Audio Engineering Society

# Convention Paper 10280

Presented at the 147th Convention
2019 October 16–19, New York, USA

# Precise Temporal Localization of Sudden Onsets in Audio Signals Using the Wavelet Approach

Yuxuan Wan[1], Yijia Chen[1], Keegan Yi Hang Sim[1], Lijia Wu[1], Xianzheng Geng[1], and Kevin Chau[1]

[1] *The Hong Kong University of Science and Technology, Hong Kong, China*

Correspondence should be addressed to Yuxuan Wan (ywanaf@connect.ust.hk) or Kevin Chau (eekchau@ust.hk)

## ABSTRACT

Presently reported is a wavelet-based method for the temporal localization of sudden onsets in audio signals with sub-millisecond precision. The method only requires $O(n)$ operations, which is highly efficient. The entire audio signal can be processed as a whole without the need to be broken down into individual windowed overlapping blocks. It can also be processed in a streaming mode compatible with real-time processing. In comparison with time-domain and frequency-domain methods, the wavelet-based method proposed here offers several distinct advantages in sudden onset detection, temporal localization accuracy and computational cost, which may therefore find broad applications in audio signal processing and music information retrieval.

## 1 Introduction

Need often arises for the determination of the exact time (with sub-millisecond precision or even within a few audio samples) when a sudden onset or burst occurs in an audio signal. This could help pinpoint specific events, e.g., gunshot, explosion, etc., that signal an emergency. In music, sudden onsets could be the result of percussion instruments. Knowing when these occur with a high precision is often useful in note, beat or tempo extraction, track synchronization, music learning, and more generally, music information retrieval (MIR). It also facilitates subsequent audio analysis and processing, e.g., audio repairs like click or cough removal, artifact-free audio compression, computational auditory scene analysis (CASA), source separation, and voice or sound triggered actions.

In general, a sudden onset is characterized by an impulsive burst in the audio signal followed by a ringing that decays over time. Various methods are available for sudden onset detection. In the time domain, the instantaneous signal power can be monitored. However, the peaks associated with the sudden onsets are often masked by the high-amplitude harmonic signals that are present, thus resulting in many false positives and negatives. In the frequency domain, the short time Fourier transform (STFT), which is well suited for stationary audio signals, suffers the well-documented time-frequency trade-off governed by the uncertainty principle. Hence the temporal resolution for sudden onsets, which is determined by the window size, is often low and inflexible in the STFT.

Wavelets, due to their highly localized transient nature, are particularly suitable for the detection of sharp transitions or contrast in a signal. As orthogonal basis functions, the multi-resolution capability of wavelets via scaling and translation, their rich frequency content, and limited time extent all contribute to creating an often more accurate and compact representation of a non-stationary transient signal [1-3]. Given these unique advantages especially for edge detection, wavelets have already been widely employed in image processing, e.g., JPEG image and fingerprint compression [4-5]. In comparison, the usage of wavelets in audio signal processing is not yet as pervasive. Nonetheless, numerous applications have been proposed, e.g., in audio denoising, compression and fingerprinting [6-8].

## 2 Principle of Operation

Using orthogonal wavelets as basis, an audio signal can be broken down into, and hence represented by its constituent wavelets. In practice, rather than dealing directly with the wavelet functions, the decomposition is implemented as a filter bank operation [9]. In the discrete wavelet transform (DWT), filter coefficients are defined for a pair of low-pass and high-pass decomposition filters, two examples of which are shown in Fig. 1. The audio signal is split into two halves by these two filters, followed by ↓2 downsampling. The filtered and decimated low-pass and high-pass outputs are known as the approximation coefficients (cA) and detail coefficients (cD), respectively. The entire process of splitting and decimation can be applied to cA, cD, and their offspring successively using the same pair of filters, eventually generating a wavelet packet tree as shown in Fig. 2 [10]. Each level of decomposition offers a different perspective in time and scale/frequency. The net result is a multi-resolution wavelet representation of the audio signal. It can be seen from Fig. 2 that the first level cA and cD offer the highest temporal resolution besides the original audio signal. By a proper choice of wavelet, it is possible that the first level decomposition can yield a set of cD coefficients that retains the relevant high frequency information associated with sudden onsets while rejecting the nonrelevant low frequency information. Since cD is a filtered output of the audio signal, the time information of the onset events is preserved except for the smearing caused by the filter length. More importantly, the filter bank implementation of the wavelet transform enables real-time processing, which is not readily achievable with the windowed overlapping blocks in the STFT.

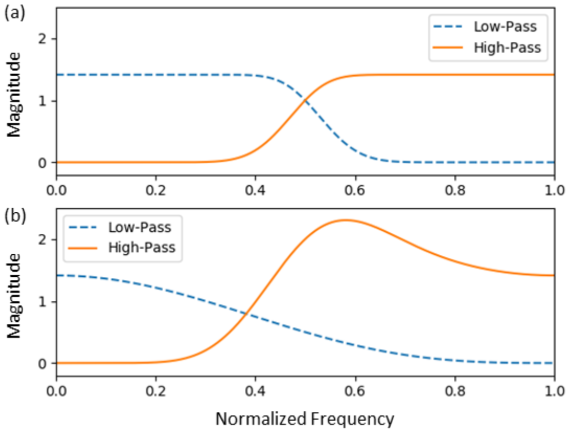

Figure 1. Frequency responses of the decomposition filters of the (a) coif10 and (b) rbio3.9 wavelets.

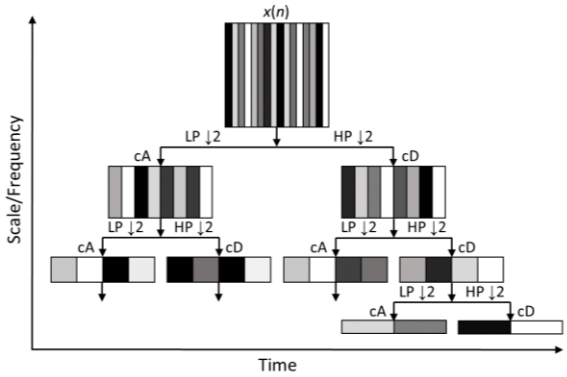

Figure 2. Successive dyadic decomposition of the audio signal $x(n)$ into a wavelet packet tree via low-pass (LP) and high-pass (HP) filtering followed by downsampling, achieving multi-resolution in time and scale/frequency with each additional level of decomposition.

## 3 Implementation

The proposed wavelet-based onset detection can be implemented in various ways. For batch mode processing, DWT can be applied to the entire audio file to extract the first-level cD. For real-time processing, the audio signal is convolved with the filter coefficients of the high-pass decomposition filter of the wavelet to yield an undecimated version of cD. It is crucial to select, e.g., empirically, a wavelet that produces the best result in singling out

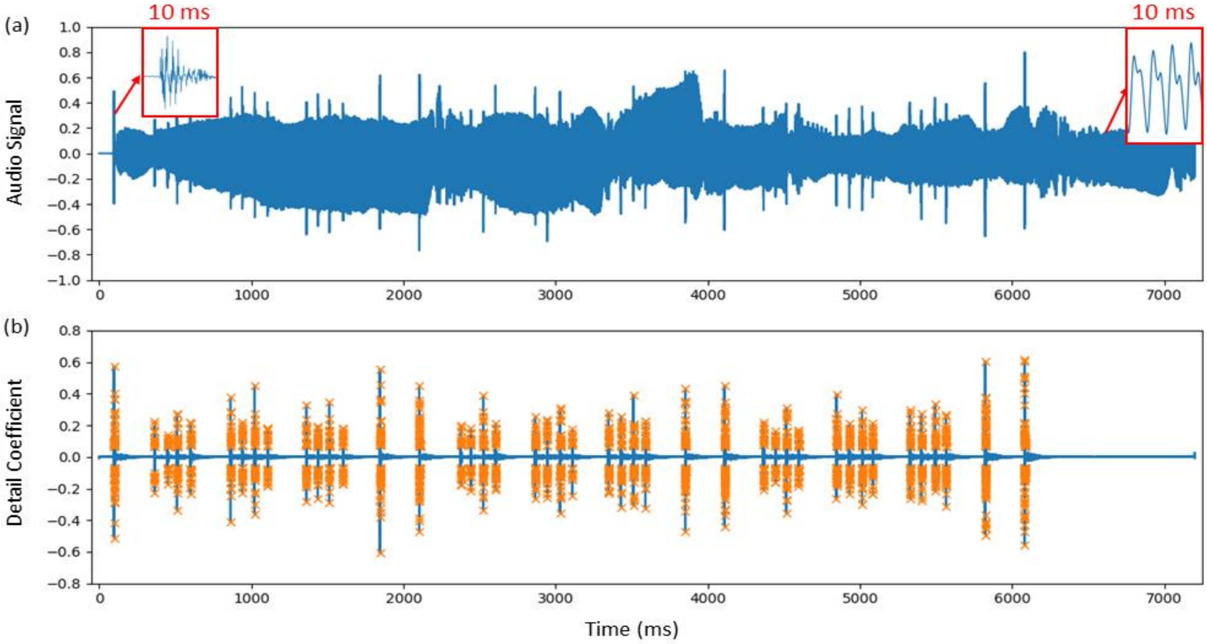

Figure 3. (a) Audio signal consisting of a castanet sequence mixed with horn. The insets show a magnified waveform of the castanets (left) and horn (right). (b) First level detail coefficients extracted with the discrete wavelet transform using the rbio3.9 wavelet. The detected castanet occurrences are marked by “×”.

the sudden onset information. If the first level cD fails to achieve that then higher levels of decomposition and the resulting cA and cD further down the wavelet packet tree should be considered. In any case, the implementation will take $O(n)$ operations, where $n$ is the length of the audio signal. Thresholding is finally applied to the cD or cA that best captures the onsets.

## 4 Results

The proposed wavelet-based sudden onset detection method was evaluated on 4 audio data sets, the results of which are described below.

*(1) Castanets Mixed with Horn*

An audio mix of castanets and horn, each obtained from the SQAM recordings [11], is shown in Fig. 3a. The sounding of the castanets is characterized by a sharp impulse of a few ms, followed by a low-level ringing that lasts over 100 ms. This is in sharp contrast to the smooth and harmonic horn sound. Using the rbio3.9 wavelet, the first level cD shows an almost complete rejection of the horn as shown in Fig. 3b. Then by hard thresholding, each of the castanet occurrences was flagged repeatedly resulting in no false positives and negatives. The cD peak detection method demonstrates an improved signal-to-background ratio (SBR) of 50 dB over that in the original audio mix and a temporal localization of the castanet onsets to within 10 audio samples (0.2 ms).

*(2) Rock Music*

The onset detection method was used as a beat finder on “One” by Metallica — a heavy metal rock song with strong percussion [12], an excerpt of which is shown in Fig. 4. The excerpt starts with regular beats that give way to 3 sets of rapid machine gun like sounds toward the end. In the audio waveform, the percussion is embedded in the electric guitar (Fig. 4a). However, both the fast and slow beats are nicely unveiled in the first level cD obtained with the coif10 wavelet (Fig. 4b). An accurate determination of the beat type and onset time is therefore possible by setting an appropriate threshold along with cluster analysis.

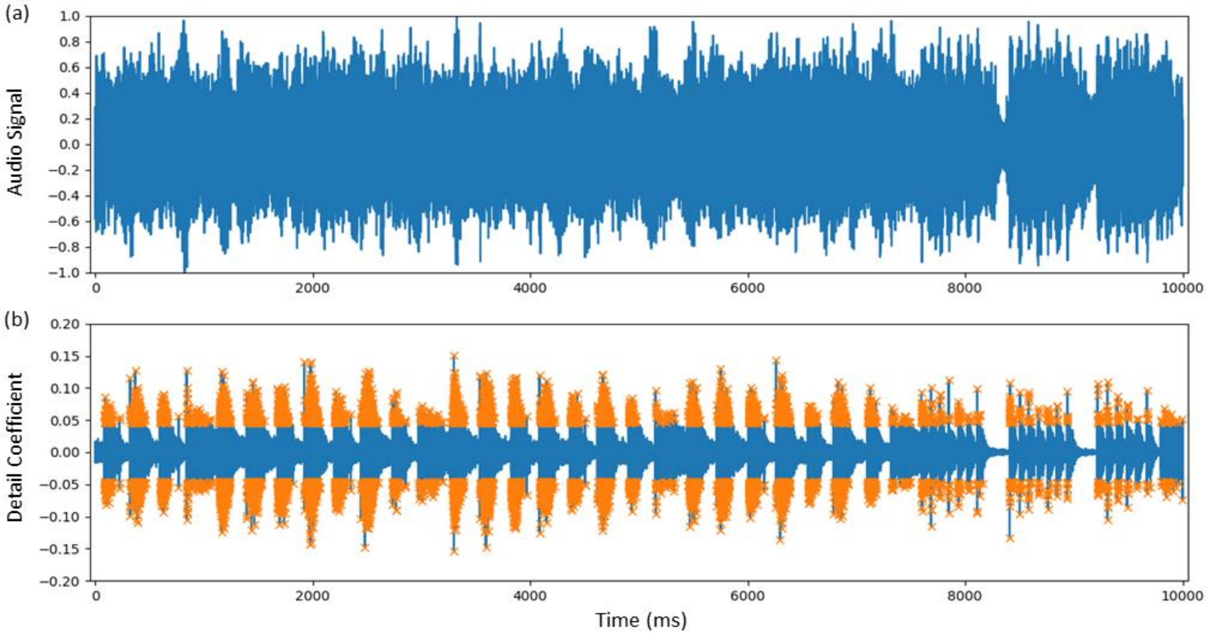

Figure 4. (a) An excerpt of the heavy metal rock song "One" by Metallica. (b) First level detail coefficients extracted with the discrete wavelet transform using the coif10 wavelet. The detected beats are marked by "×".

*(3) Gunshot in a Party*

An audio mix of party conversation, gunshot, and subsequent screaming using sound samples from Freesound [13] is shown in Fig. 5. The party conversation and screaming are almost completely rejected in the first level cD using the db4 wavelet, leaving just the gunshot for detection. The method should be equally applicable to a close-by explosion. In either case, the threshold should be set high enough to discriminate against non-emergency events like heavy door slamming or television drama.

*(4) Honking, Tire Skidding and Collision*

The final example is an audio mix of honking, tire skidding followed by the collision of an automobile using sound samples from Freesound [14]. As shown in Fig. 6, the wavelet-based method is not able to achieve a clean separation between the honking and tire skidding. Other detection methods, e.g., on the characteristic pitch, perhaps should be considered. For the collision part, the second level cD appears to offer a better detection than first level cD by using the coif10 wavelet.

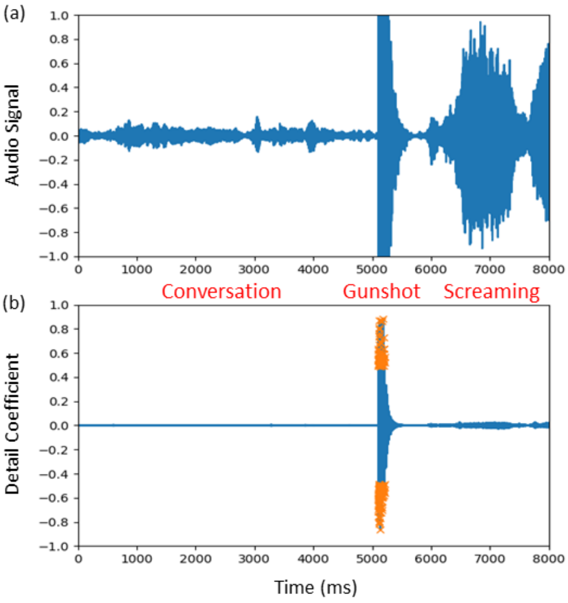

Figure 5. (a) Audio mix of a gunshot in a party. (b) First level detail coefficients extracted with the discrete wavelet transform using the db4 wavelet. The detected gunshot is marked by "×".

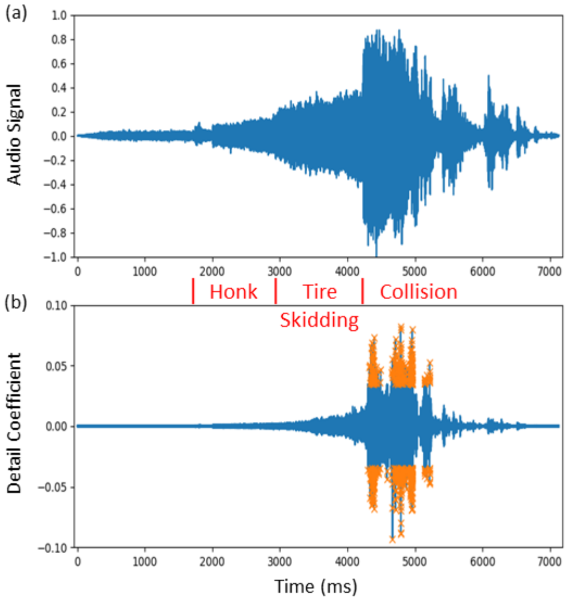

Figure 6. (a) Audio mix of honking, tire skidding and the collision of an automobile. (b) Second level detail coefficients extracted with the discrete wavelet transform using the coif10 wavelet. The detected collision peaks are marked by “×”.

## 5 Discussion

The proposed wavelet-based sudden onset detection method was evaluated with two musical and two non-musical examples. A clean separation of the sudden onsets from the background audio was obtained in 3 out of the 4 examples. The best result is achieved when the onsets to be detected are much more abrupt than the background audio. The choice of wavelet is critical as different wavelets can produce drastically different results. The task of finding the right wavelet, performed empirically in this work, is a current topic for research [15-16]. If successful, this should yield a compact wavelet representation and a good separation, and the overall sudden onset detection and temporal localization should surpass that from standard filtering without using wavelets. The separated onset events can be extracted from one of the filtered outputs (cA or cD) in the multi-resolution wavelet packet tree, and a reconstruction of the portion of the audio signal based on the chosen cA or cD is not necessary. Even though hard thresholding was employed in the examples given, more sophisticated thresholding and statistical analysis are available for the onset detection [17].

The temporal resolution of the proposed method is closely related to the sharpness of the onset. For the sharpest onset, i.e., a delta function, and by using the Haar wavelet with a filter length of two, the ultimate temporal localization can be as good as two audio samples, or 0.04 ms at a sampling rate of 48 kHz. This resolution is about two orders of magnitude better than that obtained by STFT and other spectral methods, which is therefore ideal for audio repairs like click removal.

## 6 Conclusions

A wavelet-based method is proposed for the detection and temporal localization of sudden onsets in audio signals. The method requires only $O(n)$ operations and is fully compatible with real-time processing. For the sharpest onsets, the temporal localization can achieve sub-millisecond precision by using a short enough wavelet filter. The wavelet-based method was evaluated with musical and non-musical examples. The results demonstrate the distinct advantages of the method in sudden onset detection, temporal localization accuracy and computational cost. The proposed method may therefore find broad applications in audio signal processing and music information retrieval.

## Acknowledgment

The authors would like to thank the Undergraduate Research Opportunities Program (UROP) of The Hong Kong University of Science and Technology for supporting this work.

## References

[1] S. Jaffard, Y. Meyer, and R. D. Ryan, *Wavelets: Tools for Science & Technology*, SIAM (2001).

[2] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, Academic Press, 3rd ed. (2008).

[3] I. Daubechies, *Ten Lectures on Wavelets*, SIAM (1992).

[4] A. Skodras, C. Christopoulos, and T. Ebrahimi, “The JPEG 2000 Still Image Compression Standard,” *IEEE Signal Processing Magazine*, pp. 36—58 (2001).

[5] G. A. Khuwaja and A. S. Tolba, “Fingerprint Image Compression,” *Proc. IEEE Signal Processing Society Workshop*, pp. 517—526 (2000).

[6] G. Yu, E. Bacry, and S. Mallat, “Audio Signal Denoising with Complex Wavelets and Adaptive Block Attenuation,” *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing* (2007).

[7] R. C. Guido et al., “A Study of the Best Wavelet for Audio Compression,” *Proc. 40th Asilomar Conf. Signals, Systems and Computers*, pp. 2115—2118 (2006).

[8] S. Baluja and M. Covell, “Content Fingerprinting Using Wavelets,” *Proc. 3rd European Conf. Visual Media Production*, pp. 198—207 (2006).

[9] R. C. Guido, “A Note on a Practical Relationship Between Filter Coefficients and Scaling and Wavelet Functions of Discrete Wavelet Transforms,” *App. Math. Lett.* 24, pp. 1257—1259 (2011).

[10] F. Bömers, *Wavelets in Real Time Digital Audio Processing: Analysis and Sample Implementations*, Master’s thesis, Univ. Mannheim (2000).

[11] *Sound Quality Assessment Material (SQAM) Recordings for Subjective Tests*, © EBU (2008).

[12] J. Hetfield et al., *One*, © Metallica (1988).

[13] https://freesound.org/people/Carmelomike/sounds/255215/ (2014).
https://freesound.org/people/FreqMan/sounds/23154/ (2006).

[14] https://freesound.org/people/YleArkisto/sounds/270297/ (2015).

[15] W. K. Ngui, M. S. Leong, L. M. Hee, and A. M. Abdelrhman, “Wavelet Analysis: Mother Wavelet Section Methods,” *Appl. Mechanics and Materials*, vol. 393, pp. 953—958 (2013).

[16] J. Rafiee, P. W. Tse, A. Harifi, and M. H. Sadeghi, “A Novel Technique for Selecting Mother Wavelet Function Using an Intelligent Fault Diagnosis System,” *Expert Systems with Applications*, vol. 36, pp. 4862—4875 (2009).

[17] S. G. Chang, B. Yu, and M. Vetterli, “Adaptive Wavelet Thresholding for Image Denoising and Compression,” *IEEE Trans. Image Processing*, vol. 9, pp. 1532—1546 (2000).