the early versions of TT [4], defines the attack time as the duration of the interval from $t_{on}$ until the instant in which the maximum amplitude is reached. This criterion clearly fails for the signal in Fig. 5 since the maximum of the signal envelope occurs at the very beginning and is purely due to excitation noise during the attack phase.

In Fig. 5, superimposed on the input signal are the estimates $e_h(t)$ (red curve) and $e_f(t)$ (yellow curve) of the envelopes for the harmonic and noisy components, respectively. The envelope $e_h(t)$ correctly ignores the initial noisy transient and reaches the maximum roughly when the steady-state part of the sound begins. By thresholding $e_h(t)$, the duration of the attack was correctly estimated at 182 ms, which makes more sense than the estimates for $t_{off}$ provided by thresholding the original envelope ('x' mark in the figure) and by means of the weakest effort method ('o' mark in the figure), both of which occur when the attack is still in the noisy transitory part.

The detected $t_{off}$ is only slightly larger than that of typical well-rated attacks ($\leq$160 ms). However, an additional quantity derived from the separated signals can help in assessing the clarity of the attack: the *noise ducking time* $t_{nd}$. We define this as the instant at which the initial attack noise starts to be overtaken by the harmonic components. A strategy for evaluating $t_{nd}$ that works for a large class of tones is to detect the maximum of $e_f(t)$ and then find the first subsequent instant where $e_f(t)$ falls below $e_h(t)$ by a prescribed amount, which we set at 3 dB in our experiments. If the largest peak of the noisy component is not prominent, i.e., if it falls below a prescribed threshold, which we fixed in our experiments at 15 dB below the maximum of $e_h(t)$, then we set $t_{nd}$ as the instant in which $e_h(t)$ reaches an amplitude that is 3 dB above $e_f(t)$. The justification of this conditional approach is to prevent that false detection of $t_{nd}$ is triggered at the very onset of the signal even when the noise amplitude later reaches a high level.

In some sounds with badly rated attacks, such as growling wind

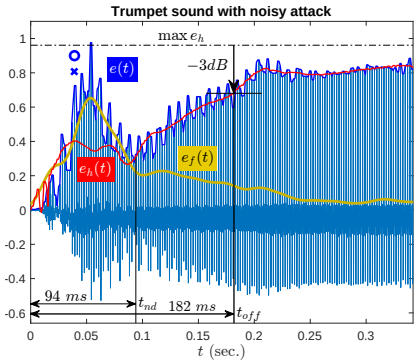



Figure 5: Trumpet sound signal, time-shifted so that $t_{on} = 0$. The estimate of the attack duration $t_{off}$ obtained by thresholding the envelope $e_h(t)$ (red curve) of the harmonic part is shown by a thick black vertical line. The estimates of $t_{off}$ obtained by means of direct thresholding (3 dB below $\max_t e(t)$) and the weakest-effort methods are also shown, respectively, by means of a cross (x) and a circle (o) mark above the input signal envelope $e(t)$. The envelope $e_f(t)$ (yellow curve) of the fluctuation components is also shown, which peaks right after the onset of the signal. An estimate of the noise ducking time $t_{nd}$ is shown, which delimits the end of the noisy part of the attack.

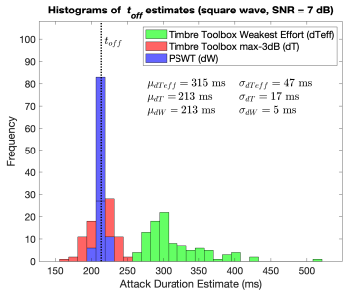



Figure 6: Histograms of the estimates of the duration of the attack of a noise-corrupted square wave synthetic sound, where dW denotes the estimate using the PSWT-based method, dT the estimate by means of input signal envelope thresholding, and dTeff by means of the weakest-effort method.

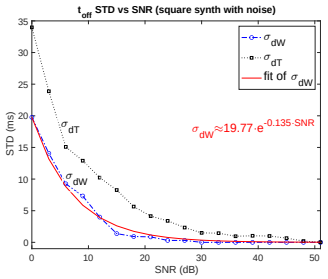



Figure 7: Behavior of the standard deviations $\sigma_{dW}$ and $\sigma_{dT}$ for measures of the duration of the attack based on wavelets and on input envelope, respectively, using square wave synthesis and noise at increasing SNR. The curve described by a decreasing exponential fit of $\sigma_{dW}$ is also plotted.

sounds or string tones played with wrong bow pressure, the initial attack is actually short and clean, but a noisy phase is initiated immediately after it. In such cases, the value of $t_{off}$ is not decisive, but large $t_{nd}$ allows us to detect a prolonged noisy activity. The $t_{nd}$ detected for the signal in Fig. 5 is 94 ms, while clean attacks show much shorter noise ducking times ($\approx$ 20 ms). Further examples of attack analysis and data tables can be found in [15].

### 4.1. Consistency and Robustness

In order to test the consistency and robustness of the PSWT-based estimate of the attack duration time and compare our method with existing ones, we generated and analyzed synthetic signals: sinusoids, band-limited square, sawtooth, and triangular waves together with a trumpet-like sound obtained from its first 10 Fourier coefficients. All sounds were corrupted by time-enveloped Gaussian random noise. In order to simplify our analysis, we used trapezoidal envelope shapes for both noise and signals, where the envelope of the noise largely covers the attack phase of the signal. In our tests, we set several values of the Signal-to-Noise Ratios (SNR), defined as $20 \log_{10}$ of the ratio between the signal level and the noise level in the flat and overlapping part of the envelopes. We also included the possibility to introduce vibrato by frequency-modulating the waves.

Given that the envelopes that we impose are programmable, it is easy to assess the duration of the attack phase from them. In order to estimate statistics of the measurement, we fed the estimation algorithms – PSWT-based, input envelope thresholding, and weakest-effort methods – sets of 100 test sounds of the same wave type but corrupted by different samples of statistically independent, equally amplitude enveloped, white noise. For each method, we plot the histogram and extract the mean $\mu$ and the standard deviation $\sigma$ of the measured samples of $t_{off}$, an example of which is shown in Fig. 6. Due to the way it is defined, it is natural that the average estimates of the duration time obtained by the weakest-effort method may differ from those of the other two methods, but its high standard deviation is of concern, which shows that the method is not very robust; as such, we will not consider it further.

Example behaviors of the standard deviation $\sigma_{dW}$ as SNR grows for the wavelet-based method and $\sigma_{dT}$ for the input envelope thresholding are reported in Fig. 7. We can see that the standard deviation of the PSWT-based method, which exponentially decreases as the SNR increases, is always much smaller than that of the classical direct thresholdung method.

To test the consistency and robustness of the estimation algorithm for $t_{nd}$, we performed statistical tests using synthetic sounds in which the noise envelope is a short linearly or exponentially decaying pulse, which overlaps with the attack of the trapezoidal wave envelope. Across all our synthesizers and various predetermined noise ducking times, the relative standard deviation (RSTD), i.e. the STD divided by the mean, always resulted less than 10% and the mean remained within a few milliseconds of the programmed value for $t_{nd}$.

Further examples of estimated statistics using synthetic sounds with and without vibrato can be found in [15].

## 5. DATA SET

To evaluate the use of the proposed algorithm for the analysis of instrumental sounds, a specific data set of isolated monophonic trumpet tones was recorded using high-end audio equipment in a

soundproof booth to minimize ambient noise and external interference. The microphone was positioned 50 cm in front of the bell of the trumpet, aligned at the same height and facing the instrument. This placement remained constant throughout the sessions to ensure consistency in the recordings.

The performer was a professional musician with a degree in music performance and a professional background in music education. The musician played isolated tones throughout the primary range of the trumpet, specifically targeting the notes B♭3, D4, F4, B♭4, D5, and F5. These pitches were chosen to cover a representative spectrum of the instrument range. For each selected note, the performer was instructed to play multiple tones, exhibiting both good and poor attack clarity. The poor attack-clarity sounds were intended to simulate common articulation errors made by novice players. No specific dynamic levels were imposed. After recording, the musician provided annotations for each selected sample, focusing on four main characteristics associated with poor attack-clarity:

- *Noisy attack*: The attack contains noticeable noise, perceived as a prolonged crack-like sound at the onset, like the sound in Fig. 5, which is taken from our data set.
- *Delayed stabilization of pitch*: The onset begins on a different harmonic than the intended pitch before settling into the intended note, resulting in distinct transient sounds depending on whether the onset starts on a higher or lower harmonic. Although playable notes in the harmonic series of trumpet resonances are equally spaced in frequency, this issue is more prevalent in the high register since the corresponding distance in cents or fractions of half-tones becomes smaller as the series ascends, requiring higher onset pitch precision.
- *Delayed stabilization of resonance*: The sound starts muffled and unstable before reaching a more resonant timbre, creating a characteristic "ti-OH" effect in the attack discussed in the pedagogical literature [13].
- *Attack with breath noise*: Despite tonguing, the sound does not start immediately; instead, there is an audible breathing noise as air passes through the instrument before the vibration begins uncontrollably late.

Our data set [15] comprises 149 labeled sounds, each annotated according to the identified attributes. It is important to note that individual recordings may exhibit more than one of these characteristics simultaneously. Although limited in size, the data set was instrumental in the development of the attack duration estimation algorithm described in Section 4, especially useful to attribute a physical meaning to the attack phase, which is absent from other definitions devised for generic signals.

## 6. RESULTS AND DISCUSSIONS

In this section, we analyze the performance of the proposed attack duration estimation method across different types of attack transients of the collected trumpet data set. The results are discussed in terms of the estimated attack duration and noise ducking time, as these are found to be the salient features.

*Good attacks* generally show a harmonic envelope that increases quite rapidly and linearly until it reaches a flatter region. The onset of the note is characterized by a short peak of the envelope of fluctuations, likely due to the tongued attack, before the envelope then sets to lower levels. The estimated $t_{nd}$ is very small as the harmonic envelope soon prevails. Depending on the slope of the attack, the estimate of $t_{off}$ can reach a range of values that are generally smaller than in attacks of lower clarity.

In *noisy attacks*, the estimated $t_{off}$ is generally only slightly higher than in cleaner articulations. Here, $t_{nd}$ emerges as the main discriminant, showing values significantly higher than in cleaner articulations. As illustrated in Fig. 5, the separation of the noisy excitation from the harmonic ensemble prevents false detections of the termination of the attack, which are induced by transient noise peaks in the original signal envelope. This is a net improvement over the TT detection methods.

In *attacks with delayed stabilization of resonance*, the harmonic envelope exhibits an initial lower amplitude, increasing until the sound stabilizes. The algorithm accurately reflects this transition, associating these cases with larger estimates of both $t_{off}$ and $t_{nd}$. Since the transient development occurs within a much shorter time scale, this behavior is distinct from a deliberate crescendo.

In *attacks with breath noise*, an onset detection algorithm based on a dynamic threshold estimated during silence ensures that the breath noise is not misclassified as background noise. Both $t_{off}$ and $t_{nd}$ are generally higher in this case.

Since the harmonic envelope may temporarily stabilize on an unintended pitch, the *attacks with delayed stabilization of pitch* do not necessarily correspond to larger $t_{off}$ and/or $t_{nd}$. This type of attack error is perceptually salient and can be easily identified by observing large standard deviations of pitch.

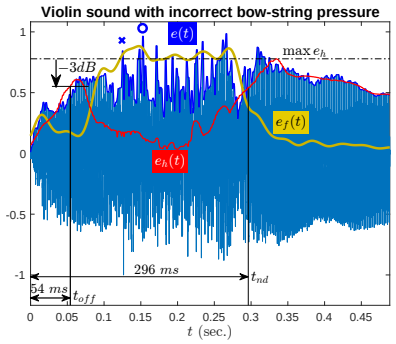
We also tested our algorithms on a wider set of recordings using the Good-sounds data set [19], which includes a collection of isolated tones of wind and bowed string instruments with a substantial number of partially annotated examples of correctly played notes and notes with attack errors. Among the annotated errors, some instances included brief descriptions of the type of attack issue, while others were generically labeled "bad attack".

Unfortunately, we could not find a systematic classification of attack mistakes across different instruments in the literature. However, we suggest that, based on the temporal evolution of the noise and of the harmonic ensemble rather than on the physical mechanism of sound production itself, the classification developed for trumpet attacks could be extended to other instruments. For example, a violin sound with a noisy attack due to incorrect bow pressure (see Fig. 8a) is physically distinct from a noisy attack of a trumpet caused by improper embouchure articulation. However, both exhibit similar behavior in terms of the interaction between the noise and the harmonic components. The analysis of a flute sound with a breathy attack is shown in Fig. 8b, which shows high values of characteristic times. Further examples shown in [15] for wind and bowed string instruments illustrate how the proposed attack duration estimation and noise ducking time have a broader applicability to describe the transient attack characteristics of various categories of musical instruments.

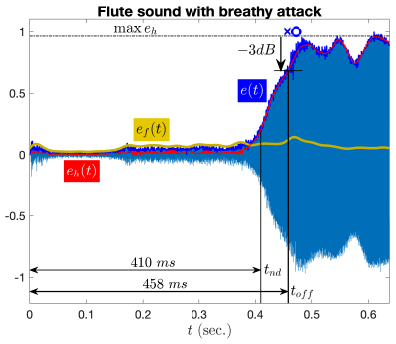
Despite variations in absolute attack times across different instruments, our algorithm consistently produce values for two parameters, $t_{off}$ and $t_{nd}$, which allow us to distinguish properly executed attacks from faulty ones. The results confirm that the integration of both temporal and spectral information is essential for accurately analyzing transient behaviors in instrumental sounds.

## 7. CONCLUSIONS

In this paper, we introduce a new method to estimate the duration of the attack of nonpercussive instruments for which, due to the




(a)




(b)

Figure 8: (a): Analysis of a violin sound with noise produced by the incorrect pressure of the bow on the string, showing a large value of $t_{nd}$; (b): Analysis of a flute sound with a breathy attack, showing a large value of both $t_{off}$ and $t_{nd}$.

excitation bow or blow noise, the classical direct estimate and the weakest-effort method are not sufficiently robust. Our method is based on an excitation/resonance separation by means of an improved PSWT. The consistency and robustness of our proposed algorithm were checked by statistical trials conducted on synthetic sounds. Qualitative checks and musical interpretation could be performed in the specially created data set and other available databases. The uses in database indexing for tone-quality related queries and in self-assisted music practice were pointed out.

Further work will extend our data set and interpretation to a broader class of instruments with annotations by experts. Furthermore, since the attack times may vary for each harmonic of the tone, we will explore the use of the Harmonic-Band Wavelet Transform (HBWT), essentially a PSWT where multiplexing is replaced by a Discrete-Cosine Transform (DCT) [20].

## 8. REFERENCES

[1] O. Romani Picas, H. Parra Rodriguez, D. Dabiri, H. Tokuda, W. Hariya, K. Oishi, and X. Serra, "A real-time system for measuring sound goodness in instrumental sounds," in *Proc. of AES 138th Conv.*, 2015, pp. 1–11.

[2] 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.

[3] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams, "The timbre toolbox: Extracting audio descriptors from musical signals," *J. Acoust. Soc. Am.*, vol. 130, pp. 2902–2916, 2011.

[4] S. Kazazis, P. Depalle, and S. McAdams, *The Timbre Toolbox User's Manual*, 2022.

[5] K. Nymoen, A. Danielsen, and J. London, "Validating attack phase descriptors obtained by the timbre toolbox and mirtoolbox," in *Proc. of the SMC Conf.*, 2017, pp. 214–219.

[6] J. Hajda, "A new model for segmenting the envelope of musical signals: the relative salience of steady state versus attack, revisited," *J. Audio Eng. Soc.*, vol. 101, 1996.

[7] G. Evangelista, "Pitch synchronous wavelet representations of speech and music signals," *IEEE Trans. on Signal Processing*, vol. 41, no. 12, pp. 3313–3330, 1993.

[8] G. Evangelista, "Comb and multiplexed wavelet transforms and their applications to signal processing," *IEEE Trans. on Signal Processing*, vol. 42, no. 2, pp. 292–303, 1994.

[9] K. Guettler and A. Askenfelt, "Acceptance limits for the duration of pre-Helmholtz transients in bowed string attacks," *J. Acoust. Soc. Am.*, vol. 101, no. 5, pp. 2903–2913, 1997.

[10] M. Pàmies-Vilà, A. Hofmann, and V. Chatziioannou, "The influence of the vocal tract on the attack transients in clarinet playing," *Journal of New Music Research*, vol. 49, no. 2, pp. 126–135, 2020.

[11] J. Meyer and U. Hansen, *Acoustics and the Performance of Music: Manual for Acousticians, Audio Engineers, Musicians, Architects and Musical Instruments Makers*, Springer, New York, USA, 2009.

[12] A. Acquilino and G. P. Scavone, "Current state and future directions of technologies for music instrument pedagogy," *Front. in Psychology*, vol. 13, Mar. 2022.

[13] A. Jacobs and B. Nelson, *Also Sprach Arnold Jacobs: A Developmental Guide for Brass Wind Musicians*, Polymnia Press, Mindelheim, Germany, 2006.

[14] K. Guettler, *The Science of String Instruments*, chapter Bows, Strings, and Bowing, pp. 279–299, Springer New York, New York, NY, 2010.

[15] A. Acquilino and G. Evangelista, "Sound Examples companion page," `https://attackdurationestimator.github.io/DAFx25`.

[16] I. Daubechies, *Ten lectures on wavelets*, Society for Industrial and Applied Mathematics, USA, 1992.

[17] S. Mallat, *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*, Academic Press, USA, 3rd edition, 2008.

[18] P. P. Vaidyanathan, "Multirate digital filters, filter banks, polyphase networks, and applications: a tutorial," *Proc. of the IEEE*, vol. 78, no. 1, pp. 56–93, 1990.

[19] G. Bandiera, O. Romani Picas, H. Tokuda, W. Hariya, K. Oishi, and X. Serra, "Good-sounds.org: a framework to explore goodness in instrumental sounds," in *Proc. of ISMIR Conf.*, Aug 2016, pp. 188–91.

[20] P. Polotti and G. Evangelista, "Fractal Additive Synthesis: a Deterministic/Stochastic Model for Sound Synthesis by Analysis," *IEEE Signal Processing Magazine*, vol. 24, no. 2, pp. 105–115, 2007.