Identification of Audio Coding Artifacts Generated due to Bandwidth Extension Schemes

Dipanjan Datta Roy*, Andreas Niedermeier+, Bernd Edler*
*International Audio Laboratories, Erlangen, Germany
+Fraunhofer Institute for Integrated Circuits (IIS), Erlangen, Germany

***Abstract*—Identification of audio coding artifacts has a wide range of applications, including encoder design, audio post-processing, and quality assessment. In this paper, we focus on identifying audio artifacts associated with (semi-)parametric audio coding, particularly in the context of audio bandwidth extension (BWE) schemes. The application of BWE can lead to artifacts in the high-frequency target area that depend on the crossover frequency and the characteristics of the signal. We aim to identify on a frame-by-frame basis two common types of artifacts: Tonality Mismatch (TM) and Unmasked Noise (UN). To achieve this, we introduce a novel method for detecting these artifact types, incorporating two key components into our model: a spectral flatness measure and harmonic analysis of the signal.**

***Index Terms*—bandwidth extension, artifact detection, spectral flatness, harmonic analysis**

I. Introduction

Modern perceptual audio codecs [1] designed for low bit rate transmission employ coding tools that enable parametric or semi-parametric representations of audio signals. In parametric representation, the audio waveform is conveyed through a sparse set of parameters rather than temporal or spectral samples [2]. These techniques often prioritize efficiency over waveform preservation, leading to reconstructed waveforms that may differ from the original. However, perceptually, the reconstructed signal can closely resemble the original audio, providing acceptable quality at very low bitrates [2]. The non-waveform preserving nature of these coding schemes can introduce audible artifacts in the reconstructed audio signal when misused or overextended. We categorize these artifacts into two main types: Tonality Mismatch (TM) and Unmasked Noise (UN) [3], which will be discussed later in this article.

Identifying coding artifacts is essential to improve audio quality, as it aids in encoder design, post-processing, and understanding the limits of compression technologies [4]. Furthermore, artifact detection can be applied in objective audio quality assessment. Methods like PEAQ [5] provide an overall score on a MOS [6] scale, but do not specify artifact locations. The approach developed in this paper effectively identifies these locations and classifies the signal frame according to the artifact type.

II. Theoretical Background

Bandwidth extension (BWE) techniques have been successfully applied in perceptual audio coding at very low bit rates. Conventional codecs that do not utilize BWE techniques are limited in the audio bandwidth they can transmit due to insufficient bit availability. In contrast, BWE-enabled codecs incorporate side information into the bitstream, allowing the decoder to reconstruct the high-frequency spectrum from the transmitted low-frequency components. Consequently, a full reconstruction of the audio spectral bandwidth can be achieved with a transmitted bandwidth as low as 4 kHz [7]. The starting frequency for bandwidth reconstruction is called the crossover frequency. One of the first BWE techniques is Spectral Band Replication (SBR) [8], which was used in the MPEG-4 High Efficiency Codec (HE-AAC) [9]. Some more advanced BWE techniques include enhanced SBR [10] and Intelligent Gap Filling (IGF) [11], schemes that are commonly used in the state-of-the-art audio codecs.

The generation of artifacts resulting from BWE techniques is influenced by the crossover frequency and the characteristics of the input signal. In this paper, we consider the audio items of the ODAQ data set [12], where the TM and UN artifacts are generated as shown in [3], where the BWE technique used is IGF. We further classify the two types of artifacts, tonality mismatch and unmasked noise, by examining the spectral properties of the signal. This subclassification is based on the observation that the perceptual quality of the reconstructed signal varies depending on the copied content. In the following paragraph, we identify the scenarios in which these artifacts are generated. Figure 1 illustrates the spectrogram representation of the different types of artifacts, the different time and frequency ranges are selected so that the artifacts are clearly visible.

Noise Substitution (NS) : If the original audio stimulus contains tonal components in the high-frequency spectrum, but the reconstructed signal appears noise-like at those frequencies, artifacts resulting in a noisy sound occur. This noise-like high-frequency spectrum may arise from copying noise-like sections from the original signal's low-frequency spectrum or from replacing the high-frequency portion with random noise while preserving the spectral envelope. For signals with a dense harmonic structure in the high frequencies, noise substitution can cause significant distortion, particularly if the crossover frequency is low. However, if the high-frequency components in the original signal lack harmonic structure, the perceptual impact of noise substitution is less pronounced. Our algorithm distinguishes between noise substitutions with severe and mild effects. Figure 1(a) illustrates this artifact,

*A joint institution of the Friedrich-Alexander-Universitat Erlangen-Nurnberg (FAU) and Fraunhofer Institute for Integrated Circuits (IIS).

where the tonal components have been replaced by noise.

Exaggerated Tonality (ET) : The original audio stimulus lacks tonal components in the high-frequency range but contains tonal components at low frequencies. In contrast, the reconstructed signal exhibits tonal components in the high-frequency range. In this context, the high-frequency section of the reconstructed signal becomes excessively tonal, resulting in artifacts that sound perceptually rough. Figure 1(b) illustrates this artifact, showing sharp horizontal lines in the reconstructed spectrum, which represent the low-frequency tonal components that have been copied to the high frequencies.

Harmonicity Mismatch (HM) : Both the original and reconstructed signals exhibit tonal characteristics, with a strong harmonic structure present throughout the spectrum. The copying of tonal components from the low-frequency region to the high-frequency segment disrupts the harmonic continuity in the reconstructed signal. When tonal components are placed too closely together in the high-frequency range, it leads to modulation or beating artifacts. This artifact is illustrated in Figure 1(c), where the difference in harmonic continuity between the reference and reconstructed signals can be observed.

Harmonicity Mismatch with Noise Filling (HMNF): The original signal maintains a harmonic structure throughout the spectrum, while the reconstructed signal preserves this harmonic structure, but introduces noise-like components in the high-frequency range. This effect typically arises in signals with a higher fundamental frequency. As these components are copied into the high-frequency spectrum, a noise-like substitution occurs among the harmonic elements. This artifact is illustrated in Figure 1(d), where noise is interspersed between the tonal components.

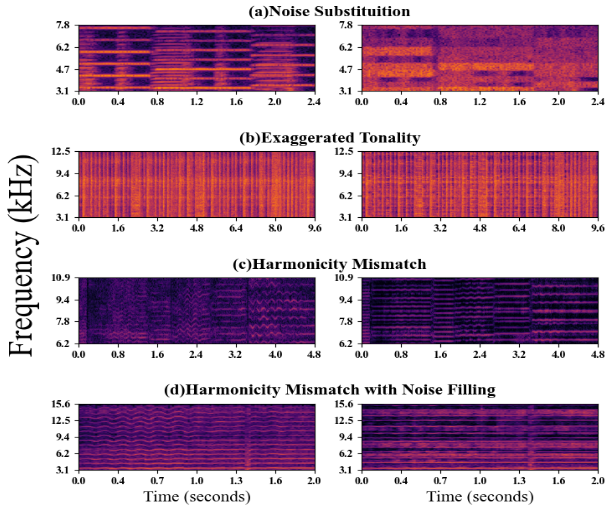


Fig. 1. Spectrogram representation of various artifact types. Each artifact shown with the original signal spectrum on the left side and the degraded spectrum on the right.

III. Detection Algorithm

In the previous section, we discussed the types of artifact, highlighting two primary factors that influence the effect of copying components from the lower to the higher frequency range of the spectrum. The first factor is the harmonic structure of the original signal. The second factor is the nature of the substitution in the high-frequency range, specifically whether it involves the substitution of noise or tonal components.

To model the two factors, we have utilized two features. The combination of both allows us to classify the audio signal frame into artifact type. In the next section, we describe how these features are calculated and used in classification.

A. ERB-Based Spectral Flatness Measure

The Spectral Flatness Measure (SFM) [13] can be used to quantify whether an audio signal is tonal in nature or noise-like. The SFM value ranges from 0 to 1, where a pure tonal signal has an SFM value of 0, and a pure white noise signal has an SFM value of 1. When comparing the SFM values of a reference signal and its coded version, if the SFM of the coded signal is less than that of the reference signal, it can be interpreted as the coded signal being more tonal compared to the reference. Conversely, if the SFM of the coded signal is greater than that of the reference, it indicates that the coded signal has become noisier than the original signal. In our case, we calculate the SFM for each time frame and for each of the 64 Equivalent Rectangular Bandwidth (ERB) bands. The grouping of the linear frequency bins into ERB bands is done based on the center frequencies of the ERB as shown in [14]. We refer to the method for calculating SFM for ERB bands as ERB-Based SFM (ESFM).

Figure 2 (a) presents the block diagram for the calculation of the ESFM feature. The discrete time-domain windowed signal x(n) is transformed into the frequency domain using the Modified Discrete Cosine Transform (MDCT) [15], as described by the equation:

$$X_k = \sum_{n=0}^{2N-1} x_n \cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}+\frac{N}{2}\right)\left(k+\frac{1}{2}\right)\right] \qquad (1)$$

where the transform length N is 512 and k is the frequency index, which ranges from 0 to 511. The window function we consider is a sine window with 50% overlap. The transform used here to convert from the time domain to the frequency domain is the MDCT, which is commonly employed in state-of-the-art audio codecs. For each time frame, the MDCT coefficients are grouped according to the number of frequency bins corresponding to each ERB band. At low frequencies, the frequency resolution for the ERB bands is quite narrow. In cases where the number of frequency bins is less than 6, we group the MDCT coefficients of 5 consecutive frequency bins corresponding to the ERB band. After grouping the coefficients, we compute the entropy-based SFM [16] for the i^{th} ERB band according to (2)

$$ESFM_i = 2^{-\sum_{m=0}^{P-1} X(m)_i \cdot \log_P X(m)_i} - 1 \qquad (2)$$

where $X(m)$ is the normalized MDCT coefficient such that the sum of the coefficients for the i^{th} band is 1, and P is the number of coefficients in the i^{th} ERB band. We do not use the classical definition for calculating flatness due to the issues outlined in [16]. Instead, the entropy-based method provides better results and greater distinction between tonal and non-tonal regions in a signal.

B. Harmonic Analysis

To evaluate the strength of the harmonic structure within the signal, we perform a harmonic analysis. Harmonic components are identified in the spectrogram of the audio signal by horizontal lines, while vertical lines typically correspond to percussive elements [17]. To quantify the strength of the harmonics in the higher frequency spectrum of a signal, we propose a feature called Harmonic Spread (HS). Harmonic Spread for any time frame is defined as the ratio of the number of harmonic peaks above the minimum crossover frequency considered relevant for BWE schemes to the total number of harmonic peaks. In our analysis, a threshold of 3 kHz is used as the minimum crossover frequency for BWE applications. We will later define what constitutes a harmonic peak.

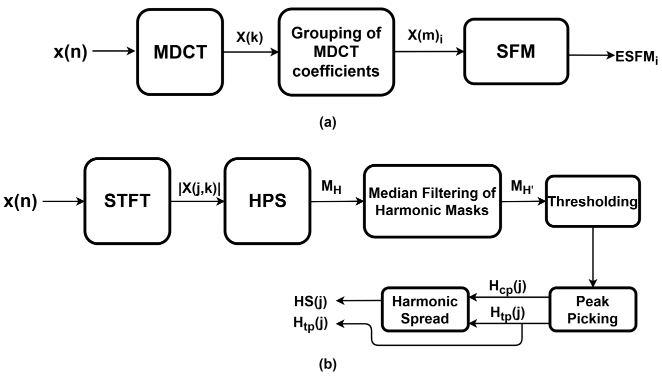


Fig. 2. (a) ESFM calculation procedure (b) Harmonic Analysis procedure

Fig. 2 (b) shows the block diagram of how the feature is calculated. The signal in the discrete time domain x (n) is converted to its spectrogram representation $X(j,k)$ by applying a Short Time Fourier Transform (STFT), with transform length of 2048 and hop size of 512. The window function used was Hann.

After the calculation of STFT, harmonic and percussive separation (HPS) is performed to generate masks according to the algorithm presented in [18] and [19]. In this paragraph, we provide a brief description of the algorithm. The magnitude spectrogram is median filtered along both the horizontal and vertical directions to obtain the harmonic-enhanced spectrogram $\mathbf{H}$ and the percussive-enhanced spectrogram $\mathbf{P}$, respectively. From the median-filtered spectrograms, soft masks are generated for the respective components, as shown in [18]. Equation (3) illustrates how masks are calculated for an arbitrary time frame index j and frequency index k; β is referred to as the separation factor, as introduced in [19], and p is the power to which each individual element of the spectrograms $\mathbf{H}$ and $\mathbf{P}$ is raised. The mask values indicate the extent to which each of the time-frequency bins belongs to the respective component, with values ranging from 0 to 1. For our experiment, the harmonic and percussive median filter lengths are 7 and 25, respectively; β is 2 and p is 4.

$$\mathbf{M}_{\mathbf{H}_{j,k}} = \frac{\mathbf{H}^{p}_{j,k}}{\left(\mathbf{H}^{p}_{j,k} + \beta \mathbf{P}^{p}_{j,k}\right)} \tag{3}$$

To obtain a smoother representation of the mask values, we applied median filtering to the mask $\mathbf{M_H}$ for each frequency bin, resulting in $\mathbf{M}^{'}_{\mathbf{H}}$. The mask values indicate the proportionate strength of the harmonic components. We only consider time-frequency masks with values greater than 0.65, while all others are set to 0. This operation is implemented in the thresholding block of Fig. 2 (b).

After thresholding is completed, we calculate the number of harmonic peaks detected in each time frame. The harmonic peaks are defined as the set of local maxima of the mask values for each time frame, determined by comparing each value of each sample with its neighboring samples. This operation is executed in the Peak Picking block. The total number of peaks for each frame is denoted by $\boldsymbol{H_{tp}}$, while the number of peaks above the minimum crossover frequency is denoted by $\boldsymbol{H_{cp}}$. Finally, we calculate HS using the formula $\boldsymbol{HS = H_{cp}/H_{tp}}$. For frames where the total number of peaks is less than 10, the HS is set to 0. The harmonic analysis provides two outputs, the harmonic spread and the total number of peaks for each frame.

C. Frame Classification

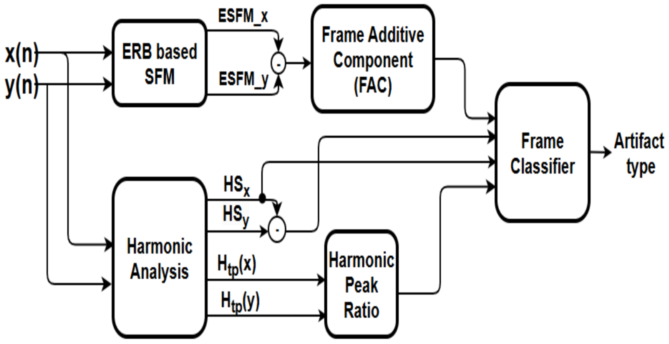


Fig. 3. Artifact identification framework

We have developed features to identify artifacts in any given frame, and incorporating all of these into a single framework is illustrated in Fig. 3. $x(n)$ and $y(n)$ represent the discrete reference signal in the time domain and the signal under test (SUT), respectively. The frame classifier takes the output from the Frame Additive Component (FAC) block as input. The difference in harmonic spread between the two signals, defined as $\boldsymbol{(HS_x - HS_y)}$, the harmonic spread of the reference signal, and the harmonic peak ratio (HPR) are also fed as inputs. HPR is defined as the ratio of the total number of harmonic peaks in the original signal to those in the SUT, expressed as $\boldsymbol{H_{tp}(x)/H_{tp}(y)}$.

The FAC block assigns a label to each frame based on the additive component introduced in the SUT. The difference in ESFM values indicates whether the SUT is more tonal or noisy because of the copying of spectral components. The difference values for each frame are analyzed from the 37^{th} to the 59^{th} ERB band, corresponding to a frequency range of 3kHz to 16kHz. This region is selected because BWE schemes operate primarily within this frequency range. Each of these ERB bands is classified as tonal or noisy. Equation (4) outlines the conditions for the classification of each ERB band for an arbitrary time index j and the ERB band i, where $diff\_ESFM_{j,i}$ represents the difference in ESFM values between $x(n)$ and $y(n)$. A small threshold value γ is introduced to ignore the ERB bands with very minimal differences, with γ set to 0.05.

$$\mathbf{ERB\_label}_{j,i} = \begin{cases} ERB\_tonal & \text{, if } diff\_ESFM_{j,i} > \gamma \\ ERB\_noisy & \text{, if } diff\_ESFM_{j,i} < -\gamma \\ \textit{No Label} & \text{, otherwise} \end{cases} \tag{4}$$

Once all the ERB bands are labeled for a frame, the final output label of the FAC block for an arbitrary time frame j is determined based on the conditions outlined in equation (5). If the number of labeled ERB bands is the same for tonal and noisy classifications, the label associated with the ERB band exhibiting the highest difference in ESFM value is selected as the FAC label output. Additionally, the overall additive component must be present in at least 5 ERB bands to be considered for artifact classification.

$$\mathbf{FAC}_j = \begin{cases} Tonal\text{, if } \#ERB\_tonal > \#ERB\_noisy \\ Noisy\text{, if } \#ERB\_tonal < \#ERB\_noisy \\ ERB\_label_{j,(max(diff\_ESFM_{j,i}))}\text{, if } same \end{cases} \tag{5}$$

In the final stage, the Frame Classifier block utilizes all the features discussed above to classify the artifact. The difference in HS values reflects how the high-frequency harmonic content of the SUT deviates from that of the reference signal. A moderate to large difference of HS values (greater than 0.2) indicates a lack of tonal components in the high-frequency region of the SUT. Additionally, if the FAC labels the frame as ‘Noisy’, the artifact is classified as NS. If both the HS values of the reference and the SUT are zero while the FAC block output is ‘Noisy’, it suggests that noise components have been added to the SUT. However, the perceptual impact of this noise substitution is less severe compared to instances where tonal components are replaced by noise. For frames that have a mild effect of noise substitution we classify them as NS_minor.

A large negative HS difference (greater than -0.2) indicates the presence of tonal components in the high-frequency region of the SUT. In such cases, if the FAC output for the frame label is‘Tonal’, then it leads to two possible artifacts: HM or ET. The harmonic spread of the reference signal is used as a discriminating feature. If the HS of the reference signal is zero, the artifact is identified as ET, since for ET artifact the reference signal does not show strong harmonic structure; otherwise, it is classified as HM.

TABLE I
SUMMARY OF FEATURE VALUES

HS_diff	FAC	HPR	HS_(x)	Artifact Type
> 0.2	Noisy	–	–	NS
0	Noisy	–	0	NS_minor
$\lvert HS\_diff \rvert <= 0.2$	Noisy	≥ 1	–	NS
$\lvert HS\_diff \rvert <= 0.2$	Noisy	< 1	–	HMNF
$\lvert HS\_diff \rvert <= 0.2$	Tonal	–	–	HM
< -0.2	Tonal	–	0	ET
< -0.2	Tonal	–	$\neq 0$	HM

The magnitude of the HS difference is also influenced by the crossover frequency. When the crossover frequency is around 3-4 kHz, the difference between the HS values is typically greater compared to when the crossover frequency ranges from 7-10 kHz. When the HS difference is small (with an absolute value less than 0.2) and the FAC output is ‘Noisy’, it becomes challenging to distinguish between noise substitution at higher crossover frequencies and harmonicity mismatch with noise-filling artifacts. In such cases, the HPR is used as a distinguishing feature. Noise substitution artifacts usually result in a significant reduction in the number of harmonics in the SUT compared to the original signal, leading to an HPR value greater than or equal to 1. In HMNF, since the harmonic structure is maintained in the SUT, the HPR value will be less than 1.

Table I summarizes how various feature values are used to identify the type of artifact.

IV. Experiment and Results

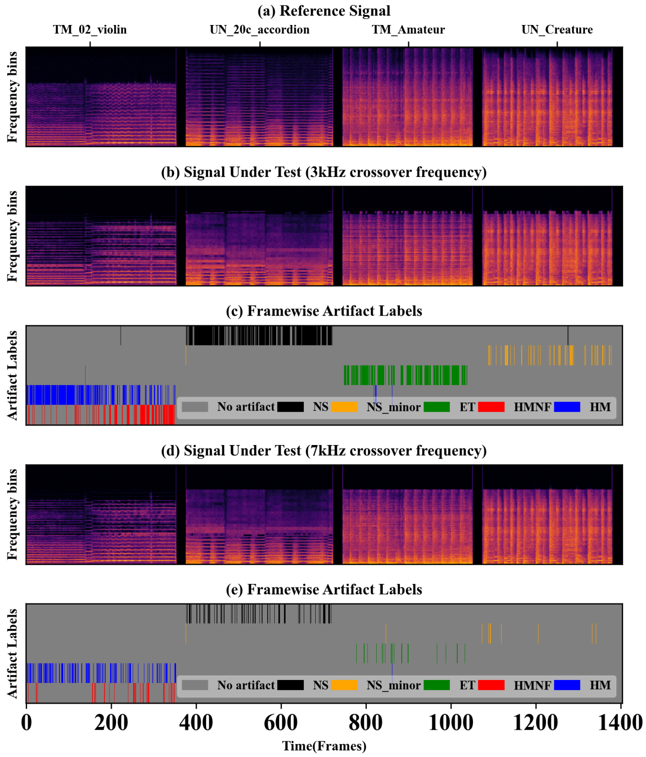
To validate our algorithm, we tested it on the publicly available Open Dataset of Audio Quality (ODAQ) [12]. From this dataset, we selected items affected by TM and UN artifacts. The items were generated for five different crossover frequencies; however, we focus on presenting our results for only two crossover frequencies (3kHz and 7kHz). At higher crossover frequencies, audible artifacts are reduced.

To demonstrate the use of the artifact detection algorithm, we created a custom audio file from the ODAQ dataset containing all artifacts. All signals were down-mixed to mono and sampled at 48kHz. The audio files considered include “TM_02_violin”, “UN_20c_accordion”, “TM_Amateur”, and “UN_Creature”. The item names are prefixed with the artifact type (TM, UN) as mentioned in [12]. Each segment is separated by 0.25 seconds of silence to clearly distinguish the different sections and their associated artifacts. The spectrogram representation of the resulting reference signal is shown in Fig. 4 (a), while Fig. 4 (b) displays the degraded version at a crossover frequency of 3 kHz.

Fig. 4 (c) presents the results of the artifact detection algorithm, where the graph shows the frames labeled with specific artifact types. The HM and HMNF artifacts are shown for the same audio excerpt i.e “TM_02_violin”. As a result of which the first segment contains frame labels belonging to the HM and HMNF artifact types. We further categorize the NS artifact by highlighting the severe and less severe noise substitution, marked as NS and NS_minor, respectively. This distinction is validated by the higher average MUSHRA [20] rating for “UN_Creature” compared to “UN_20c_accordion.”

For a higher crossover frequency of 7kHz, the resulting signal can be seen in Fig. 4 (d). The results of the artifact detection are shown in Fig. 4 (e). At these crossover frequencies, the perceptual quality improves, resulting in fewer frames being labeled with artifacts. This improvement in perceptual quality is corroborated by the increased mean MUSHRA rating of all individual items, as observed in [12].

Based on the dataset that we have used, the items affected by TM artifact would be classified as HM, HMNF or ET; and those affected by UN would be classified as NS or NS_minor. From the results, it can be seen that our algorithm incorrectly classifies only a single frame of TM artifact as a UN one at both the crossover frequencies.


(a) Reference Signal

(b) Signal Under Test (3kHz crossover frequency)

(c) Framewise Artifact Labels

(d) Signal Under Test (7kHz crossover frequency)

(e) Framewise Artifact Labels

Fig. 4. Framewise artifact identification of custom made audio file for crossover frequency of 3kHz and 7kHz.

V. Conclusion and Future work

Our work introduces a novel approach for identifying artifacts produced by bandwidth extension schemes, with a further classification of these artifacts based on the spectral structure of both the reference signal and the signal under test. Our method effectively detects multiple distinct artifacts within an audio file. In future work, this artifact identification framework has the potential to be integrated into an objective audio quality assessment system, thereby enhancing the accuracy and reliability of the quality metric.

References

[1] T. Painter and A. Spanias, “Perceptual coding of digital audio,” *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–515, 2000.

[2] Jürgen Herre and Sascha Dick, “Psychoacoustic models for perceptual audio coding—a tutorial review,” *Applied Sciences*, 2019.

[3] Sascha Dick, Nadja Schinkel-Bielefeld, and Sascha Disch, “Generation and evaluation of isolated audio coding artifacts,” *journal of the audio engineering society*, , no. 9809, october 2017.

[4] Chi-Min Liu, Han-Wen Hsu, and Wen-Chieh Lee, “Compression artifacts in perceptual audio coding,” *Trans. Audio, Speech and Lang. Proc.*, vol. 16, no. 4, pp. 681–695, May 2008.

[5] International Telecommunication Union (ITU-R), “Method for the objective measurement of perceived audio quality,” ITU-R Recommendation BS.1387-1, International Telecommunication Union, Geneva, Switzerland, 2001.

[6] International Telecommunication Union (ITU-T), “Methods for subjective determination of transmission quality,” ITU-T Recommendation P.800, International Telecommunication Union.

[7] herre juergen and dick sascha, “introducing the free web edition of the “perceptual audio coders – what to listen for” educational material,” *journal of the audio engineering society*, , no. 87, may 2023.

[8] Per Ekstrand, “Bandwidth extension of audio signals by spectral band replication,” 2002.

[9] J. Herre and M. Dietz, “Mpeg-4 high-efficiency aac coding [standards in a nutshell],” *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 137–142, 2008.

[10] Frederik Nagel and Sascha Disch, “A harmonic bandwidth extension method for audio codecs,” in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 145–148.

[11] disch sascha, niedermeier andreas, helmrich christian r., neukam christian, schmidt konstantin, geiger ralf, lecomte jérémie, ghido florin, nagel frederik, and edler bernd, “intelligent gap filling in perceptual transform coding of audio,” *journal of the audio engineering society*, , no. 9661, september 2016.

[12] M. Torcoli, C. W. Wu, S. Dick, P. A. Williams, M. M. Halimeh, W. Wolcott, and E. A. P. Habets, “ODAQ: Open dataset of audio quality,” in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Seoul, Korea, April 2024.

[13] A. Gray and J. Markel, “A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 22, no. 3, pp. 207–217, 1974.

[14] Brian R Glasberg and Brian C.J Moore, “Derivation of auditory filter shapes from notched-noise data,” *Hearing Research*, vol. 47, no. 1, pp. 103–138, 1990.

[15] J. Princen, A. Johnson, and A. Bradley, “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*, 1987, vol. 12, pp. 2161–2164.

[16] Nilesh Madhu, “Note on measures for spectral flatness,” *Electronics Letters*, vol. 45, pp. 1195–1196, 2009.

[17] Nobutaka Ono, Kenichi Miyamoto, Jonathan Le Roux, Hirokazu Kameoka, and Shigeki Sagayama, “Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram,” in *2008 16th European Signal Processing Conference*, 2008, pp. 1–4.

[18] Derry Fitzgerald, “Harmonic/percussive separation using median filtering,” in *13th International Conference on Digital Audio Effects (DAFX10)*, 2010.

[19] Jonathan Driedger, Meinard Müller, and Sascha Disch, “Extending harmonic-percussive separation of audio signals,” in *International Society for Music Information Retrieval Conference*, 2014.

[20] International Telecommunication Union (ITU-T), “Method for the subjective assessment of intermediate quality level of audio systems,” ITU-R Recommendation BS.1534-3 (10/2015), International Telecommunication Union.