Audio Engineering Society

Convention Paper 7541

Presented at the 125th Convention
2008 October 2–5 San Francisco, CA, USA

The Effects of Lossy Audio Encoding on Onset Detection Tasks

Kurt Jacobson[1], Matthew Davies[1], and Mark Sandler[1]

[1] *Centre for Digital Music, Queen Mary University of London, Mile End Road, London, E1 4NS, UK*

Correspondence should be addressed to Kurt Jacobson (kurt.jacobson@elec.qmul.ac.uk)

ABSTRACT

In large audio collections, it is common to store audio content with perceptual encoding. However, encoding parameters may vary from collection to collection or even within a collection - using different bit rates, sample rates, codecs, etc. We evaluate the effect of various audio encodings on the onset detection task. We show that audio-based onset detection methods are surprisingly robust in the presence of MP3 encoded audio. Statistically significant changes in onset detection accuracy only occur at bit-rates lower than 32kbps.

1. INTRODUCTION

There is a large body of research related to audio-based music information retrieval methods. The research community has developed a variety of techniques [1, 2, 3] and standards [4, 5] for performing content-based retrieval using digital signal analysis. However, applying these methods in practical systems means dealing with a variety of real-world issues including scalability, usability and audio encoding quality. Here we are concerned with the issue of digital audio encoding quality and the subsequent effect on music information retrieval tasks.

In practical systems, perceptual encoding of audio content is nearly ubiquitous. To reduce storage and bandwidth requirements while maintaining a relatively high standard of audio quality large stores of music content are usually encoded using an MP3 codec or a similar perceptual codec . However, audio encoding formats are not uniform across different collections. For example, Myspace[1] hosts audio content encoded at 96kbps MP3 while iTunes[2] provides content encoded as 128kbps AAC. Even within one music collection, audio encodings maybe heterogeneous. This is often the case for an individual's personal music collection where the audio content may have been collected from a variety of disparate sources.

[1] http://myspace.com

[2] http://www.apple.com/itunes/

In this paper we address the onset detection task – automatically detecting note onsets that are musically salient in an audio signal. Onset detection is an important task in music information retrieval research as it is a pre-processing step for many other tasks including rhythmic analysis [6], bar boundary detection [7], segmentation [8], and music similarity [9]. To perform onset detection we use some recent spectral and energy-based methods [10, 11], each of which is described in section 2.

To determine how these methods are effected by lossy audio encoding we re-encode an existing corpus of audio content at different bit-rates. We use the LAME MP3 codec[3] to create five data sets encoded as PCM, 128kbps / 44.1kHz MP3, 96kbps / 32kHz MP3, 32kbps / 16kHz MP3, and 8kbps / 16kHz MP3 respectively. The corpus of audio content is hand-annotated with accurate 'ground truth' onset locations and was first used in the MIREX 2005 onset detection competition [12].

2. ONSET DETECTION METHODS

Most automatic approaches to note onset detection make use of a two-stage process. The first is the generation of a mid-level feature representation, commonly referred to as an *onset detection function*. This signal exhibits local maxima at likely onset locations. The subsequent stage of analysis is to extract the temporal locations of these peaks using a *peak-picking* algorithm. A comprehensive review of methods for generating onset detection functions can be found in [10, 11]. For our experiment we focus on the derivation of five conceptually simple onset detection functions Γ_x, where x is used to identify each function, these are:

- Γ_{E}: Energy based [10]
- Γ_{H}: High Frequency Content [13]
- Γ_{S}: Spectral Difference [14]
- Γ_{P}: Phase Deviation [15]
- Γ_{C}: Complex Spectral Difference [16]

The detection function $\Gamma_x(m)$ with samples m is generated from analysis of one or more short term frames of length N from an input signal $s(n)$ with sampling rate $f_{\mathbf{s}}$ Hz. To create a 50% temporal overlap between frames (as in [10, 11]) we use a *hop size*, $h = N/2$. To maintain a fixed time resolution t=11.6ms [11] for each detection function sample m, we set $h = f_{\mathbf{s}} \times t$. For an audio signal $s(n)$ sampled at $f_{\mathbf{s}} = 44.1$kHz this equates to a frame size of $N = 1024$ with hop size $h = 512$ audio samples.

2.1. Energy based

The energy based onset detection function $\Gamma_{\mathrm{E}}(m)$ measures signal energy within each short term analysis frame:

$$\Gamma_{\mathrm{E}}(m) = \sum_{n=1}^{N} s(mh - n). \tag{1}$$

Variations on $\Gamma_{\mathrm{E}}(m)$ include finding the squared energy (power) of each frame [10], taking the half-wave rectified first derivative of the detection function to measure the change in energy [17] and applying logarithmic compression [18].

2.2. Spectral Approaches

The remaining detection functions are based on analysis of *spectral* properties of the input and require the calculation of the Short Term Fourier Transform (STFT). For their derivation, we take the windowed STFT of $s(n)$ using an N length Hanning window $w(m)$ and find the k^{th} bin of the m^{th} short term spectral frame $S_k(m)$ as

$$S_k(m) = \sum_{n=-\infty}^{\infty} s(n) w(mh - n) e^{-j2\pi nk/N} \tag{2}$$

which has magnitude

$$R_k(m) = |S_k(m)| \tag{3}$$

where $S_k(m) \in C$ and $R_k(m) \in R$.

2.3. High Frequency Content

The high frequency content (HFC) [13] onset detection function $\Gamma_{\mathrm{H}}(m)$ is an energy based approach where strongest emphasis is given to the highest frequency bands. The magnitude spectrum $R_k(m)$ is multiplied by a linear weighting, corresponding to the bin number, k,

[3] http://lame.sourceforge.net/

$$\Gamma_{\mathrm{H}}(m) = \sum_{k=0}^{N/2} k R_k(m). \tag{4}$$

The main attribute of the HFC approach is to boost the contribution made by wide-band, noise-like percussive events such cymbals and snare drums.

2.4. Spectral Difference

An alternative spectral approach to measuring the frequency-weighted energy (as in the HFC method), is to generate an onset detection function based on the measurement of spectral change. During steady-state regions of the signal, i.e. not at a note onset, the magnitude spectrum should remain approximately constant [14]. Therefore a good prediction of the magnitude spectrum $\hat{R}_k(m)$ at frame m is the observed magnitude spectrum of the previous frame, $R_k(m-1)$. A spectral difference detection function $\Gamma_{\mathrm{S}}(m)$ can then be formed by measuring the Euclidean distance between the observed and predicted magnitude spectra

$$\Gamma_{\mathrm{S}}(m) = \sum_{k=1}^{K} \left| R_k(m) - \hat{R}_k(m) \right|^2. \tag{5}$$

Regions where $R_k(m)$ is dissimilar to $R_k(m-1)$ lead to peaks in $\Gamma_{\mathrm{S}}(m)$. This need not be the result of a change in signal energy, for which Γ_{E} and Γ_{H} are sensitive, but a change in spectral profile.

2.5. Phase Deviation

In addition to the steady-state model of the magnitude spectrum used to form the spectral difference onset detection function, a second model exists for the properties of the phase spectrum. This leads to the generation of the phase deviation onset detection function $\Gamma_{\mathrm{P}}(m)$ [15]. During steady-state regions, the phase velocity at the k^{th} bin should be approximately constant,

$$\phi_k(m) - \phi_k(m-1) \approx \phi_k(m-1) - \phi_k(m-2). \tag{6}$$

By adopting a short-hand notation for the left hand side of equation 6 we can write

$$\Delta\phi_k(m) = \phi_k(m) - \phi_k(m-1). \tag{7}$$

Then by substituting equation 7 into equation 6 and rearranging terms, we can make a prediction about the phase of the k^{th} bin for frame m given the observations of the previous two frames:

$$\hat{\phi}_k(m) = \mathrm{princarg}[\phi_k(m-1) + \Delta\phi_k(m-1)] \tag{8}$$

where princarg maps the phase value into the range [-π, π]. The difference between the predicted phase value $\hat{\phi}_k(m)$ and the observed phase value $\phi_k(m)$ at frame m is then used to form the phase deviation onset detection function

$$\Gamma_{\mathrm{P}}(m) = \frac{1}{K} \sum_{k=1}^{K} \mathrm{princarg}[\hat{\phi}_k(m) - \phi_k(m)]. \tag{9}$$

Instability in the phase spectrum can be the result of both percussive onsets, characterised by a significant change in signal energy, and *softer* tonal onsets [14] where there may be a negligible change in energy. The principal weakness of the phase based approach is a sensitivity to noise [14], which can mask the locations of onsets within the detection function.

2.6. Complex Spectral Difference

The motivation for the complex spectral difference detection function [16] is to combine the magnitude spectral difference $\Gamma_{\mathrm{S}}(m)$ and phase deviation $\Gamma_{\mathrm{P}}(m)$ approaches into a single method which is sensitive to percussive and tonal onsets, while at the same time, insensitive to noise.

The predictions of the magnitude spectrum $\hat{R}_k(m)$ used to generate the spectral difference detection function $\Gamma_{\mathrm{S}}(m)$, and the phase spectrum $\hat{\phi}_k(m)$ used in the phase deviation detection function $\Gamma_{\mathrm{P}}(m)$, can be represented in polar form to give a combined spectral prediction $\hat{S}_k(m)$ in the complex domain

$$\hat{S}_k(m) = \hat{R}_k(m) e^{j\hat{\phi}_k(m)}. \tag{10}$$

By comparing this to the observed complex spectrum in polar form, $S_k(m) = R_k(m)e^{j\phi_k(m)}$ we can derive the complex spectral difference detection function $\Gamma_{\mathrm{C}}(m)$, at frame m, as the sum of the Euclidean distance between the predicted and observed spectra for all k bins

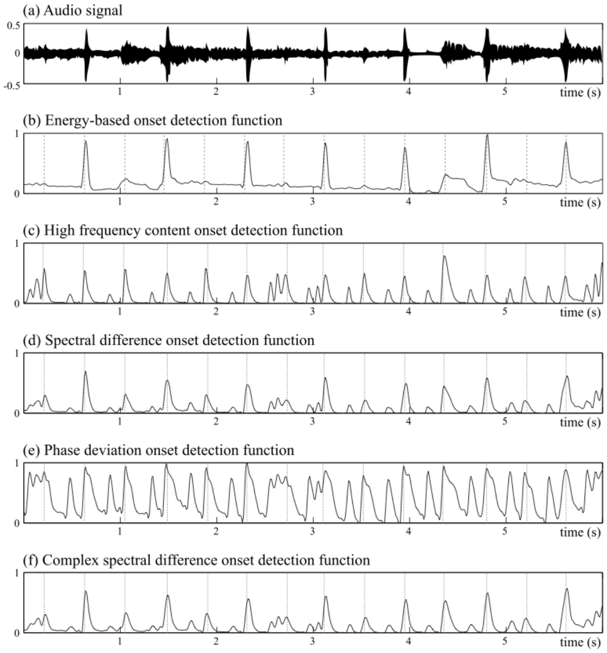

Fig. 1: Onset detection functions for a short excerpt of rock music. Dotted vertical lines indicate beat locations. (a) Audio signal. (b) Energy-based. (c) High Frequency Content. (d) Spectral Difference. (e) Phase Deviation. (f) Complex Spectral Difference.

$$\Gamma_{\mathrm{C}}(m) = \sum_{k=1}^{K} |S_k(m) - \hat{S}_k(m)|^2. \qquad (11)$$

Dixon [19] also presents a combined magnitude and phase based onset detection function, where the contribution of each bin in equation 9 is weighted by the magnitude spectrum $R_k(m)$.

An overview of each of the onset detection functions is shown in figure 1.

3. EXPERIMENTAL SETUP

To evaluate the effects of lossy audio encoding on onset detection performance we require a set of uncompressed audio files that include ground-truth onset annotations. For this we selected the Music Information Retrieval Exchange (MIREX) 2005 data set for the onset detection task [12]. This data set contains 85 audio clips encoded in PCM 16-bit 44.1kHz that have been hand-annotated by between 3 and 5 listeners (depending on the perceived complexity of the annotation) with ground truth onset times.

group	codec	bit-rate	sample-rate
PCM	PCM-16bit	1.4Mbps	44.1kHz
128kbps	LAME MP3	128kbps	44.1kHz
96kbps	LAME MP3	96kbps	32.0kHz
32kbps	LAME MP3	32kbps	16.0kHz
8kbps	LAME MP3	8kbps	8kHz

Table 1: Audio encoding schemes used in the test.

To create our test groups, each track in the MIREX data set is re-encoded at various bit-rates using the LAME MP3 codec. The LAME encoding is done using constant bit-rates and the default settings. The specifics of the encoding methods used are summarized in Table 1. Our re-encoding process results in five groups of audio files: PCM, 128kbps, 96kbps, 32kbps, and 8kbps.

To evaluate how onset detection performance degrades as lossy encoding bit-rates are decreased we follow the evaluation procedures outlined in the MIREX onset detection competition [12]. The detected onset times are compared with the ground-truth onset annotations. For a given ground-truth onset time, if there is a single detection in a tolerance time-window around it, it is considered as a *correct detection* (or true positive). If there is no detected onset for a given annotation this is counted as a *false negative.* All detections outside the tolerance windows (or multiple detections within a single allowance window) are counted as *false positives.* A tolerance time window of +/- 50ms is used in the evaluation.

We use the standard F-measure to evaluate the results of the onset detection. The F-measure is based on the calculation of two quantities, *precision*

$$P = \frac{O_{CD}}{O_{CD} + O_{FP}} \qquad (12)$$

and *recall*

$$R = \frac{O_{CD}}{O_{CD} + O_{FN}} \qquad (13)$$

where O_{CD} is the number of correctly detected onsets, O_{FP} is the number of false positive onsets, and O_{FN} is the number of false negative onsets.

Each of the files in the test database have been annotated by at least three different annotators. To account for the cross-annotation, we find the mean Precision and Recall per file. When then calculate the F-measure as

$$\mathrm{F} = \frac{2PR}{P+R} \tag{14}$$

To address the balance between false positives (over-detection) and false negatives (under-detection) within the peak-picking process a static threshold δ is subtracted from each onset detection function Γ_x. We define a range of threshold values such that $-0.1 \leq \delta \leq 0.25$ with increments of 0.05. Given each δ we extract the sequence of detected onsets and evaluate against our ground truth data and calculate a δ-dependent F-measure.

4. RESULTS

In Figures 2 to 6 the F-measure performance is plotted against δ for each onset detection method. The best F measure results and corresponding δ thresholds are also shown in table 2. The plots confirm that F-measure performance is fairly consistent between the test groups. A significant drop in performance is only seen in the 8kbps group for each onset detection method.

The results indicate that all onset detection methods tested are surprisingly robust in the presence of lossy audio encoding. Statistically significant changes in the F-measures only occur at an encoding of 8kbps ($0.0011 \leq p \leq 0.0492$). For the energy based method the changes in F-measure performance are only just within the 95% confidence interval at $p = 0.0492$ as calculated by the Wilcoxon rank-sum test [20].

The energy based method seems to be the most robust in the presence of lossy audio encoding. However, the energy based method has the lowest F-measure performance of all the methods overall. A visual inspection of the plots suggests that the complex spectral difference onset detection method has the best combination of F-measure performance and robustness in the presence of lossy audio encoding.

5. CONCLUSION

We have tested the effects of lossy audio encoding on the performance of audio-based onset detection.

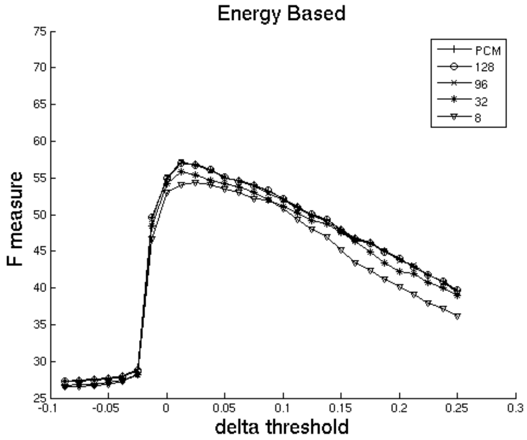

Fig. 2: Energy based onset detection performance

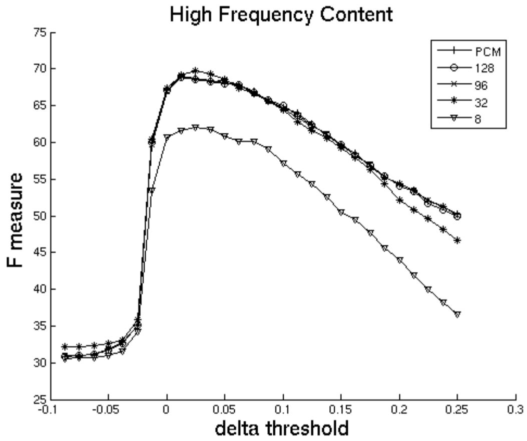

Fig. 3: High frequency content performance

	PCM		128kbps		96kbps		32kbps		8kbps	
method	F	δ	F	δ	F	δ	F	δ	F	δ
Γ_{E}	57.04	0.0125	57.05	0.0125	57.15	0.0125	55.83	0.0125	54.30	0.0250
Γ_{H}	68.98	0.0125	68.91	0.0125	68.99	0.0125	69.73	0.0250	62.13	0.0250
Γ_{S}	72.44	0.1000	72.34	0.1000	72.39	0.1000	70.63	0.0750	64.14	0.0750
Γ_{P}	71.58	0.0625	71.31	0.0750	70.50	0.0750	71.90	0.1000	63.89	0.1250
Γ_{C}	**74.29**	0.1125	**74.25**	0.1125	**74.27**	0.1125	**74.69**	0.0875	**68.30**	0.0625

Table 2: The best F-measure results and the corresponding δ thresholds for each onset detection method and each test group.

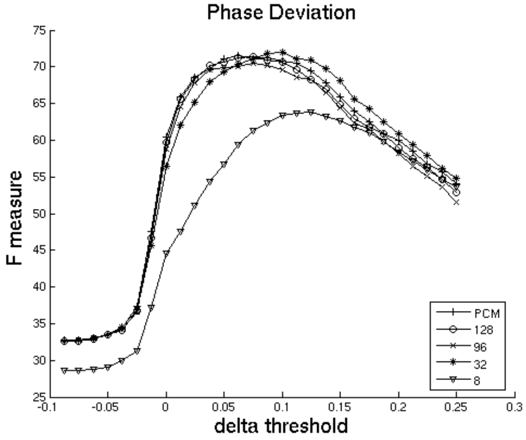

Fig. 4: Phase deviation performance

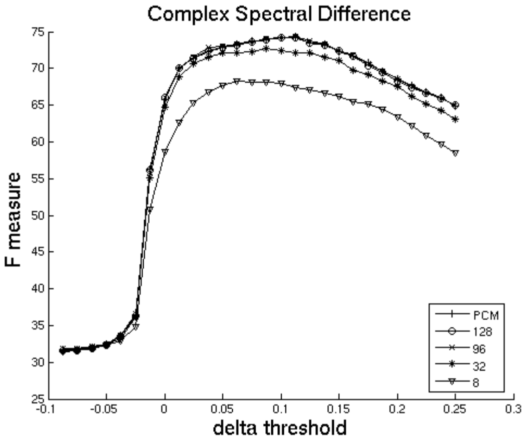

Fig. 6: Complex spectral difference performance

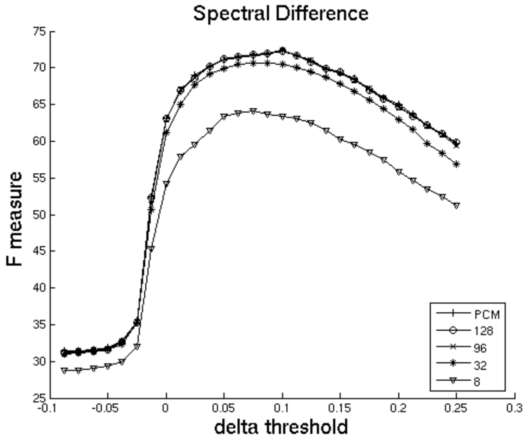

Fig. 5: Spectral difference performance

Using a standard data set and standard performance measures we have shown that five different state-of-the-art onset detection functions suffer no significant performance loss when operating on audio encoded at 32kbps / 16kHz LAME. Audio encoding at 8kbps / 8kHz LAME caused significant drops in performance for all onset detection methods tested.

Given that such low audio encoding bit rates result in audio quality that is likely to be unacceptable to even the most casual listener, these very-low bit rate encodings are almost non-existant in the world of digital audio. Therefore, we conclude that onset detection can reasonably be applied to almost any lossy encoded audio one is likely to encounter.

In future work we hope to examine other onset detection functions such as the one proposed by Stowell et al. [21] and examine other audio analysis tasks that use onset detection as a first step such as beat tracking and chord extraction. We also hope to use test groups created with different audio codec technology.

6. ACKNOWLEDGEMENTS

This work is supported as a part of the OMRAS2 project, EPSRC number EP/E017614/1.

7. REFERENCES

[1] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, July 2002.

[2] J.-J. Aucouturier and F. Pachet, "A scale-free distribution of false positives for a large class of audio similarity measures," *Pattern Recogn.*, vol. 41, no. 1, pp. 272–284, 2008. [Online]. Available: http://portal.acm.org/citation.cfm?id=1285179

[3] A. Berenzweig, B. Logan, D. Ellis, and B. Whitman, "A large-scale evaluation of acoustic and subjective music-similarity measures," *Computer Music Journal*, 2004.

[4] *ISO 15938-4:2001 MPEG-7: Multimedia Content Description Interface*, 2001.

[5] M. Casey, "Mpeg-7 sound-recognition tools," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 737–747, 2001.

[6] S. Dixon, "Automatic extraction of tempo and beat from expressive performances," *Journal of New Music Research*, vol. 30, pp. 39–58, 2001.

[7] T. Jehan, "Downbeat prediction by listening and learning," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Mohonk, NY, USA, 2005, pp. 267–270.

[8] K. West and S. Cox, "Finding an optimal segmentation for audio genre classification," in *Proceedings of 6th International Conference on Music Information Retrieval*, 2005, pp. 680–685.

[9] J. P. Bello and J. Pickens, "A robust mid-level representation for harmonic content in music signals," in *Proceedings of 6th International Conference on Music Information Retrieval*, London, United Kingdom, 2005, pp. 304–311.

[10] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, "A tutorial on onset detection in music signals," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, part 2, pp. 1035–1047, 2005.

[11] N. Collins, "A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions," in *Proceedings of the 118th AES Convention*, Barcelona, Spain, May, 28–31 2005, pre-print 6363.

[12] [Online]. Available: http://www.music-ir.org/mirex/2005/index.php/Main_Page

[13] P. Masri and A. Bateman, "Improved modelling of attack transients in music analysis-resynthesis," in *Proceedings of International Computer Music Conference (ICMC)*, Hong Kong, 1996, pp. 100–103.

[14] C. Duxbury, "Signal models for polyphonic music," Ph.D. dissertation, Department of Electronic Engineering, Queen Mary, University of London, 2004.

[15] J. P. Bello and M. Sandler, "Phase-based note onset detection for music signals," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2003, pp. 441–444, volume V.

[16] J. P. Bello, C. Duxbury, M. E. Davies, and M. B. Sandler, "On the use of phase and energy for musical onset detection in the complex domain," *IEEE Signal Processing Letters*, vol. 11, no. 6, pp. 553–556, 2004.

[17] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *Journal of Acoustical Society of America*, vol. 103, no. 1, pp. 588–601, 1998.

[18] A. Klapuri, “Sound onset detection by applying psychoacoustic knowledge,” in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. VI, Phoenix, USA, March 15–19 1999, pp. 3089–3092.

[19] S. Dixon, “Onset detection revisited,” in *Proceedings of 9th International Conference on Digital Audio Effects (DAFx)*, Montreal, Canada, 2006, pp. 133–137.

[20] J. L. Devore, *Probability and Statistics for Engineering and the Science 5th Edition.* California Polytechnic State University, 2000.

[21] D. Stowell and M. Plumbley, “Adaptive whitening for improved real-time audio onset detection,” in *Proceedings of International Computer Music Conference*, 2007.