Audio Engineering Society

Convention Express Paper 131

Presented at the 155th Convention
2023 October 25–27, New York, USA

Transient Detection Methods for Audio Coding

Senyuan Fan[1], Emily Kuo[1], Sneha Shah[1], and Marina Bosi[1]

[1] *Center for Computer Research in Music and Acoustics (CCRMA), Stanford University*

Correspondence should be addressed to Senyuan Fan (senyuanf@ccrma.stanford.edu)

ABSTRACT

Transient detection is an important algorithm in perceptual audio codecs that enables adaptation in filterbank resolution to effectively mitigate artifacts in encoded audio signals. We present a curated selection of transient detection methods tailored for audio coding purposes, namely high frequency energy (HFE), block perceptual entropy (BPE), time-frequency spectral flatness measure (TFSFM), and sub-block peak energy (SPE). The methods are then evaluated in a MUSHRA listening test using selected critical materials from the EBU-SQAM dataset. This paper provides insights into perceptual audio coding and paves the way for further optimization in transient detection.

1 Introduction

Perceptual audio coding is a crucial technology that has transformed the way audio content is stored, transmitted, and consumed. Widely adopted audio coding standards such as MPEG Advanced Audio Coding (AAC) [1] and MPEG Layer III (MP3) [2] efficiently compress audio signals while maintaining high quality sound. By exploiting the limitations of human hearing, these perceptual audio codecs allocate data in a way that minimizes perceptible quantization noise.

Despite their effectiveness in achieving high compression ratios, perceptual audio codecs can suffer from coding artifacts due to inherent limitations and trade-offs in the internal signal representation. For instance, unwanted artifacts could be introduced when sudden signal variation exceeds the fixed time resolution of the signal representation. One common artifact is pre-echo distortion, which occurs when a faint replica of a sound is heard just before the sound itself [3]. Pre-echo is caused by the spread of quantization noise in time prior to the onset of the attack.

To mitigate such effects, audio codecs need to have adaptive filterbank resolution. A commonly used method is block switching, which enables perceptual audio codecs to adaptively change the block length of the time to frequency mapping stage based on the characteristics of the input audio signal [4]. For transient-like segments in a signal, shorter blocks are employed to enhance time resolution and thus reduce the spread of quantization noise in the time domain. Conversely, for harmonic and steady state signals, longer blocks are employed to achieve higher frequency resolution, improving the representation of closely spaced frequency components.

Therefore, a robust and accurate transient detection method is crucial to the quality of audio codecs. However, related research in musical onset detection [5, 6] aim to detect the start of musical notes, which include a wide variety of non-transient onsets in audio signals. There is still a lack of published research that explains available transient detection methods for audio coding purposes.

Thus, in this paper, we present a selection of transient detection methods that aim to detect time-domain transients in the most critical situation for audio coding. The different transient detection methods are presented with theoretical explanations and listening test evaluations.

2 Baseline Codec

We implemented a baseline perceptual audio codec with the guidance of [4]. Figure 1 shows the structure of the codec. The finite-length blocks of the input signal are mapped to the frequency domain via the critically sampled modified discrete cosine transform (MDCT) [7]. The output of the MDCT is then quantized at a reduced data rate while keeping the quantization noise below the masked threshold, which is computed on a block-by-block basis in the Psychoacoustics Model stage. This method allows for efficient audio compression without sacrificing perceived audio quality.

Representing audio signals in the frequency domain with a fixed resolution, i.e., constant block size transform, may cause undesired artifacts. Typically, audio codecs operate at high frequency resolution (long block transform). However, when a transient occurs, in order to avoid the pre-echo artifacts, high temporal resolution (short block transform) is necessary.

Our baseline codec implements a method called block switching based on [8] to mitigate these artifacts. When a transient is detected, the time resolution of the transform is increased by adopting a short block transform. In this fashion, the spreading of the quantization noise in the time domain is significantly reduced and the pre-echo distortion is virtually eliminated. As shown in Figure 2, the coded signal using an adaptive block transform (Figure 2C) does not show significant energy spreading before the transient occurs.

In order to successfully implement the dynamic behavior of the time-frequency transform, it is essential to accurately detect the occurrences of transients in the input signal. Therefore, we present four methods to detect transients in an audio codec in Section 3.

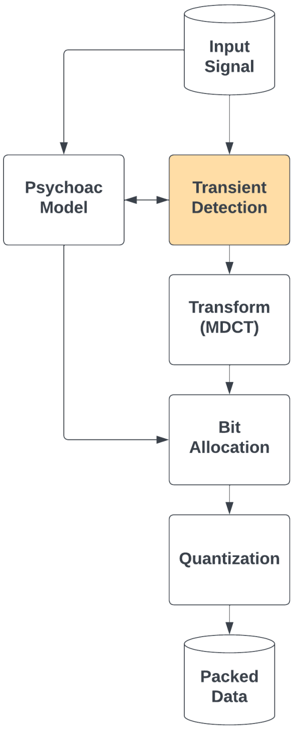

Fig. 1: *Structure of the baseline codec.*

3 Transient Detection Methods

We anaylze the following four transient detection methods:

1. High Frequency Energy (HFE)
2. Block Perceptual Entropy (BPE)
3. Time Frequency Spectral Flatness Measure (TFSFM)
4. Sub-block Peak Energy (SPE)

3.1 High Frequency Energy (HFE)

Typically, sudden variations in high frequency energy (HFE) are associated with transients. The HFE method is defined by Equation 1, where K is the high frequency

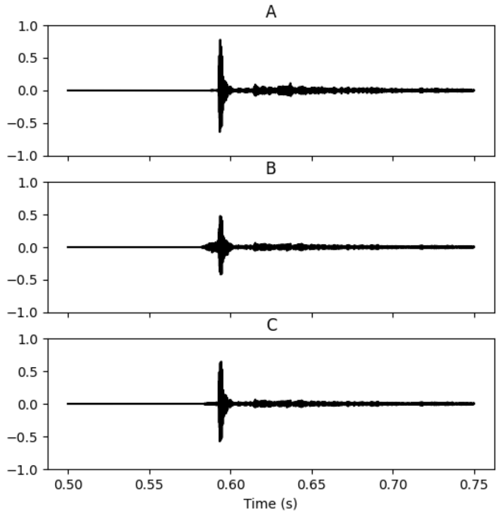

Fig. 2: *A. Original castanets signal. B. Castanets coded at 96 kbps with block size N = 1024. C. Castanets coded at 96 kbps with adaptive block size.*

cutoff index, N is the block size, and $\frac{8}{N^2}$ is the normalization factor based on the use of the Kaiser-Bessel derived (KBD) window [9].

In our implementation, the cutoff frequency is set at 8kHz. HFE is computed as the high frequency sound pressure level (SPL) as follows:

$$\begin{gathered}\mathrm{HFE} = \mathrm{SPL}\Big(\sum_{K}^{N/2-1} \frac{8}{N^2}|\mathrm{block_{FFT}}|^2\Big) \\ \mathrm{K} = \mathrm{N} \times \frac{\text{cutoff frequency}}{\text{sampling rate}} \\ \mathrm{block_{FFT}} = \mathrm{FFT}(\mathrm{KBD}(\mathrm{block}))\end{gathered} \tag{1}$$

The motivation behind the HFE is that transients typically exhibit high energy in high frequency ranges. To illustrate this, we can examine the signals from the EBU-SQAM dataset (SQAM) [10]. As shown in Figure 3, the castanets signal, characterized by a multitude of pure transients, has high energy in frequencies above 8kHz when the transients occur. In contrast, the violin signal Figure 4, characterized by smooth tonal sounds, has significantly reduced energy in the higher frequency bands.

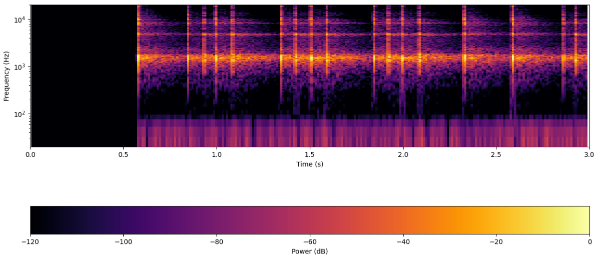

Fig. 3: *Spectrogram of the castanets signal.*

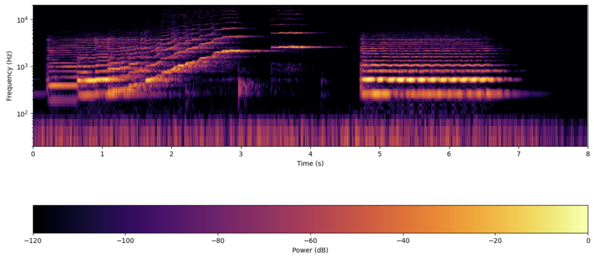

Fig. 4: *Spectrogram of the violin signal.*

Therefore, by calculating the HFE difference between the current block and the previous block, transient occurrences in an audio signal can be detected fairly accurately.

3.2 Block Perceptual Entropy (BPE)

Block perceptual entropy (BPE) stems from the idea of perceptual entropy (PE), which measures the amount of perceptually relevant information in an audio signal [11]. PE is used to detect transients in MPEG Psychoacoustic Model 2 [2].

BPE represents the minimum number of bits per block required to achieve transparency in an encoded signal, given the masking curve of the block. A simplified BPE calculation is introduced in [4], as defined in Equation 2, where c is the critical band index, B is the number of critical bands in the block, N_c is the number of frequency components in critical band c, and SMR_c is the maximum signal to mask ratio (SMR) at band c.

$$\mathrm{BPE} = \frac{1\text{ bit}}{6.02\text{ dB}} \sum_{c=0}^{B-1} N_c \times \mathrm{SMR}_c,\ \mathrm{SMR}_c > 0 \tag{2}$$

As shown in Equation 2, BPE is dependent on the signal to mask ratio (SMR), which represents the difference between the SPL of the signal and the the energy of the masked threshold measured in dB. The masked threshold is computed in the Psychoacoustics Model stage as a combination of the hearing threshold and the masking thresholds pertaining to various signal components. The quantization noise positioned in the spectral region below the masked threshold is considered inaudible [4]. In a perceptual audio codec, signals with a larger SMR values require a higher number of bits allocated to ensure that the quantization noise falls below the masked threshold. Because transient signals typically have a flat frequency spectrum that require more bits to represent, an increase in BPE values indicates transient occurrences.

3.3 Time Frequency Spectral Flatness Measure (TFSFM)

The spectral flatness measure (SFM) was introduced as a measure of whiteness of a speech signal [12]. For discrete-time audio signals, SFM is defined in [13] as

$$\mathrm{SFM} = \frac{(\prod_{k=0}^{K-1} |X(k)|^2)^{1/K}}{\frac{1}{K}\sum_{k=0}^{K-1} |X(k)|^2} \tag{3}$$

where $|X(k)|^2$ is the power spectral density (PSD) of the signal and K is the number of frequency bins in a block. In the SFM implementation, $K = N/2$.

SFM represents the ratio between the geometric mean and arithmetic mean of the signal PSD. Notice that SFM varies between 0 and 1, where SFM = 1 implies a signal with a perfectly flat spectrum. SFM can be used to distinguish between steady state conditions and transients [14]. However, a randomly distributed signal could have similar SFM to a transient signal. Thus, to improve the robustness and accuracy of the transient detection task, we propose to consider the temporal flatness measure (TFM) in addition to SFM.

When a transient is present, the signal fluctuates in the time domain but remains flat in the frequency domain. Inspired by [15], we define TFM as

$$\mathrm{TFM} = \frac{(\prod_{n=0}^{N-1} x(n)^2)^{1/N}}{\frac{1}{N}\sum_{n=0}^{N-1} x(n)^2} \tag{4}$$

where $x(n)$ represents the discrete-time input signal. Because transients exhibit spectral energy evenly distributed throughout the spectrum (as shown in Figure 3), given a transient, SFM would approach 1 while TFM would approach 0. Thus, we can calculate the ratio between SFM and TFM, namely time-frequency spectral flatness measure (TFSFM), as

$$\mathrm{TFSFM} = \frac{\mathrm{SFM}}{\mathrm{TFM}} \tag{5}$$

The TFM for a randomly distributed signal would be higher than that for a transient signal, leading to a reduced TFSFM value. Similar to HFE, the TFSFM difference between blocks serves as a detection function, and a sudden increase indicates the presence of a transient.

3.4 Sub-block Peak Energy (SPE)

Sub-block peak energy (SPE) is inspired by the AC-3 codec [16]. Transients are identified by finding sudden increases in amplitude between sub-blocks in a signal. If N is the block size, SPE treats the left N/2 samples as previous data and the right N/2 samples as current data and detects transients in the right half of the block.

Similar to HFE, the signal is first high pass filtered at 8kHz as transients generally occur at high frequencies. Each block is then divided into sub-blocks of size N/2, N/4, and N/8 respectively. The peaks in each sub-block are calculated in Equation 6, where $x(n)$ is the n^{th} sample in the block, j = 1, 2, 3 is the layer number, and k is the segment number within layer j.

$$\begin{gathered} P[j][k] = abs(max(x(n))) \\ \text{for } k = 1,...,2^{j-1} \\ n = (N\times(k-1)/2^j), (N\times(k-1)/2^j)+1,..., \\ (N\times k/2^j)-1 \end{gathered} \tag{6}$$

The peak amplitudes are first checked to be above a zero threshold to avoid catching amplitude changes in very soft signals. The zero threshold is set to 1500 when samples are represented as int16. Then, the peak amplitude of each sub-block is compared to that of the previous sub-block, as shown in Equation 7. If the peak amplitude is greater than the zero threshold and the amplitude increase in each layer is greater than each layer's threshold, a transient is flagged. The thresholds ensure that only sudden increases in amplitude are accounted for, preventing false detections. The threshold for each layer can be found in Table 1.

Layer	Sub Block Size	Threshold
1	512 (N/2)	0.4
2	256 (N/4)	0.4
3	128 (N/8)	0.07

Table 1: SPE thresholds

Method	Threshold
HFE	10
BPE	300
TFSFM	0.6

Table 2: Pre-tuned thresholds for transient detection methods

$$P[j][k] \times T[j] > P[j][k-1] \tag{7}$$

where $T[j]$ is the pre-tuned threshold for level j

4 Results

4.1 Detection Result Analysis

To evaluate the transient detection methods proposed in Section 3, we present a detailed comparison of each method's transient detection performance.

For each block of input data, our methods yield a raw output value. For BPE, HFE and TFSFM, we compute the detection function according to the raw output difference between the current block and the previous block. Then, transients are identified by applying a pre-tuned threshold to the detection function, as summarized in Table 2. In other words, if the raw output of the current block is larger than that of the previous block by the pre-tuned threshold, we determine that a transient is present in the current block. For SPE, a block is flagged as a transient if any of its sub-layers detects a transient. Thus, its raw detection function outputs a boolean value that combines the detection results of all layers.

Figure 5 shows each method's detection function output and its pre-tuned threshold for the castanets signal. The peak locations in the detection functions correspond to the transient locations in the original waveform. Figure 6 also shows that for a steady tonal signal, the pre-tuned threshold is not prone to false positives.

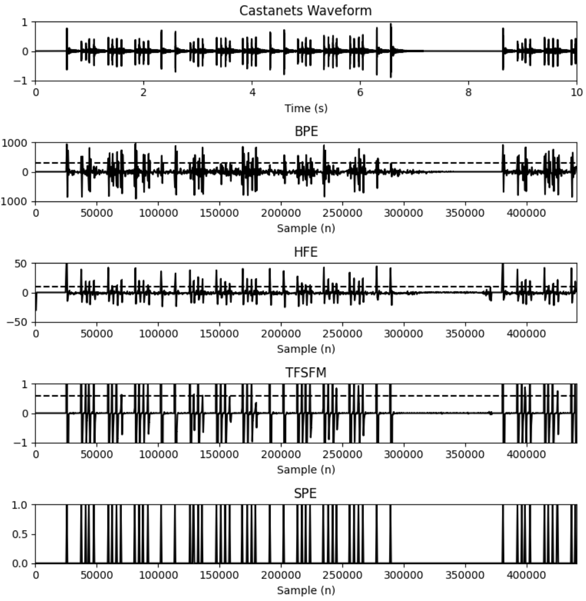

Fig. 5: *Detection function output for the castanets signal. The dashed line is the pre-tuned threshold.*

Figure 7, 8, and 9 show a side-by-side comparison of the detection methods. Each input signal is processed with a block size of 1024 and 50% overlap. The castanets signal from SQAM [10] presents distinct transients. In Figure 7, it is observed that nearly all detection methods accurately identify each attack, with the exception of BPE missing a few. In contrast, for the violin signal from SQAM [10] depicted in Figure 8, it is expected that no transients would be detected because the signal consists of smooth note transitions. However, sparse false detections are observed in the outputs of HFE and BPE.

When the detection methods are used on more complex signals like the rock song presented in Figure 9, variations among the detection methods become more evident. In particular, while HFE detects transients for each hi-hat sound (which contains high energy in high frequency ranges), BPE tends to detect transients that align with the loud attacks of snare sounds.

4.2 Listening Test Evaluation

We conducted a MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) [17] listening test with 15 participants using the webMUSHRA platform [18].

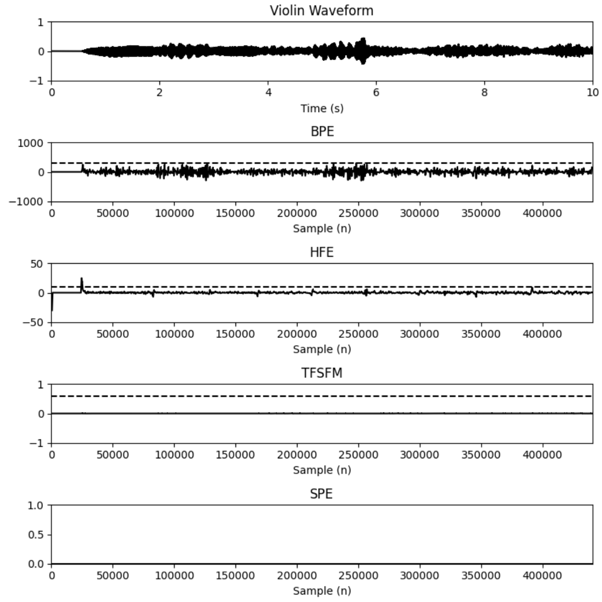

Fig. 6: *Detection function output for the violin signal. The dashed line is the pre-tuned threshold.*

The listening test included critical materials from EBU-SQAM [10] (castanets, violin, glockenspiel, harpsichord, and German speech) as well as samples of applause and a rock song. These signals were chosen as they include percussive, tonal, and mixed sounds. Following the MUSHRA specifications, for each critical material, subjects were presented with the reference, two low-pass anchors with cutoff frequency at 7kHz and 3.5 kHz, and excerpts coded at 96 kbps utilizing the four detection methods presented (BPE, HFE, TFSFM, SPE). Each test also included excerpts coded at 96kbps using a bypass version of the codec, where the transient detection stage is bypassed and only the highest filterbank resolution is used. In our codec implementation, the long and short blocks are of size 1024 and 256 samples respectively. At a sampling frequency of 44.1 kHz, the highest frequency resolution is 43.06 Hz, and the highest time resolution is 5.8 ms.

Subjects underwent training before the listening test, which included getting familiar with the test setup and transient coding artifacts. Post-screening was applied following the ITU-R BS.1534 recommendation. Based on these criteria, data from 2 subjects were excluded. The results were analyzed with a 95% confidence interval.

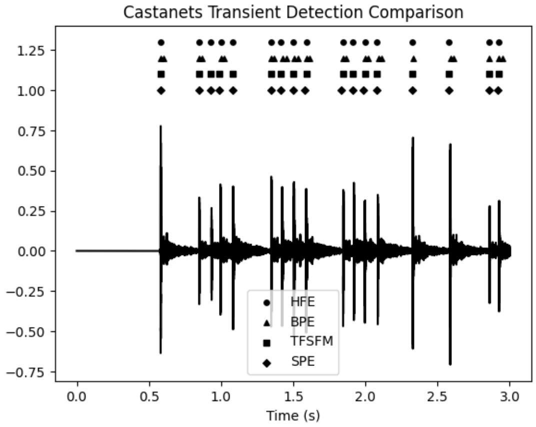

Fig. 7: *Detection results for the castanets signal.*

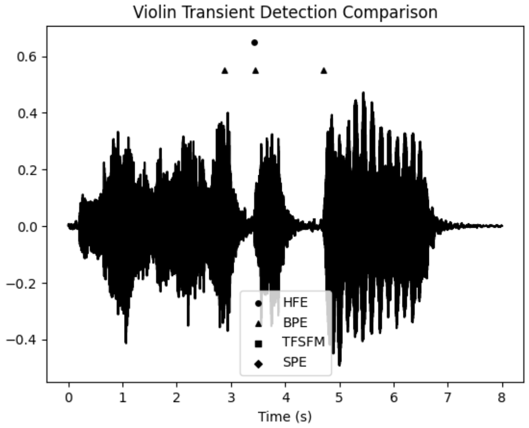

Fig. 8: *Detection results for the violin signal.*

The listening test results for the castanets, violin, and rock song signals are summarized in Figure 10, 11, and 12. Figure 10 demonstrates that for percussive sounds that have many transients, using HFE, SPE, and TFSFM for block switching yield better perceptual audio quality than bypassing block switching. At the same time, Figure 11 shows that these methods have no negative impact on the perceptual quality of tonal signals. The results for mixed signals, like the rock song in Figure 12, are not as clear and show no statistically significant improvement or degradation in perceived audio quality.

5 Conclusion

This paper presents a comprehensive overview of a selection of transient detection methods for audio coding,

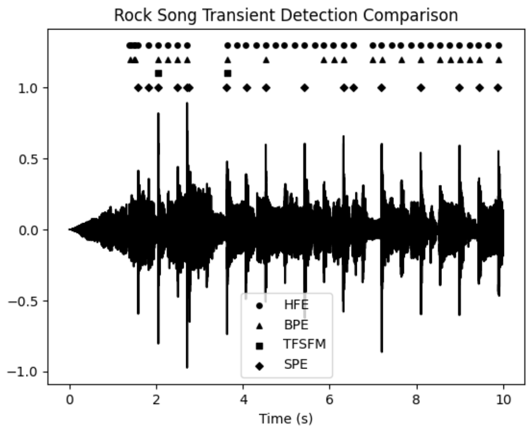

Fig. 9: *Detection results for a standard rock song.*

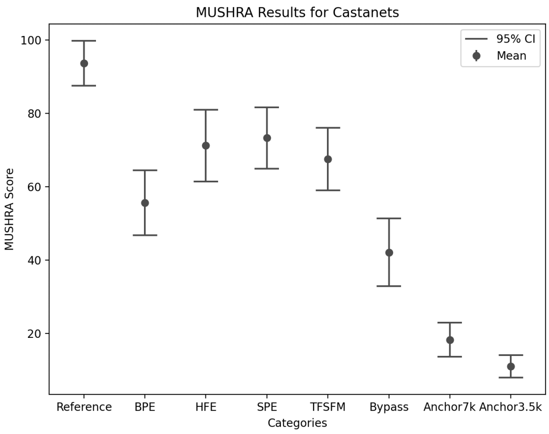

Fig. 10: *Listening test results for the castanets signal.*

including high frequency energy, block perceptual entropy, time frequency spectral flatness measure, and sub-block peak energy.

A MUSHRA [17] listening test was conducted to evaluate differences in human perception of the transient detection methods. Results show that listeners prefer at least 3 of the methods (HFE, SPE, TFSFM) over coding with high frequency resolution for transient-like signals. Furthermore, the methods come with no significant degradation in the perceived quality of coded tonal and mixed signals.

In the future, we would like to explore the application of convolutional neural networks (CNN) and other machine learning-based approaches for the purpose of transient detection in perceptual audio coding.

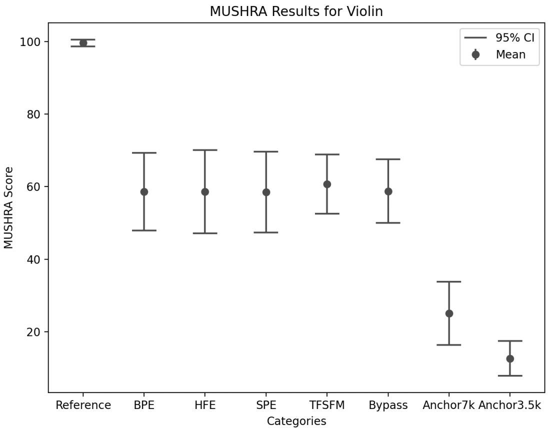

Fig. 11: *Listening test results for the violin signal.*

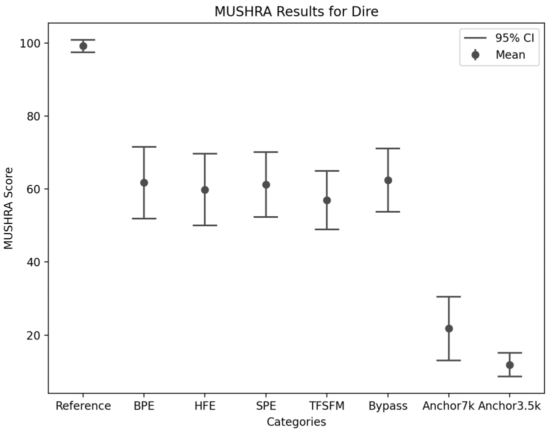

Fig. 12: *Listening test results for a standard rock song.*

References

[1] Bosi, M., Brandenburg, K., Quackenbush, S., Fielder, L., Akagiri, K., Fuchs, H., and Dietz, M., "ISO/IEC MPEG-2 advanced audio coding," *Journal of the Audio Engineering Society*, 45(10), pp. 789–814, 1997.

[2] ISO/IEC 11172-3, "Information technology — Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s — Part 3: Audio," Standard, International Organization for Standardization/International Electrotechnical Commission, 1993.

[3] Erne, M., "Perceptual Audio Coders "What to listen for"," in *AES 111th Convention*, 2001.

[4] Bosi, M. and Goldberg, R. E., *Introduction to Digital Audio Coding and Standards*, volume 721, Springer Science & Business Media, 2003.

[5] Bello, J., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., and Sandler, M., "A tutorial on onset detection in music signals," *IEEE Transactions on Speech and Audio Processing*, 13(5), pp. 1035–1047, 2005, doi:10.1109/TSA.2005.851998.

[6] Schlüter, J. and Böck, S., "Improved musical onset detection with Convolutional Neural Networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6979–6983, 2014, doi:10.1109/ICASSP.2014.6854953.

[7] Princen, J., Johnson, A., and Bradley, A., "Subband/Transform coding using filter bank designs based on time domain aliasing cancellation," in *ICASSP '87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 12, pp. 2161–2164, 1987, doi:10.1109/ICASSP.1987.1169405.

[8] Bosi, M. and Davidson, G., "High-Quality, Low-Rate Audio Transform Coding for Transmission and Multimedia Applications," in *AES 93rd Convention*, 1992.

[9] Fielder, L. D., Bosi, M., Davidson, G., Davis, M., Todd, C., and Vernon, S., "AC-2 and AC-3: Low-Complexity Transform-Based Audio Coding," in *AES Conference: Collected Papers on Digital Audio Bit-Rate Reduction*, 1996.

[10] EBU Tech 3253, "Sound Quality Assessment Material recordings for subjective tests," Technical report, European Broadcasting Union, 2008.

[11] Johnston, J., "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on Selected Areas in Communications*, 6(2), pp. 314–323, 1988, doi:10.1109/49.608.

[12] Gray, A. and Markel, J., "A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 22(3), pp. 207–217, 1974, doi:10.1109/TASSP.1974.1162572.

[13] Bosi, M., "Filter Banks in Perceptual Audio Coding," in *AES 17th International Conference: High-Quality Audio Coding*, 1999.

[14] Taghipour, A., Jaikumar, M. C., and Edler, B., "A psychoacoustic model with Partial Spectral Flatness Measure for tonality estimation," in *2014 22nd European Signal Processing Conference (EUSIPCO)*, pp. 646–650, 2014.

[15] Zhang, X., Cai, C., and Zhang, J., "A transient signal detection technique based on flatness measure," in *2011 6th International Conference on Computer Science Education (ICCSE)*, pp. 310–312, 2011, doi:10.1109/ICCSE.2011.6028641.

[16] Doc. A/52:2012, "ATSC Standard: Digital Audio Compression (AC-3, E-AC-3)," Standard, Advanced Television Systems Committee, 2012.

[17] ITU-R Recommendation BS.1534-3, "Method for the subjective assessment of intermediate quality level of audio systems," 2015.

[18] Schoeffler, M., Bartoschek, S., Stöter, F.-R., Roess, M., Westphal, S., Edler, B., and Herre, J., "webMUSHRA — A Comprehensive Framework for Web-based Listening Tests," *Journal of Open Research Software*, 6(1), p. 8, 2018, doi:10.5334/jors.187.