ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747575

DEEP NEURAL NETWORK (DNN) AUDIO CODER USING A PERCEPTUALLY IMPROVED TRAINING METHOD

Seungmin Shin[†], *Joon Byun*[†], *Youngcheol Park*[†], *Jongmo Sung*[‡], *Seungkwon Beack*[‡]

[†] Intelligent Signal Processing Lab., Yonsei University, Wonju, Korea
[‡]Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea

ABSTRACT

A new end-to-end audio coder based on a deep neural network (DNN) is proposed. To compensate for the perceptual distortion that occurred by quantization, the proposed coder is optimized to minimize distortions in both signal and perceptual domains. The distortion in the perceptual domain is measured using the psychoacoustic model (PAM), and a loss function is obtained through the two-stage compensation approach. Also, the scalar uniform quantization was approximated using a uniform stochastic noise, together with a compression-decompression scheme, which provides simpler but more stable learning without an additional penalty than the softmax quantizer. Test results showed that the proposed coder achieves more accurate noise-masking than the previous PAM-based method and better perceptual quality then the MP3 audio coder.

Index Terms— DNN-based Audio Coder, PAM, Perceptual Loss Function

1. INTRODUCTION

Although deep neural network (DNN) technology was successfully introduced to speech and audio processing applications, improving the perceptual quality without increasing the complexity of the DNN has always been an issue of importance. As an effort to resolve this, it was tried to use the human auditory perception models for designing trainable loss functions mainly in the forms of partial perceptive models [1, 2, 3]. For speech and audio coding, in particular, excellent perceptual sound quality could be achieved using the loss function designed based on the psychoacoustic model (PAM) [4, 5, 6]. This approach exploits the irrelevance of masked signal components, which is key to achieving a significant reduction in bitrate while preserving subjective audio quality. However, the masking of quantization noise still needs to be improved to attain perceptual transparancy.

In speech/audio coding, entropy implies the minimum bitrate achievable with lossless entropy coding. The DNN is trained to transform the input to latent features with the smallest possible entropy under a certain level of distortion, which results in a non-trivial rate-distortion (R-D) optimization problem. One more issue associated with the DNN coder is how to approximate the non-differentiability of quantizer. Various methods have been proposed to relax the R-D optimization comprising the nonlinear quantization process [3, 7, 8, 9]. Among those, the so-called softmax quantization [3, 9] demonstrated effectiveness in a trainable end-to-end audio coder based on DNN. However, this approach requires a penalty to create a proper latent vector [10] and an annealing process to convert a soft assignment to a hard one. On the other hand, in image coding, an efficient R-D optimization was achieved by replacing the quantization process with uniform stochastic noise [11, 12]. Since neither a penalty nor additional processing is required, it allows a more intuitive and better approach to training the DNN coder than the softmax.

This paper proposes an end-to-end DNN audio coder trained using a perceptually motivated loss function. Distortions for the R-D optimization are measured in both signal and perceptual domains to implement a perceptual compensation. The loss in the perceptual domain is designed using PAM as in the previous study [6]. However, it is further elaborated through the two-stage compensation approach: the noise-to-mask-ratio (NMR) function is rescaled first, and perceptually important bands are emphasized using perceptual entropy (PE) later. Also, to relax the nonlinearity issue of the quantizer, we adopt the method used for image coding: replacing the quantization process with a uniform stochastic noise, with an additional compression-decompression scheme. Through tests, we confirm the superior noise-masking performance of the proposed DNN coder to both the previous PAM-based coder and better subjective quality than the conventional MP3 audio coder.

2. DNN AUDIO CODER

2.1. DNN Model

An autoencoder type time-domain end-to-end audio/speech coder was constructed. A block diagram of the constructed coder is shown in Fig. 1. Input and output vectors consist

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. [21ZH1200, The research of the basic mediacontents technologies]

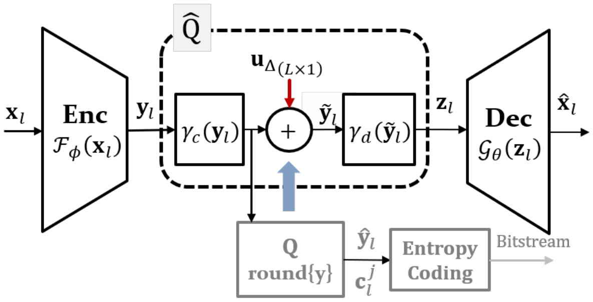


Fig. 1. Schematic diagram of the training model for the proposed DNN audio coder.

Table 1. Architecture of the ResGLU-based encoding network.

Layer	Input shape	Kernel shape	Output shape
Channel change	(512, 1)	(9, 1, 100)	(512, 100)
ResGLU	(512, 100)	$\left.\begin{matrix}(9,100,30)\\(9,30,30)\\(9,30,30)\\(9,30,100)\end{matrix}\right\}\times 6$	(512, 100)
Downsampling	(512, 100)	(9, 100, 100)	(256, 100)
ResGLU	(256, 100)	$\left.\begin{matrix}(9,100,30)\\(9,30,30)\\(9,30,30)\\(9,30,100)\end{matrix}\right\}\times 6$	(256, 100)
Channel change	(256, 100)	(9, 100, 1)	(256, 1)

of T time samples, i.e., $\{\mathbf{x}_l, \hat{\mathbf{x}}_l\} \in \mathbb{R}^T$, where l denotes the frame index. The encoder transforms the input vector to a latent vector with a size of $T/2$: $\mathcal{F}_\phi(\mathbf{x}_l) \rightarrow \mathbf{y}_l \in \mathbb{R}^{T/2}$. The latent vector $\mathbf{y}_l$ is then quantized and losslessly compressed using entropy coding techniques such as arithmetic coding. However, in this study, entropy coding is not realized. The decoder recovers the input audio as $\mathcal{G}_\theta(\mathbf{z}_l) \rightarrow \hat{\mathbf{x}}_l$.

For the encoding and decoding networks, we adopted a ResNet-type gated linear unit (ResGLU) [13, 14]. Detailed dimensions of the encoding network are summarized in Table 1. The decoding network has an inverted form of the encoding network. The encoding network consists of six layers of ResGLUs implementing 1D convolution with dilation factors 2, 4, 8, 16, and 32, respectively. The channel changer is a convolutional block, transforming the single-channel input to multiple channel features or vice versa. As a result, the constructed DNN model in Fig. 1 consists of $1.8M$ trainable network parameters in total. The input frame has a size of $512 (= T)$ samples, and it is obtained using a window with half-sines at both ends. The frame overlap is 32 samples that are weighted using the half-sines.

2.2. Quantization Model

Previously, softmax scalar quantizer was popularly adopted in DNN-based speech and audio coders [3, 4, 5, 6]. The softmax quantizer relaxes the nonlinear quantization process through a soft assignment based on the softmax function. The softmax quantizer has shown reasonable learning characteristics and nice performance as well. However, it is known that the soft assignment process can not create a proper latent vector without a penalty term in the loss function [3, 10].

On the other hand, in image coding, replacing the uniform quantizer with an additive stochastic noise was popularly adopted because, if the latent features are quantized uniformly, the quantization error can be modeled as a uniform distribution. Since this method doesn't involve a soft-to-hard annealing process nor an additional penalty for obtaining latent vectors, it provides a more intuitive approach to quantization approximation. In this study, we adopted this approach for training the constructed DNN audio coder.

The quantization model used in this study is shown in Fig. 1, as the block $\hat{Q}$. First, the encoder $\mathcal{F}_\phi$ transforms the input to a latent vector $\mathbf{y}_l$. Then, a uniform stochastic noise is added to the latent vector after compression $\gamma_c(\cdot)$, which in turn is delivered to the decoding network $\mathcal{G}_\theta$ after decompression $\gamma_d(\cdot)$. The purpose of the compression and decompression is to limit the maximum range of the input samples of the quantizer and recover the original range through dequantization. We use the $\tanh$ function for the compression and decompression, i.e.,

$$\gamma_c(\mathbf{y}_l) = \tanh(a\mathbf{y}_l), \;\; \gamma_d(\tilde{\mathbf{y}}_l) = \frac{1}{a}\tanh^{-1}(b\tilde{\mathbf{y}}_l) \qquad (1)$$

where a and b are scalers. The scaler b, in particular, is to prevent the input to the $\tanh^{-1}$ function from being greater than 1 due to adding the uniform noise. Since the magnitude of latent vector $\mathbf{y}_l$ is compressed into the range of ± 1 by the $\tanh$ compressor, the scalar quantization with a step-size Δ is approximated using a noise $\mathbf{u}_\Delta$ with a uniform distribution on $[-\frac{\Delta}{2}, +\frac{\Delta}{2}]$. In this study, we use a 32-level uniform quantizer so that $\Delta = \frac{1}{32}$.

In the test stage, the actual quantization is performed by the element-wise rounding to the nearest integer, as $\mathrm{round}(\gamma_c(\mathcal{F}_\phi(\mathbf{x}_l)))$. Then, we obtain one-hot code vectors $\mathbf{c}_l^j \in \mathbb{R}^{(T/2)\times 2^B}$ where B is the bit-dimension of the code vector, i.e., $B = 5$ when a 32-level uniform quantizer is used. The major advantages of this approach are simple structure and easy training. Also, unlike the softmax quantizer, it is not subject to an additional penalty for proper latent vectors.

3. DNN TRAINING

3.1. R-D loss function

The error introduced by the quantization is tolerated under the limited bitrate condition, which creates a rate-distortion optimization problem. To solve this problem, the DNN encoder $\mathcal{F}_\phi$ and decoder $\mathcal{G}_\theta$ are trained to minimize the rate-distortion being expressed using a loss function:

$$\mathcal{L} = \mathcal{R} + \lambda\mathcal{D}, \qquad (2)$$

where $\mathcal{R}$ is a term comprising the discrete entropy of the quantized feature vector, and $\mathcal{D}$ is a term corresponding to the distortion measured using the original and reconstructed audio vectors. λ is a combination parameter. The rate term can be estimated using the entropy [3] as $\mathcal{R} = -\sum_i p_i \log_2 p_i$ where p_i is the probability mass function at the quantized level i. For practical consideration, p_i's are approximated using the normalized histogram of the noise-added latent features contained in $\tilde{\mathbf{y}}_l$. [11, 12],

Distortions are measured both in the signal domain and perceptual domain, and combined as

$$\mathcal{D} = \mathcal{D}_s + \alpha\mathcal{D}_p. \tag{3}$$

As a distortion in the signal domain, a mean-square error (MSE) between the reconstructed and original signals is used: $\mathcal{D}_s = \|\hat{\mathbf{x}}_l - \mathbf{x}_l\|_2^2$. The perceptual domain distortion is measured as $\mathcal{D}_p = d\left(\mathcal{P}(\hat{\mathbf{x}}_l), \mathcal{P}(\mathbf{x}_l)\right)$ where $\mathcal{P}$ and $d(\cdot,\cdot)$ denote a perceptual transform and distance metric, respectively.

3.2. Perceptual domain loss

In previous studies [5, 6], psychoacoustic model (PAM) was used as a perceptual transform $\mathcal{P}$. Based on PAM, a loss function was achieved using the log noise-to-mask ratio (NMR), defined as the difference between the quantization noise power and the global masking threshold (GMT) in the logarithmic scale. In [5], a priority weighting scheme was jointly combined with an NMR loss in the linear scale. In [6], NMR in the logarithmic scale was weighted using perceptual entropy (PE) and integrated on multi-scale Mel-frequency bands. Both approaches aimed the DNN coder to focus on the frequency bands where the quantization noise exceeds the masking level. However, experimental results show that the PAM-based loss function in [5] often fails to mask the quantization noise properly, which results in audible distortions. Thus, in this paper, we propose an improved PAM-based loss function as a measure of the perceptual domain distortion. The loss function proposed in this paper closely follows the one in [6]. But it is further elaborated for better control of the quantization noise.

The proposed method works in two steps: first, the quantization noise spectrum is rescaled according to the masking threshold, and second, the perceptually important bands are further emphasized using perceptual entropy (PE). In the first step, frequency-dependent perceptual weighting factors are calculated using NMR computed as $\mathbf{N}_p = \mathbf{C}_n - \mathbf{C}_t$, where $\mathbf{C}_n$ and $\mathbf{C}_t$ denote the quantization noise power and GMT in the logarithmic scale, respectively. Then, a scaling vector is obtained as

$$\mathbf{S}_p = \mathbf{N}_p - \bar{N}_p, \tag{4}$$

where $\bar{N}_p$ denotes the mean of elements in the vector $\mathbf{N}_p$. Now, $\mathbf{S}_p$ indicates relative perceptual vulnerability, meaning that the quantization noise power is more likely to cross GMT. Thus, the DNN coder must focus more on such perceptually vulnerable frequencies. To implement this, we use a method of inversely rescaling the quantization noise power as much as $10^{-\mathbf{S}_p/10}$, which can be simplified in the logarithmic scale:

$$\hat{\mathbf{C}}_n = \mathbf{C}_n - \mathbf{S}_p. \tag{5}$$

Then, the quantization noise powers that cross the masking threshold are collected into a vector:

$$\mathbf{M}_p = \max\left\{\left(\hat{\mathbf{C}}_n - \mathbf{C}_t\right), \mathbf{0}\right\}, \tag{6}$$

where $\mathbf{0}$ is a zero vector.

In the second step, perceptually important frequencies are selected using PE, as in the previous study [6]. Thus, a weighting vector is determined in a normalized form as

$$\mathbf{w}_{PE} = \left(\frac{\mathcal{E}}{\|\mathcal{E}\|_\infty}\right)^\gamma, \tag{7}$$

where $\mathcal{E}$ is the vector comprising PEs computed in each frequency bin, $\|\cdot\|_\infty$ denotes the maximum norm and $\gamma \geq 0$ is a constant for controlling the relative ratio to the maximum PE.

Also, as noted in the previous studies [3, 6], simultaneously providing coarse and detailed spectral features helps the DNN model learn the perceptual attributes better. Thus, all parameters in $\mathbf{N}_p$ and $\mathbf{w}_{PE}$ are transformed into Mel-frequency spectra of multiple band-resolutions, and the final loss in the perceptual domain is obtained by summing Mel-frequency parameters of all band resolutions as

$$\mathcal{D}_p = \frac{1}{K}\sum_{i=1}^{K}\left\|\mathbf{w}_{PE}^i \odot \mathbf{M}_p^i\right\|_1, \; i = 1, ...K, \tag{8}$$

where the index i denotes the i-th Mel filterbank having B_i bands resolution, and K is the total number of considered Mel filterbanks.

4. EXPERIMENTAL RESULTS

4.1. Test setup

Tests were conducted using an audio dataset. We construct the dataset using $1,000$ commercial music clips using a $32kHz$ sampling rate. 700 clips were selected for training, 200 clips and another 100 clips were chosen to create the validation and test sets, respectively. All sources were normalized with standard deviations. The mini-batch size was 128 throughout the training procedure. Learning rates were adjusted using cosine annealing between 0.0002 and 0.0001. The model was trained in PyTorch using Adam optimizer [15]. We used initial values of $\lambda = 60$ and $\alpha = 0.00005$ but dynamically adjusted through training until the DNN coder reaches the pre-set range of bitrate. We determined the number of Mel filterbanks (K) through informal listening tests, and the best result was obtained when the four Mel filterbanks of 16, 32, 64, 256 were jointly used. We also set the parameter γ in Eq. (7) as 2.4, which produced the highest perceptual quality.

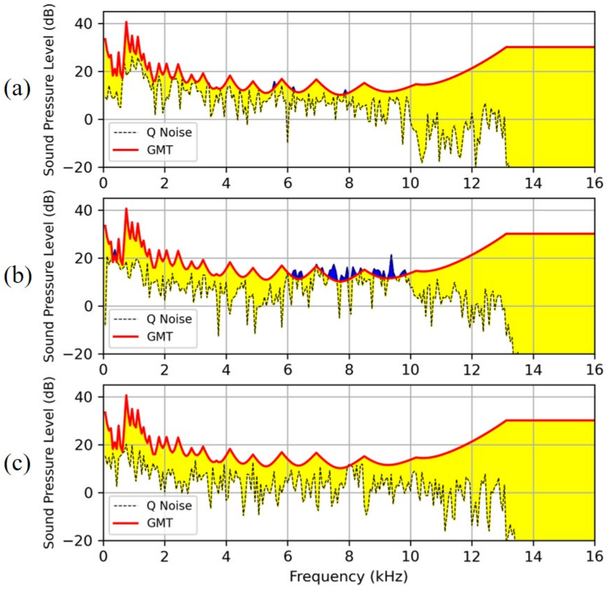


Fig. 2. Quantization noise comparison at $56kbps$: (a) MP3 audio coder, (b) the method in [5], (c) proposed. The signal bandwidth was limited by $13kHz$.

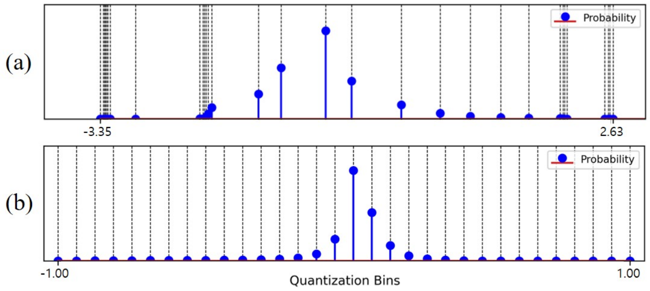
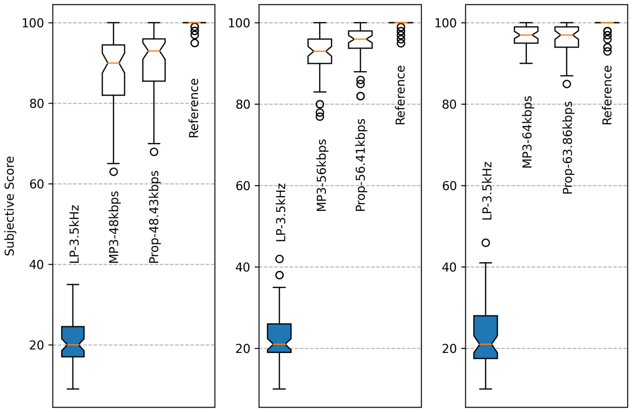


Fig. 3. Histograms of the assigned codes using a 32-cluster/level quantizer ($B = 5$): (a) softmax quantizer [6], (b) white noise model.

4.2. Noise masking and quantizaion performance

We compared the masking performance between the proposed loss function in Eq. (8) and the one in [5]. After training the DNN coder for 200 epochs using the respective loss functions, quantization noises were measured using a test audio clip whose bandwidth was limited to $13kHz$. The target bitrate was $56kbps$. Results are shown in Fig. 2. As a reference, the quantization noise of an MP3 audio coder used in the MUSHRA test is also presented. MP3 shows that it can accurately control the quantization noise along the masking curve, but the margin is small. The proposed method shows a sufficient margin between the noise spectrum and the masking curve, while the method in [5] fails to control the noise spectrum below down to the masking curve, especially in mid-frequency bands.



(a) $48kbps$ (b) $56kbps$ (c) $64kbps$

Fig. 4. MUSHRA test results.

In Fig. 3, we compare histograms of the assigned code vectors $\mathbf{c}_l^j$ using the quantization model in Section 2.2, and the softmax model [6]. Both models converged adequately to the target bitrate. One particular note about the softmax quantizer is that it continuously increases magnitudes of latent features due to updating the quantization clusters.

4.3. Subjective test results

The MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) [16] test was conducted. The proposed DNN audio coder was compared with the existing MP3 audio coder at $48kbps$, $56kbps$, and $64kbps$, respectively. Nine experienced subjects participated in the test, and seven test clips were selected from different genres. We used the commercial MP3 codec from Adobe Audition, licensed from Fraunhofer IIS and Thomson. Test results are shown in Fig. 4. Both the proposed and MP3 coders show almost transparent quality at $64kbps$. Several listeners reported that they had difficulties in identifying the original in the $64kbps$ test. At lower bitrates, $48kbps$ and $56kbps$, the proposed coder obtains higher scores than the MP3 coder. Although MP3 limits the signal bandwidth when the bitrate is not sufficiently high, occasional artifacts were still perceptible. In summary, the proposed DNN audio coder shows almost transparent sound quality at $64kbps$ and better quality than MP3 at $48kbps$ and $56kbps$.

5. CONCLUSIONS

We proposed an end-to-end DNN audio coder optimized using a rate-distortion functional. Distortions in the signal and perceptual domains were used to train the DNN coder, and the quantization process was approximated using a uniform stochastic noise. Test results show the superior performance of masking the quantization noise and better subjective quality than an MP3 audio coder. Test samples used in this study are available online[1].

[1] https://sites.google.com/view/isplab-yonsei/research/listening

6. REFERENCES

[1] Szu-Wei Fu, Yu Tsao, Hsin-Te Hwang, and Hsin-Min Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on blstm," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2018, pp. 1873–1877.

[2] Juan Manuel Martin-Doñas, Angel Manuel Gomez, Jose A. Gonzalez, and Antonio M. Peinado, "A deep learning loss function based on the perceptual evaluation of the speech quality," in *IEEE Signal Processing Letters*, 2018, vol. 25, pp. 1680–1684.

[3] Srihari Kankanahalli, "End-to-end optimized speech coding with deep neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2521–2525.

[4] Kai Zhen, Mi Suk Lee, Jongmo Sung, Seungkwon Beack, and Minje Kim, "Efficient and scalable neural residual waveform coding with collaborative quantization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 361–365.

[5] Kai Zhen, Mi Suk Lee, Jongmo Sung, Seungkwon Beack, and Minje Kim, "Psychoacoustic calibration of loss functions for efficient end-to-end neural audio coding," in *IEEE Signal Processing Letters*, 2020, vol. 27, pp. 2159–2163.

[6] Joon Byun, Seungmin Shin, Youngcheol Park, Jongmo Sung, and Seungkwon Beack, "Development of a psychoacoustic loss function for the deep neural network (DNN)-based speech coder," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2021, pp. 1694–1698.

[7] Song Han, Huizi Mao, and William J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," in *4th International Conference on Learning Representations (ICLR)*, 2016.

[8] Yoojin. Choi, Mostafa. El-Khamy, and Jungwon. Lee, "Towards the limit of network quantization," in *5th International Conference on Learning Representations (ICLR)*, 2017.

[9] Eirikur Agustsson, Fabian Mentzer, Michael Tschannen, Lukas Cavigelli, Radu Timofte, Luca Benini, and Luc Van Gool, "Soft-to-hard vector quantization for end-to-end learning compressible representations," in *31st Conference on Neural Information Processing Systems (NIPS)*, Dec. 2017, pp. 1141–1151.

[10] Eirikur Agustsson and Lucas Theis, "Universally quantized neural compression," in *34th Conference on Neural Information Processing Systems (NIPS)*, 2020.

[11] Johannes Balle, Valero Laparra, and Eero P. Simoncelli, "End-to-end optimization of nonlinear transform codes for perceptual quality," in *IEEE Picture Coding Symposium (PCS)*, 2016.

[12] Johannes Balle, Valero Laparra, and Eero P. Simoncelli, "End-to-end optimized image compression," in *International Conference on Learning Representations (ICLR)*, 2017.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[14] Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier, "Language modeling with gated convolutional networks," in *Proceedings of the 34th International Conference on Machine Learning Research (PMLR)*, 2017, pp. 933–941.

[15] Diederik P. Kingma and Jimmy Lei Ba, "Adam: A method for stochastic optimization," in *arXiv preprint arXiv:1412.6980*, 2014, vol. 19, pp. 2046–2057.

[16] ITU-R Recommendation BS 1534-1, *Method for the subjective assessment of intermediate quality levels of coding systems (MUSHRA)*, 2003.