INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Development of a Psychoacoustic Loss Function for the Deep Neural Network (DNN)-Based Speech Coder

Joon Byun[1], *Seungmin Shin*[1], *Youngcheol Park*[1], *Jongmo Sung*[2], *Seungkwon Beack*[2]

[1]Intelligent Signal Processing Lab., Yonsei University, Wonju, Korea
[2]Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea
{bj9407, nicesin97, young00}@yonsei.ac.kr, {jmseong, skbeack}@etri.re.kr

Abstract

This paper presents a loss function to compensate for the perceptual loss of the deep neural network (DNN)-based speech coder. By utilizing the psychoacoustic model (PAM), we design a loss function to maximize the mask-to-noise ratio (MNR) in multi-resolution Mel-frequency scales. Also, a perceptual entropy (PE)-based weighting scheme is incorporated onto the MNR loss so that the DNN model focuses more on perceptually important Mel-frequency bands. The proposed loss function was tested on a CNN-based autoencoder implementing the softmax quantization and entropy-based bitrate control. Objective and subjective tests conducted with speech signals showed that the proposed loss function produced higher perceptual quality than the previous perceptual loss functions.

Index Terms: Loss Function, DNN-based Speech Coder, PAM, CNN-based Autoencoder

1. Introduction

In recent years, various speech and audio signal processing algorithms based on deep neural network (DNN) models have shown promising improvements over the classical approaches [1, 2, 3, 4, 5]. Despite the great potentials of such DNN-based models, a major challenge remaining is improving the perceptual quality of the processed speech and audio. Thus, designing perceptually meaningful loss functions is an issue of importance. Traditionally, the mean squared error (MSE) is the most commonly used loss function to train DNN models. However, the MSE alone cannot faithfully reflect human auditory perception, and even if a high signal-to-noise ratio (SNR) is successfully achieved, low perception quality is often measured.

In the previous speech processing studies, efforts have been made to obtain trainable loss functions that reflect human auditory perception. For speech enhancement, metrics such as short-time objective intelligibility (STOI) [6] and perceptual assessment of speech quality (PESQ) [7, 8] were popularly considered for inclusion in the loss function. On the other hand, speech coding is associated with the quantization process, which is not directly differentiable. To achieve a differentiable estimate, various methods have been proposed to approximate the quantization process [9, 10, 11]. Among those, the so-called softmax quantization [12] based on the soft relaxation scheme of quantization [11], combined with a convolutional neural network (CNN) autoencoder, demonstrated its effectiveness for the trainable end-to-end speech coding model [13].

Also, by augmenting the Mel frequency spectra-based loss to the total loss function, it achieved higher PESQ scores than the Adaptive Multi-Rate Wideband (AMR-WB) [14] speech coder. But there is more potential for further improvements in perceptual quality since its outputs are still not close to the perceptual transparency.

This paper proposes a new loss function that can calibrate the perceptual quality for the DNN-based speech coder, where the input and reconstructed output speech frames can be synchronized. To this end, we utilize the psychoacoustic model (PAM) that has been successfully used to achieve the compression of the audio signals with a minimum loss of quality [15, 16]. A PAM-based frequency-dependent weighting for speech signal was previously tested in [17], but it was to achieve a more efficient network structure with less trainable parameters. Thus, the direct use of PAM to train the DNN-based compression model is still a potential direction for enhancing perceptual quality. The loss function in this paper is designed to maximize the mask-to-noise ratio (MNR) using the global masking threshold (GMT) computed by PAM. The loss function is further weighted by the perceptual entropy (PE) [15, 18] to drive the DNN to focus more on the perceptually important frequency bands. The rest of this paper is organized as follows. Section 2 briefly explains the DNN baseline model and incorporating loss functions for speech coding. Section 3 presents the proposed PAM-based loss function in detail. Experimental results are presented in Section 4. Section 5 draws conclusions.

2. Baseline DNN Model and Loss Functions for Speech Coding

This section describes the baseline DNN model for testing the perceptual loss functions and fundamental loss terms required for speech coding on the baseline DNN model.

2.1. Baseline DNN Model

We construct a unified end-to-end speech compression model based on a CNN autoencoder. A block diagram of the constructed model is shown in Fig. 1. The model consists of stacks of four bottleneck residual blocks [19], and it has 1.5 million trainable network parameters in total. The input and out vectors of the encoder and decoder subnetworks consist of T(=512) time-domain samples: $\{\mathbf{x}_l, \mathbf{y}_l\} \in \mathbb{R}^T$ where l is the frame index. We use a window function of length T with half sines of 32-sample lengths at both ends of overlapping regions. The same window is used to synthesize the output through the overlap-add procedure.

The operational process of the model in Fig. 1 similarly follows that in [12], where network parameters, quantization levels, and entropy for the bitrate control are jointly trained. The encoder extracts $T/2$ feature values from the input vector, i.e., $\mathbf{z}_l \leftarrow \mathcal{F}(\mathbf{x}_l)$ where $\mathbf{z}_l \in \mathbb{R}^{T/2}$. Each residual learning block contains four bottleneck structures with the dilation of 1, 2, 4, 8. A 2:1 downsampling is achieved by striding during the 1D convolution. The feature vector $\mathbf{z}_l$ is then quantized via the

http://dx.doi.org/10.21437/Interspeech.2021-2151

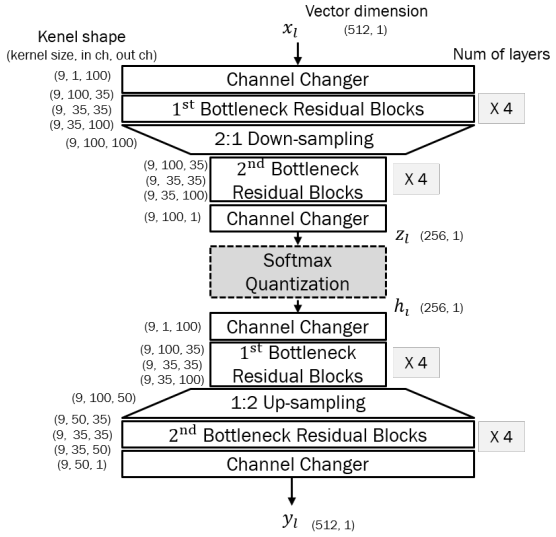


Figure 1: *Schematics of the baseline speech coder.*

softmax quantization process. The decoding process is a reverse of the encoding process, where a sub-pixel convolutional layer [20] is adopted for upsampling. The decoder recovers the original signal from the quantized feature vector: $\mathbf{h}_l \in \mathbb{R}^{T/2}$, as $\mathbf{y}_l \leftarrow \mathcal{G}(\mathbf{h}_l)$.

2.2. Basic Loss Terms and Trainable Parameters

Since the end-to-end model in Fig. 1 works in the time domain, the network can be straightforwardly trained using the MSE between the input and the recovered samples, as given by $\mathcal{L}_{mse} = \frac{1}{L}\sum_{l=1}^{L} \|\mathbf{y}_l - \mathbf{x}_l\|_2^2$ where L is the number of frames in a training minibatch and $\|\cdot\|_2$ denotes l_2 norm. Considering the overlap-and-add process between the frames, $\mathbf{y}_l$ and $\mathbf{x}_l$ were reconstructed via the exact time-domain signal reconstruction (eTDR) scheme [21].

Through quantization, the real-valued elements of the feature vector are converted to the nearest one among K discrete bins: $\mathbf{b} = [\,b_1, ..., b_K\,]$. The softmax quantization relaxes the quantization process using a soft assignment as $\hat{\mathbf{c}}_{l,i} = \mathrm{softmax}(\sigma \mathbf{d}_{l,i})$ where $\mathbf{d}_{l,i} = [\,|z_{l,i} - b_1|, ..., |z_{l,i} - b_K|\,]$. Then, the quantized code is acquired as $h_{l,i} = \hat{\mathbf{c}}_{l,i}^T \mathbf{b}, i = 1, ..., T/2$ and a hard assignment is obtained by increasing σ: $\mathbf{c}_{l,i} = \lim_{\sigma\to\infty} \hat{\mathbf{c}}_{l,i}$. As the standard softmax operator is differentiable, so is the quantization process. When we train the network, the bins in $\mathbf{b}$ are also updated to minimize the quantization error [11]. To prevent the network from generating unintended values, $\hat{\mathbf{c}}_{l,i}$ needs to be constrained to be close to a one-hot vector using a loss term [12], $\mathcal{L}_c = \frac{1}{L\cdot(T/2)} \sum_{l=1}^{L}\sum_{i=1}^{T/2} \left(\|\hat{\mathbf{c}}_{l,i}^{1/2}\|_1 - 1\right)$ where $\|\cdot\|_1$ denotes l_1 norm.

For the sake of rate-distortion optimization, a loss term for the entropy control is also added, which is defined as $\mathcal{L}_e = -\|\mathbf{p} \odot \log_2(\mathbf{p})\|_1$ where $\odot$ denotes an element-wise product. $\mathbf{p}$ is the probability distribution over the quantized symbol $\mathbf{c}$, which can be estimated by averaging the soft assignments generated by the encoder over the training frames or minibatch [11, 12], as $\mathbf{p} \approx \frac{1}{L\cdot(T/2)} \sum_{l=1}^{L}\sum_{i=1}^{T/2} \hat{\mathbf{c}}_{l,i}$. The average bitrate can be estimated as $\frac{f_s}{T-32} \times \frac{T}{2} \times \mathcal{L}_e$ *bps* where f_s is the sampling frequency.

2.3. Perceptually-Motivated Loss Functions

In [12], to enhance the perceptual quality, basic loss functions were augmented using the difference between the log Mel spectra of the input and output samples. The log Mel spectra were measured using four filterbanks with different resolutions to allow coarse and fine differentiation, as $\mathcal{L}_p = \frac{1}{4}\sum_{i=1}^{4} \|\mathcal{M}_y^i - \mathcal{M}_x^i\|_2$ where $\mathcal{M}_y^i$ and $\mathcal{M}_x^i$ denote the input and output log Mel spectrum vectors, respectively. In [8], a perceptual metric for speech quality evaluation, referred to as PMSQE, was proposed for speech enhancement. The PMSQE was designed using two disturbance terms, symmetrical and asymmetrical, inspired by the PESQ algorithm [22].

3. Proposed Loss Function

3.1. Psychoacoustic Model

There have been many studies to design loss functions to improve the perceptual quality of the speech signal [6, 7, 8, 12]. Psychoacoustic models (PAMs) are widely used in audio coding to exploit the simultaneous masking effect. PAM calculates GMT using the input power spectrum by accumulating all masking levels along with the absolute hearing threshold. There are many versions of PAM, from simple to complex. Although the accuracy of the PAM strongly affects the performance of the audio coder, a simple version still suffices since this study aims to verify that the PAM-based loss function has the potential to enhance the perceptual quality of the DNN-based speech coder. Therefore, in this paper, we use a simple version, PAM-1 [15], and the use of a more accurate PAM may result in better performance. In the PAM-1 implementation, a Sound Pressure Level (SPL) normalization was performed for the training data. The GMT was obtained by combining individual masking curves of tonal and noise maskers, plus the absolute hearing threshold. Detailed implementation of PAM-1 can be found in [15].

3.2. Loss Function for Noise Masking

This section presents a loss function based on the PAM that can make the DNN model focus more on perceptual meaningful features. Performing the short-time Fourier transforms (STFTs) on the input and output frames of the autoencoder, $\mathbf{x}_l$ and $\mathbf{y}_l$, we have the frequency coefficients, $\{\mathbf{X}, \mathbf{Y}\} \in \mathbb{C}^F$. In the following, we omit the frame index l for simplicity.

Using $\mathbf{X}$ and $\mathbf{Y}$, the input and output power spectra are calculated as $\mathbf{P}_x = [\ |X_0|^2, ..., |X_{F-1}|^2\ ]^T$ and $\mathbf{P}_y = [\,|Y_0|^2, ..., |Y_{F-1}|^2\ ]^T$, respectively. Then, the power spectra are vector transformed into the Mel frequency scale of different resolutions using transformation matrices $\mathbf{H}_i$, as follows,

$$\mathbf{C}_x^i = \mathbf{H}_i \mathbf{P}_x, \ \ \mathbf{C}_y^i = \mathbf{H}_i \mathbf{P}_y, \ \ i = 1, ..., N, \qquad (1)$$

where $\{\mathbf{C}_x^i, \mathbf{C}_y^i\} \in \mathbb{R}^{B_i}$ are the input and output Mel power spectra obtained using B_i-band Mel filterbanks. The index i is used to differentiate the Mel scale of different resolutions. We introduce the multi-scale observations to the loss function as in [12], because incorporating the coarse and detailed spectral features often helps the DNN model learn the perceptual attributes for human hearing better.

The power of the noise incurred by the softmax quantization is obtained as $\mathbf{P}_n = |\mathbf{Y} - \mathbf{X}|^2$. In addition, by applying PAM-1 to the input power spectrum, we obtain the GMT in the

power domain, $\mathbf{T}$. Similar to Eq. (1), $\mathbf{P}_n$ and $\mathbf{T}$ are also transformed onto the Mel-frequency scale as

$$\mathbf{C}_n^i = \mathbf{H}_i \mathbf{P}_n, \;\; \mathbf{C}_t^i = \mathbf{H}_i \mathbf{T}, \;\; i = 1, ..., N. \tag{2}$$

Bit allocation algorithms in most high-quality audio coders try to maximize the mask-to-noise ratio (MNR), resulting in the quantization noise power being controlled down below the masking curve. Thus, we want to focus on the frequency bands where the quantization noise exceeds the masking level. To this end, we define a log noise-to-mask ratio (NMR) vector to be minimized, as

$$\mathbf{D}_{nm}^i = \max\left\{\mathcal{C}_n^i - \mathcal{C}_t^i, \;\; \mathbf{0}_i\right\}, \;\; i = 1, ..., N, \tag{3}$$

where $\mathcal{C}_n^i = 10\log_{10}\mathbf{C}_n^i$, $\mathcal{C}_t^i = 10\log_{10}\mathbf{C}_t^i$, and $\mathbf{0}_i$ denotes a $(B_i \times 1)$ zero vector. Using $\mathbf{D}_{nm}^i$, we finally obtain a loss function,

$$\mathcal{L}_{nm} = \frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{w}_i \odot \mathbf{D}_{nm}^i\right\|_1, \tag{4}$$

where $\{\mathbf{w}_i\}$ are frequency-dependent weighting vectors.

Elements of the weighting vector $\mathbf{w}_i$ in Eq. (4) are determined based on the perceptual entropy (PE) [18] computed in each Mel frequency band. The PE is calculated using PAM-1, at each frequency bin f, as [15]

$$E_f = \log_2\left(2\frac{|\Re(X_f)|}{\sqrt{6T_f}} + 1\right) + \log_2\left(2\frac{|\Im(X_f)|}{\sqrt{6T_f}} + 1\right), \tag{5}$$

where $\Re\{\cdot\}$ and $\Im\{\cdot\}$ are the operators to take the real and imaginary parts, respectively. Again, after transforming the PE vector $\mathbf{E} = [E_0, ..., E_{F-1}]^T$ into the Mel-frequency scale as $\hat{\mathbf{E}}_i = \mathbf{H}_i\mathbf{E}$, the weight vector is determined by $\hat{\mathbf{E}}_i$ normalized to its maximum element:

$$\mathbf{w}_i = \left(\frac{\hat{\mathbf{E}}_i}{\|\hat{\mathbf{E}}_i\|_\infty}\right)^\gamma, \tag{6}$$

where $\|\cdot\|_\infty$ denotes the maximum norm and $\gamma \geq 0$ is a constant for controlling the relative ratio to the maximum PE. The significance of the loss function in Eq. (4) is that, by incorporating the PE weighting into the loss function, we can drive the DNN model to focus on the noise masking in perceptually more important frequency bands.

Finally, the total loss function for training the DNN model is obtained:

$$\mathcal{O} = \alpha_1\mathcal{L}_{mse} + \alpha_2\mathcal{L}_c + \alpha_3\mathcal{L}_e + \alpha_4\mathcal{L}_{nm}, \tag{7}$$

where $\alpha_i, i = 1, 2, 3, 4$ are the combination weights.

4. Experiments

Datasets were constructed using utterances from the TIMIT corpus [23] of the 16kHz sampling rate. In total, $3,000$ utterances were selected for training. Additional 200 utterances and another 500 utterances were chosen to create the validation and test sets, respectively. All sources were normalized by min-max normalizatoin. 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 [24]. We initialized the combination weights by considering the dynamic range of each loss term as $\alpha_1 = 60$, $\alpha_2 = 10$, and $\alpha_4 = 0.003$. α_3 is initiated as 0.5 and can be adjusted by 0.025 if the estimated bitrate is not in the target bitrate region.

Table 1: *Objective evaluation results for each coder.*

Coder	Bitrate	PESQ	SNR	APS	IPS	OPS
AMR-WB	12.65	3.771	11.31	69.84	82.79	89.45
LMS	13.32	3.919	15.09	63.32	83.54	89.99
PMSQE	13.34	4.049	14.53	60.59	77.62	84.61
Prop	12.81	3.878	14.47	65.30	82.68	89.11
AMR-WB	19.85	3.970	12.43	82.53	90.91	93.33
LMS	20.43	4.139	16.61	76.78	91.35	96.40
PMSQE	20.36	4.217	16.35	75.98	89.84	95.04
Prop	20.16	4.175	16.68	81.14	92.49	96.38
AMR-WB	23.85	3.999	12.69	84.97	94.65	97.16
LMS	24.23	4.219	17.44	81.22	94.25	97.93
PMSQE	24.44	4.264	16.67	79.49	93.44	95.73
Prop	24.42	4.228	17.13	83.32	94.78	97.91

4.1. Hyperparameter Optimization

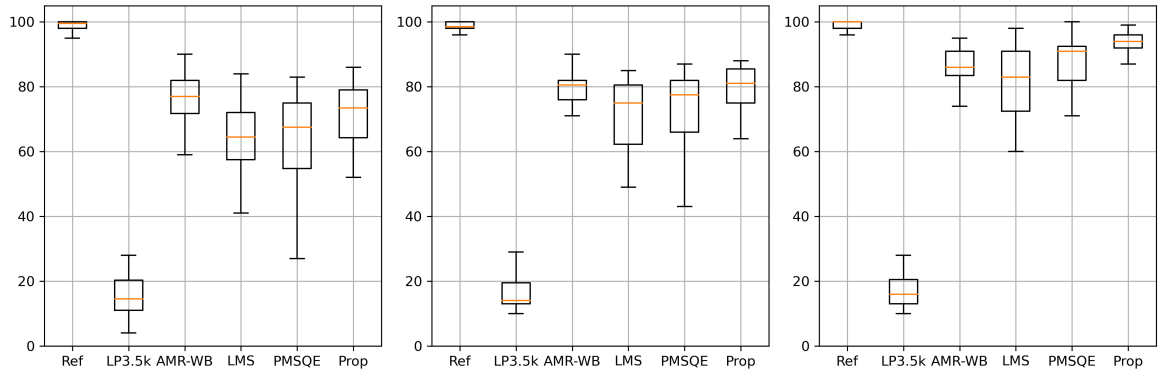
We determine the number of frequency scales (N) through experiment. The best PESQ result (4.175) during the validation at 20kbps was obtained when the three Mel-frequency scales of 16, 32 and 64 bands were integrated. Thus, in the following tests, three Mel filterbanks ($N = 3$) of 16, 32 and 64 bands were jointly used. We also set the parameter γ in Eq. (6) as 0.8 through experiments, which produced the highest PESQ during the validation. When no PE-based weighting was applied, i.e., $\gamma = 0$, we obtained a PESQ of 4.027. Through hyperparameter optimization, we could confirm a clear correlation between PE and PESQ. We also could observe that the perceptual attributes of the source signal need to be considered when we train the DNN.

4.2. Objective Performance Evaluation

The performance of the proposed loss function was assessed using the PESQ and SNR metrics. In addition, we measured the interference-related perceptual score (IPS), artifact-related perceptual score (APS), and overall perceptual score (OPS) using the Perceptual Evaluation methods for Audio Source Separation (PEASS) toolkit [25] to complement the PESQ score.

For comparison, we trained the same DNN model using the conventional perceptual loss terms: log Mel spectra (LMS) in [12] and PMSQE in [8]. After training the baseline DNN model at 13kbps, 20kbps and 24kbps, we computed PESQs and SNRs using the test dataset, which was also processed by the AMR-WB standard coder. The results are compared in Table 1. LMS, PMSQE, and Prop (proposed) in the table indicate training the baseline coder using corresponding loss functions. In terms of PESQ and SNR, DNN-based coders outperformed AMR-WB by a significant margin at all bitrates. However, in terms of artifact and interference-related scores (APS, IPS), AMR-WB achieved higher or comparable scores to DNN-based coders. In terms of OPS, none achieved consistent superiority over the others. Among the DNN models, PMSQE achieved slightly higher PESQs than the other two. However, even for high PESQs, the output speech obtained using PMSQE and LMS sometimes contained disturbing artifacts, which implies the fact that high PESQs do not always guarantee the high perceptual quality, as shown in the previous cases [3, 26]. To observe more on this, we compared the spectral distribution of the quantization noise with the GMT, in Fig. 3, for a frame encoded at 20kbps.

The spectral plots in Fig. 3 show that both PMSQE and LMS produce significant quantization noises, especially in the high-frequency region. But, the proposed loss function is relatively successful for modulating the quantization noise down



(a) *13kbps* (b) *20kbps* (c) *24kbps*

Figure 2: *MUSHRA test results*

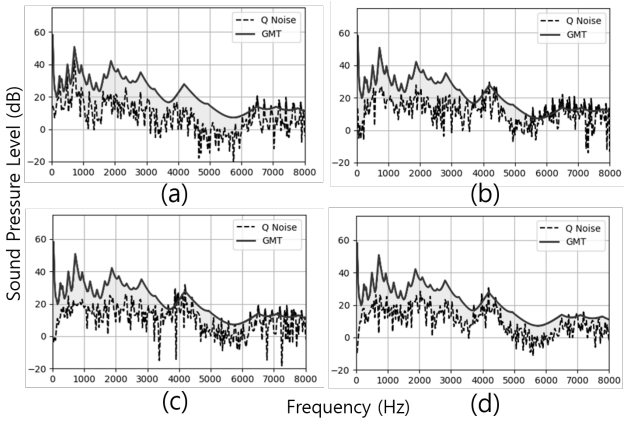


Figure 3: *Quantization noise (dashed) versus GMT (solid): (a) AMR-WB, (b) LMS, (c) PMSQE, and (d) Prop.*

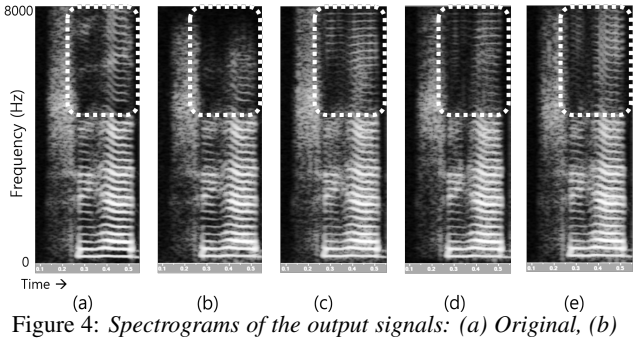


Figure 4: *Spectrograms of the output signals: (a) Original, (b) AMR-WB, (c) LMS, (d) PMSQE, and (e) Prop.*

below GMT. Spectrograms of the outputs are also shown in Fig. 4, where we can make similar observations. Spectral artifacts of LMS and PMSQE are visible in the high-frequency region, indicated by squares, while Prop faithfully reconstructs the spectral components. AMR-WB misses high-frequency signal components, which resulted in high quantization noise in the high-frequency region.

4.3. Subjective Quality Measure

To assess the perceptual quality of each loss function, we conducted a MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA) [27] test with 7 experienced listeners. 10 decoded files were randomly selected from the test set and processed using the DNN-based coders and AMR-WB, respectively, at 13kbps, 20kbps, and 24kbps. For each trial, the anchor was a low-pass filtered signal with a cutoff frequency at 3.5kHz.

Test results are shown in Fig. 2. We can see from the plots that the proposed loss function (Prop) always obtained higher scores at all bitrates over the other loss functions, LMS and PMSQE, which is resulted from that fact that the proposed loss function helps the baseline coder suppress the quantization noise below the GMT. However, overall MUSHRA scores of the baseline coder were lower than the AMR-WB standard coder. The main reason was spectral artifacts in the high-frequency region shown in Figs. 2 and 3 were sometimes audible. However, those artifacts were significantly reduced by using the proposed loss function. As a result, at 24kbps, it shows higher scores with smaller variance than the AMR-WB code, which is mainly because the proposed loss function enables us to train the baseline coder to suppress the quantization noise below the GMT.

5. Conclusions

This paper incorporated a PE-based weighting scheme into the MNR loss to achieve a perceptually more relevant loss function for the DNN-based speech coder. The proposed loss function was tested on a baseline DNN speech coder where the softmax quantization and entropy-based bitrate control are implemented. For comparison purpose, other popularly employed perceptual functions in [6, 7, 8, 12] were also tested on the same baseline speech coder. Objective test results showed that the baseline coder produced higher scores than AMR-WB in PESQ and SNR. MUSHRA test results showed that the proposed loss function (Prop) always obtained higher scores at all bitrates over the other loss functions. However, the perceptual quality of the baseline coder was lower than the AMR-WB standard coder. The proposed loss function could show comparable performance to AMR-WB at 20kbps and obtain higher quality at 24kbps. The proposed loss function can be easily amendable to DNN-based general audio coders, where the input and output audio frames are synchronized, and it will be our future work.

6. Acknowledgements

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

7. References

[1] Y. Xu, J. Du, L. Dai, and C. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[2] A. van den Oord, S. Dieleman, K. Simonyan H. Zen, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[3] W. Kleijn, F. Lim, A. Luebs, J. Skoglund, F. Stimberg, Q. Wang, and T. Walters, "Wavenet based low rate speech coding," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 676–680.

[4] J. Klejsa, P. Hedelin, C. Zhou, R. Fejgin, and L. Villemoes, "High-quality speech coding with sample RNN," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7155–7159.

[5] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, 2019.

[6] S. Fu, T. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1570–1584, 2018.

[7] S. Fu, C. Liao, and Y. Tsao, "Learning with learned loss function: Speech enhancement with quality-net to improve perceptual evaluation of speech quality," *IEEE Signal Processing Letters*, vol. 27, pp. 26–30, 2020.

[8] J. M. Martin-Doñas, A. M. Gomez, J. A. Gonzalez, and A. M. Peinado, "A deep learning loss function based on the perceptual evaluation of the speech quality," *IEEE Signal Processing Letters*, vol. 25, no. 11, pp. 1680–1684, 2018.

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

[10] Y. Choi, M. El-Khamy, and J. Lee, "Towards the limit of network quantization," in *5th International Conference on Learning Representations (ICLR)*, 2017.

[11] 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 *Neural Information Processing Systems (NIPS)*, Dec. 2017, pp. 1141–1151.

[12] S. 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.

[13] K. Zhen, M. S. Lee, J. Sung, S. Beack, and M. Kim, "Psychoacoustic calibration of loss functions for efficient end-to-end neural audio coding," *IEEE Signal Processing Letters*, vol. 20, pp. 2159–2163, 2020.

[14] Bruno Bessette, Redwan Salami, Roch Lefebvre, Milan Jelinek, Jani Rotola-Pukkila, Janne Vainio, Hannu Mikkola, and Kari Jarvinen, "The adaptive multirate wideband speech codec (AMR-WB)," *IEEE transactions on speech and audio processing*, vol. 10, no. 8, pp. 620–636, 2002.

[15] A. Spanias, T. Painter, and V. Atti, *Audio Signal Processing and Coding*, John Wiley Sons, Inc., 2007.

[16] J. Herre and S. Dick, "Psychoacoustic models for perceptual audio coding—a tutorial review," *Applied Sciences*, vol. 9, no. 14, pp. 1–22, 2019.

[17] Kai Zhen, Aswin Sivaraman, Jongmo Sung, and Minje Kim, "On psychoacoustically weighted cost functions towards resource-efficient deep neural networks for speech denoising," in *Proc. the Seventh Annual Midwestern Cognitive Science Conference*, May 2018.

[18] James D. Johnston, "Estimation of perceptual entropy using noise masking criteria," in *1988 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1988, pp. 2524–2527.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[20] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1874–1883.

[21] J. Lee and H.-G. Kang, "A joint learning algorithm for complex-valued tf masks in deep learning-based single-channel speech enhancement systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 1098–1108, 2019.

[22] ITUT Rec, "P. 862.2: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs," *International Telecommunication Union, CH–Geneva*, 2005.

[23] J. W. Lyons, *DARPA TIMIT acoustic-phonetic continuous speech corpus*, Nat. Inst. Stand. Technol., 1993.

[24] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, vol. 19, no. 7, pp. 2046–2057, 2014.

[25] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[26] J. Valin and J. Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5891–5895, 2019.

[27] B Series, "Method for the subjective assessment of intermediate quality level of audio systems," *International Telecommunication Union Radiocommunication Assembly*, 2014.