A MDCT BASED HARMONIC SPECTRAL BANDWIDTH EXTENSION METHOD

Christian Neukam, Frederik Nagel

Fraunhofer IIS
Am Wolfsmantel 33
91058 Erlangen, Germany

Gerald Schuller, Michael Schnabel

Ilmenau University of Technology
Ehrenbergstr. 29
98693 Ilmenau, Germany

ABSTRACT

Modern audio coding technologies apply methods of bandwidth extension (BWE) to efficiently represent audio data at low bitrates. An established method is the well-known spectral band replication (SBR) that is part of MPEG High Efficiency Advanced Audio Coding (HE-AAC). However, if the signal features a distinct harmonic spectral structure, the use of these methods tends to result in audible artifacts, because the harmonic structure is not reconstructed correctly. In this paper a bandwidth extension method is proposed which eliminates the undesirable effects and allows for an efficient implementation in the Modified Discrete Cosine Transform (MDCT) domain. The proposed Harmonic Spectral Bandwidth Extension (HSBE) method uses arbitrary frequency shifts for modulating the replicated spectrum in a way that the harmonic structure of the signal is preserved. A listening test demonstrates the advantage of the proposed method compared to the state of the art.

Index Terms— Bandwidth extension, audio coding, MDCT, MDST

1. INTRODUCTION

Bandwidth extension is a standard technique within contemporary audio codecs to efficiently code audio signals at low bitrates. The main idea of bandwidth extension is to exploit correlations between the low frequency part (LF) and the higher frequencies (HF) of the signal [1]. Hence it is possible to code only the LF part of the signal with a core coder. The high frequency band is represented by additional side information parameters which are estimated for the HF band on encoder side. On the decoder side, the HF band is reproduced by shifting the LF band into the HF band with additional post processing steps. The newer methods Harmonic Bandwidth Extension (HBE) and Continuously Modulated Bandwidth Extension (CM-BWE) use either spectral stretching or arbitrary frequency shifts for the reproduced HF band to maintain the correct harmonic structure of the signal. In this way, artifacts for harmonic signals which could result from a fixed, non-adaptive frequency shift into the HF band can be avoided or reduced. CM-BWE uses bandpass filters to separate the LF and HF bands and single sideband (SSB) modulation in the time domain to obtain arbitrary frequency shifts to regenerate the HF band. To implement the SSB modulation, a 90 degree phase shifted version of the signal is needed, which is obtained with a Hilbert transform filter.

1.1. Goal

In this paper an algorithm is proposed to implement the CM-BWE method right in the MDCT domain of an audio coder. In this way, the computational complexity and the encoding/decoding delay due to the Hilbert transform can be reduced. It is also possible to take advantage of the steep bandpass filters of the MDCT to better separate the LF and HF bands.

1.2. Problem to Solve

To reach the mentioned goal some problems have to be solved:

The MDCT provides a frame length dependent frequency resolution. Here a solution has to be found to achieve a finer frequency shift than the bandwidth of one MDCT band.

The shift of the MDCT spectrum by arbitrary frequency shifts produces aliasing. Here a solution has to be found to reduce the amount of aliasing.

2. PRIOR WORK, STATE OF THE ART

The most commonly used BWE method is SBR as used in HE-AAC [2]. SBR uses a Pseudo-Quadrature Mirror Filter (PQMF) description of the signal and improves the compression efficiency of perceptual audio codecs [3]. This is achieved by simply copying the LF bands to the HF bands within the used filter bank, followed by post processing (inverse filtering, adaptive noise addition, sinusoidal regeneration, shaping of the spectral envelope) [4].

For harmonic signals e.g. a pitch pipe, SBR and equivalent BWE methods produce artifacts because the harmonic structure of the signal is not preserved. These artifacts are evoked by fixed frequency shifts to reproduce the HF bands. Hence, two BWE methods were developed to maintain the harmonic structure: the phase vocoder driven HBE [5] and the CM-BWE using single sideband modulation [6].

HBE employs a signal representation in either the DFT [5] or the PQMF [7] domain where the DFT-based HBE algorithm is more complex. The basic idea here is to spread the spectrum in such a way that the harmonic structure of the signal is preserved. In addition a transient handling is necessary in order to avoid pre- and post-echoes caused by the phase vocoder [8]. The CM-BWE uses a single sideband modulation, which uses an analytic signal. This signal is generated by a Hilbert transform [5], which introduces an additional signal delay. Here it is necessary to find a trade-off between quality and produced delay. The separation of the LF and HF bands is done with bandpass filters.

The proposed approach targets the reduction of the computational complexity and the encoding/decoding delay of CM-BWE.

3. NEW APPROACH: HARMONIC SPECTRAL BANDWIDTH EXTENSION

The basic idea is to implement the single sideband modulation of CM-BWE in the MDCT domain. As a critically sampled filter bank providing perfect reconstruction [9], the MDCT is the most popular transform in today's transform codecs like mp3 and AAC [10]. Since the MDCT is already computed in the core encoder, it is not necessary to compute a different filter bank, as for instance SBR does.

3.1. Notation

In this paper the following notation is assumed: lower case letters are used for time domain signals, like the input audio signal $x(n)$ with the sample index n. Bold face letters represent vectors and matrices, for instance the sequence of blocks of size N of the input audio signal:

$$\mathbf{x}(m) = [x(mN), x(mN+1), \ldots, x(mN+N-1)] \quad (1)$$

with N being the block size and also the number of subbands of the critical sampled MDCT, m being the block index at the lower, downsampled rate of the filter bank, k being the subband number. $\mathbf{y}_{\mathrm{k}}(m)$ denotes the k^{th} subband of the MDCT at the down-sampled time index m.

Capital letters denote z-transforms. For instance, the sequence of blocks of an audio signal is transformed into the z-domain using the block number m as time variable at the lower sampling rate:

$$\mathbf{X}(z) = \sum_{m=0}^{\infty} \mathbf{x}(m) \cdot z^{-m} \quad (2)$$

3.2. The MDCT in the z-Domain

If the introduced notation is applied, the MDCT analysis filter bank operation in the z-domain can be written as a matrix product with the analysis polyphase matrix $\mathbf{P}_{\mathrm{a}}(z)$:

$$\mathbf{Y}_{\mathrm{C}}(z) = \mathbf{X}(z) \cdot \mathbf{P}_{\mathrm{a}}(z) \quad (3)$$

where $\mathbf{Y}_{\mathrm{C}}(z)$ is the vector containing all subband signals from the MDCT analysis filter bank, in the z-domain. The MDCT with its modulation function has symmetries, which enable us to simplify the polyphase matrix $\mathbf{P}_{\mathrm{a}}(z)$ into a product of matrices. First a size $N \times N$ 'folding matrix' $\mathbf{F}_{\mathrm{C}}$ with entries mostly zero, except along a diamond shaped pattern in the matrix, as defined in [11]:

$$\mathbf{F}_{\mathrm{C}} = \begin{bmatrix} 0 & & h_{2\mathrm{N}-1} & h_{\mathrm{N}-1} & & 0 \\ & \iddots & & & \ddots & \\ h_{1.5\mathrm{N}} & & 0 & 0 & & h_{0.5\mathrm{N}} \\ h_{1.5\mathrm{N}-1} & & 0 & 0 & & -h_{0.5\mathrm{N}-1} \\ & \ddots & & & \iddots & \\ 0 & & h_{\mathrm{N}} & -h_0 & & 0 \end{bmatrix} \quad (4)$$

where the coefficients h_i are the samples of the baseband impulse response (the time-reversed window) of the MDCT. Second, the 'delay matrix' $\mathbf{D}(z)$ of size $N \times N$ is defined as follows:

$$\mathbf{D}(z) = \begin{bmatrix} z^{-1} & & & & & \\ & \ddots & & & 0 & \\ & & z^{-1} & & & \\ & & & 1 & & \\ & 0 & & & \ddots & \\ & & & & & 1 \end{bmatrix} \quad (5)$$

Finally the DCT-4 matrix $\mathbf{T}_C$ of size $N \times N$ is required.

Observe that the only one of these matrices with a dependence on z is $\mathbf{D}(z)$. A multiplication with z^{-1} in the z-domain corresponds to a delay of 1 sample in the time domain at the lower sampled block index m. Hence this is the only matrix with associated memory. Using these simpler matrices the MDCT can be written as follows:

$$\mathbf{Y}_{\mathrm{C}}(z) = \mathbf{X}(z) \cdot \mathbf{F}_{\mathrm{C}} \cdot \mathbf{D}(z) \cdot \mathbf{T}_{\mathrm{C}} \quad (6)$$

To obtain arbitrary frequency shifts independent from the subband structure of the MDCT, complex subbands values a required. This provides a 90 degree phase shift in the imaginary part with the complex MDCT/MDST, also known as CMDCT [12]:

$$\mathbf{Y}(z) = \mathbf{X}(z) \cdot (\mathbf{F}_{\mathrm{C}} \cdot \mathbf{D}(z) \cdot \mathbf{T}_{\mathrm{C}} + j\,(\mathbf{F}_{\mathrm{S}} \cdot \mathbf{D}(z) \cdot \mathbf{T}_{\mathrm{S}})) \quad (7)$$

with $\mathbf{T}_{\mathrm{S}}$ being the DST-IV matrix and $\mathbf{F}_{\mathrm{S}}$ being the folding matrix corresponding to the MDST. The difference between $\mathbf{F}_{\mathrm{C}}$ and $\mathbf{F}_{\mathrm{S}}$ is the sign change of the coefficients from $h_{1.5\mathrm{N}}$ to $h_{2\mathrm{N}-1}$ and from h_0 to $h_{0.5\mathrm{N}-1}$.

For arbitrary frequency shifts, the total shift is divided into a part with an integer multiple of a subband width and a fine tuned part within one subband. The fine tuned part can be obtained with the following modulation:

$$\mathbf{y}_{\mathrm{k,mod}}(m) = \mathbf{y}_{\mathrm{k}}(m) \cdot \mathrm{e}^{-\mathrm{j} \cdot m \cdot f_{\varphi} \cdot \pi} \quad (8)$$

with f_{φ} being the fine tuned modulation frequency, normalized to the bandwidth of one CMDCT subband. Since f_{φ} realizes a shift of the complex spectrum within a subband, this frequency is limited to the interval $f_{\varphi} \in [-0.5; 0.5)$.

For the frequency shift with integer multiples of the bandwidth of an MDCT band, the spectrum is shifted and copied the desired number of bands. If the integer shift corresponds to an even spectral distance the patched spectrum has to be multiplied with 1 otherwise with -1. This additional modulation corresponds to the reversed spectrum in every second subband of the MDCT.

The modulation frequency f_{φ} is calculated on the basis of the time varying fundamental frequency f_0 of the signal. It can be either calculated on encoder side using the full spectrum or on decoder side using only the core band. To avoid additional side information to be transmitted, one solution is to estimate f_0 in the decoder.

The basic principle of HSBE is illustrated in Figure 1. The first step is to copy the core band into the HF region (Figure 1.2). Afterwards the copied spectrum will be shifted by the two methods mentioned above and finally the spectral envelope will be reconstructed (Figure 1.4) in order to obtain a spectrum close to the original (Figure 1.1). The modulation frequencies are chosen such that the highest harmonic of the core band matches with the lowest harmonic of the replicated spectrum as shown in Figure 1.3.

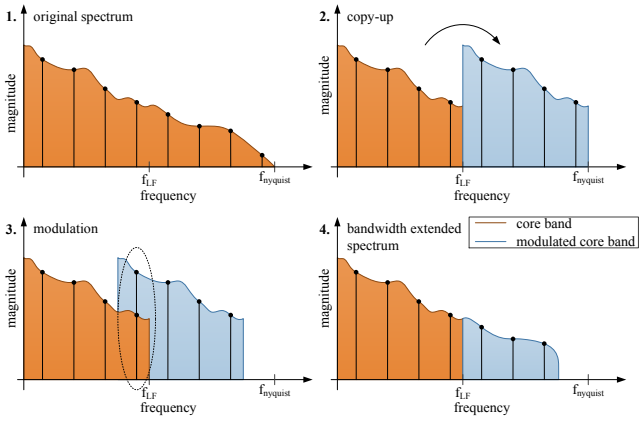

Fig. 1. Basic principle of HSBE

3.3. MDCT-MDST-transformation

The easiest way to get the MDST coefficients for the complex representation would be to simply transform the MDCT coefficients $\mathbf{Y}_{\mathrm{C}}$ back to the time domain and then subsequently do the forward MDST transform in order to derive the MDST transformed signal $\mathbf{Y}_{\mathrm{S}}$. As this would be quite complex, a more efficient method is proposed to use, which is explained in the following.

Taking the inverse MDCT transform, followed by a forward MDST transform, can be written in the matrix notation as follows:

$$\mathbf{Y}_{\mathrm{S}}(z) = \mathbf{Y}_{\mathrm{C}}(z) \cdot \mathbf{T}_{\mathrm{C}}^{-1} \cdot \mathbf{D}^{-1}(z) z^{-1} \cdot \mathbf{F}_{\mathrm{C}}^{-1} \cdot \mathbf{F}_{\mathrm{S}} \cdot \mathbf{D}(z) \cdot \mathbf{T}_{\mathrm{S}} \quad (9)$$

Observe the factor z^{-1}, which is to make $\mathbf{D}^{-1}(z)$ causal. At first a 'conversion matrix' $\mathbf{H}(z)$ according to [13] is defined to derive an efficient MDCT-to-MDST transform:

$$\mathbf{H}(z) = \mathbf{T}_{\mathrm{C}}^{-1} \cdot \mathbf{D}^{-1}(z) z^{-1} \cdot \mathbf{F}_{\mathrm{C}}^{-1} \cdot \mathbf{F}_{\mathrm{S}} \cdot \mathbf{D}(z) \cdot \mathbf{T}_{\mathrm{S}} \quad (10)$$

Next, the imaginary part is obtained from the real part with the following multiplication:

$$\mathbf{Y}_{\mathrm{S}}(z) = \mathbf{Y}_{\mathrm{C}}(z) \cdot \mathbf{H}(z) \quad (11)$$

Hence, just the conversion matrix $\mathbf{H}(z)$ has to be simplified. Since the delay matrix $\mathbf{D}(z)$ contains polynomials of 1^{st} order, a polynomial of 2^{nd} order for the conversion matrix is obtained:

$$\mathbf{H}(z) = H_0 z^0 + H_1 z^{-1} + H_2 z^{-2} \quad (12)$$

It turns out that H_1 is already a sparse matrix with only $2N$ coefficients. Only the matrices H_0 and H_2 have to be simplified. By a closer look at the values of these matrices, it can be seen that the most values are close to zero. Hence it is possible to reduce the amount of needed operations for calculating equation (11) by setting these smallest values to zero and to keep only the larger coefficients near the main diagonal. Although this simplification produces an error, an SNR of 70 dB can be achieved by using only 10 % of the coefficients per matrix.

3.4. Aliasing cancelation

In the last section a method is derived to shift a complex spectrum by arbitrary frequency shifts. But especially the shifts within one MDCT subband width compromises its aliasing cancelation properties.

Figure 2 shows the typical appearance of the aliasing components in the Fourier-spectrum after the inverse CMDCT of a shifted sinusoidal tone. It is noticeable that the undesired artifacts occur equidistant in every second band of the used filter bank. The magnitude of each aliasing component increases with higher values of the modulation frequency f_φ. Despite of the fact that no closed solution

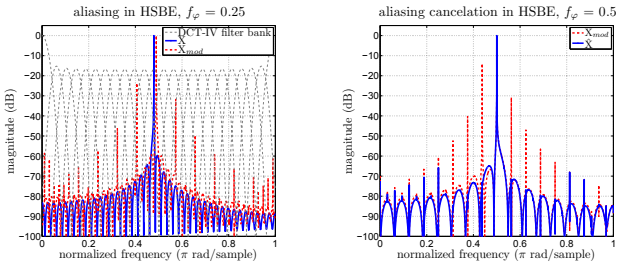

Fig. 2. left: Introduced aliasing by arbitrary frequency shifts, right: aliasing cancelation up to 4^{th} order aliasing terms

could be found to solve this problem, these aliasing terms can be canceled or reduced by the developed butterfly structure in Figure 3. According to this structure, each aliasing component will be reduced by a weighed sum of neighboring CMDCT bins.

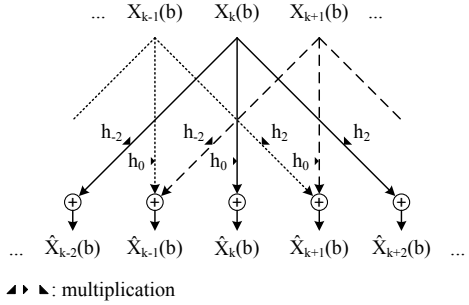

Fig. 3. Diagram of the anti-aliasing butterfly structure in HSBE

Since the magnitude of each aliasing component is dependent on the modulation frequency f_φ, the required coefficients h_α of the anti-aliasing butterfly are also frequency shift dependent. These coefficients are obtained once using an optimization algorithm, and then stored. The result of the aliasing cancelation is shown in Figure 2. In this example the aliasing components are canceled up to order $\alpha = 4$.

4. LISTENING TEST

For subjective evaluation of the perceptual quality, the proposed Harmonic Spectral Bandwidth Extension is compared to two other BWE methods, SBR and CM-BWE. For a better comparison of the bandwidth extension methods all techniques were applied to a pulse code modulated (PCM) signal without any data compression. SBR is the basis for the comparison, i.e. all the compared methods use and share the same methods from SBR for envelope shaping, inverse filtering, adaptive noise addition, sinusoidal regeneration. Only the patching algorithm itself, the reproduction of the HF band, is replaced for each method. So only the patch algorithms themselves are compared. For the test the SBR post processing is done with a separate PQMF filter bank, operating on the time domain signal.

The test consists of 10 items, 8 music items and 2 speech items listed in Table 1. The sample rate of all items is 24 kHz. The crossover frequency for the bandwidth extension (the high end of the LF band) is $f_{\mathrm{LF}} = 4\,\mathrm{kHz}$ and the signals are bandwidth extended by either SBR, CM-BWE or HSBE to 12 kHz. For HSBE the two mentioned methods for estimation of the fundamental frequency f_0 are compared. For 'HSBE (f0 encoder)' f_0 is estimated in the encoder, for 'HSBE (f0 decoder)' f_0 is estimated in the decoder.

12 expert listeners participated in the listening test, which was conducted using the MUSHRA methodology [14]. The test items were presented together with the original and one lowpass filtered anchor with the cut-off frequency being $f_{\mathrm{lp}} = 3.5\,\mathrm{kHz}$. The sound was reproduced via Stax headphones.

item	description
brahms	classical music
es01	voiced speech
fanfare	brass ensemble
judas	violin
phi7	pitch pipe
sm01	bag pipe
sm03	harpsichord
te15	classical music
te1	speech
vivaldi	classical music

Table 1. List of testitems

5. RESULTS

Difference ratings with its means and confidence intervals were calculated for every listener and every item pairwise between SBR, CM-BWE and HSBE. Figure 4 presents the results. The statistical evaluation is based on a non-parametric Wilcoxon signed-rank test since a standard normal distribution of the ratings cannot be considered. The results show that HSBE is significant better $(p < .05)$ than SBR for 3 items. For the other 7 items there is no statistical significance, but HSBE tends to outperform SBR. In comparison to CM-BWE the proposed method is significant better $(p < .05)$ for 6 items. For the other 4 items there is no significant difference. The overall performance of HSBE is significant better than SBR and than CM-BWE. The perceived quality of HSBE driven with the estimation of f_0 either on encoder side or on decoder side is very similar as the results show.

6. DISCUSSION

The discussion of the results is based on the fact that all BWE methods are applied on a PCM signal. So the shown results are not in conflict with the results in [6] where the core band was coded with USAC.

The proposed bandwidth extension method HSBE has an improved performance compared to CM-BWE in several points. At first, HSBE outperforms CM-BWE in perceived audio quality as the results of the listening test show. Second, the computational complexity and the produced delay decrease. The resulting signal delay of the proposed algorithm decreases by a factor of 2.7 from CM-BWE. The Hilbert transform for generating the analytic signal produces the major part of the resulting delay in CM-BWE. As this is not needed in HSBE, the delay reduces to one frame due to the

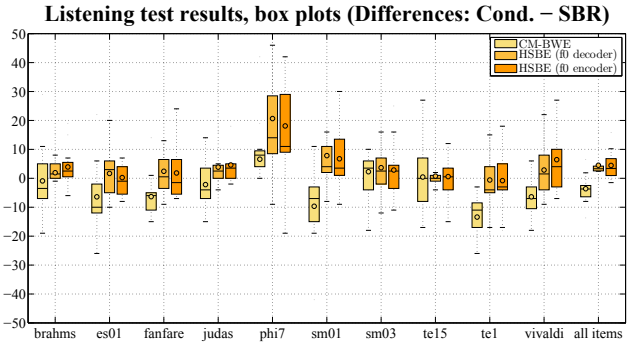

Fig. 4. Listening test results: Box plots of differences of individual ratings of HSBE - SBR and CM-BWE - SBR. HSBE outperforms SBR significantly $(p < .05)$ in 3 of 10 cases (significant positive values) and overall means for both f_0 estimation methods.

MDCT-to-MDST transform. The comparison of the complexity is done by evaluating the number of required additions and multiplications for both BWE methods. For CM-BWE only the signal processing for the SSB modulation and the Hilbert transform is taken into account while for HSBE the mentioned techniques plus an additional inverse MDST are taken into account. For both methods it is considered that the fundamental frequency f_0 is known a priori. It turns out that the number of required operations per frame of the proposed algorithm decreases by a factor of 19.2 from CM-BWE.

7. CONCLUSION

The work presented here shows, that it is possible to perform harmonic bandwidth extension in the MDCT domain. An algorithm is derived to shift the spectrum by arbitrary frequencies. Occurring problems due to the modulation of the complex CMDCT spectrum within the bandwidth of one MDCT band are canceled by applying an anti-aliasing butterfly structure. In comparison to CM-BWE the proposed technique features less delay and reduced computational complexity. The conducted listening test shows that the proposed MDCT approach does not fall back in perceived audio quality compared to SBR or CM-BWE. For some items it even performs significantly better. So the goal of getting a functional and efficient implementation of CM-BWE in the MDCT domain is achieved.

8. ACKNOWLEDGMENT

At the end of this paper we would like to thank Sascha Disch and Ralf Geiger for reviewing the paper, Stephan Wilde for his support and the listeners for participating the listening test and their patience.

9. REFERENCES

[1] E. Larsen and R. M. Aarts, *Audio Bandwidth Extension: Application of Psychoacoustics, Signal Processing and Loudspeaker Design*, John Wiley & Sons Ltd., Chichester, 2004.

[2] ISO/IEC 14496-3:2005, *Information technology - Coding of audio-visual objects - Part 3: Audio*, Geneva, 2005.

[3] M. Dietz, L. Liljeryd, K. Kjorling, and O. Kunz, "Spectral Band Replication, a Novel Approach in Audio Coding," in *Audio Engineering Society Convention 112*, Munich, April 2002.

[4] P. Ekstrand, "Bandwidth Extension of Audio Signals by Spectral Band Replication," in *Proceedings of 1st IEEE Benelux Workshop on MPCA*, Leuven, November 2002, vol. 1.

[5] F. Nagel and S. Disch, "A harmonic bandwidth extension method for audio codecs," in *Acoustics, Speech and Signal Processing (ICASSP), 2009 IEEE International Conference on*, April 2009.

[6] F. Nagel, S. Disch, and S. Wilde, "A continuous modulated single sideband bandwidth extension," in *Acoustics, Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, March 2010.

[7] H. Zhong, L. Villemoes, P. Ekstrand, S. Disch, F. Nagel, S. Wilde, K. S. Chong, and T. Norimatsu, "QMF Based Harmonic Spectral Band Replication," in *Audio Engineering Society Convention 131*, New York, 10 2011.

[8] F. Nagel, S. Disch, and N. Rettelbach, "A Phase Vocoder Driven Bandwidth Extension Method with Novel Transient Handling for Audio Codecs," in *Audio Engineering Society Convention 126*, Munich, 5 2009.

[9] J. Princen, A. Johnson, and A. Bradley, "Subband/Transform coding using filter bank designs based on time domain aliasing cancellation," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '87.*, April 1987, vol. 12.

[10] K. Brandenburg, "MP3 and AAC Explained," in *Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding*, August 1999.

[11] G. D. T. Schuller and M. J. T. Smith, "New framework for modulated perfect reconstruction filter banks," *Signal Processing, IEEE Transactions on*, vol. 44, no. 8, August 1996.

[12] M. Mathew, V. Bhat, S.M. Thomas, and Changhoon Y., "Modified MP3 encoder using complex modified cosine transform," in *Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on*, july 2003, vol. 2.

[13] G. Schuller, M. Gruhne, and T. Friedrich, "Fast audio feature extraction from compressed audio data," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, no. 6, Oktober 2011.

[14] ITU-R BS.1534-1, *Method for subjective assessment of intermediate sound quality (MUSHRA)*, Geneva, 2003.