Using Perceptual Quality Features in the Design of the Loss Function for Speech Enhancement

Nicholas Eng*, Yusuke Hioka*, Catherine I Watson†
* Acoustics Research Centre, Department of Mechanical and Mechatronics Engineering, University of Auckland, Auckland, 1010 New Zealand
† Department of Electrical, Computer and Software Engineering, University of Auckland, Auckland, 1010 New Zealand
E-mail: neng668@aucklanduni.ac.nz, yusuke.hioka@ieee.org, c.watson@auckland.ac.nz

Abstract—**Although deep learning has shown success in the domain of speech enhancement, there is still much research that can be undertaken in finding alternative loss functions to train the speech enhancement models. An approach is to utilise features in objective quality metrics, which aim to judge the quality of enhanced or transmitted speech, as part of the loss function. One such objective quality metric utilises perceptual quality dimensions, which identify perceptual quality features that can be derived from the signal and have shown to contribute to perceptual speech quality. In this study, we investigate two such features, cepstrum statistics and MFCC statistics, to be used alongside a baseline loss function for a DNN speech enhancement method. The results from experimentation show that addition of cepstrum statistics to the loss function is detrimental to the scores of objective quality metrics compared to a baseline loss function, however using MFCC statistics in the loss function improves speech quality scores.**

I. Introduction

With the advent of machine learning, there have been many interesting and effective methods of speech enhancement which have shown to outperform traditional speech enhancement methods, especially under more adverse conditions such as under non-stationary noise or reverberation. However, there is still much that could be done to further improve the quality of these speech enhancement methods to increase speech quality.

One such improvement is to develop better objective functions that are utilised to train the speech enhancement models. Many recent speech enhancement methods aim to reduce the L1-loss or Mean-Squared Error (MSE) between the estimate and true speech waveforms or spectrograms. However, previous studies have shown that these typical objective functions may not accurately reflect human perception of speech quality [1], [2], [3], [4]. Several studies have investigated the use of alternative loss functions, for example, using perceptual correlates such as bark spectral distortion [2] or weighting filters based on code-excited linear prediction [5]. Additionally, other studies have investigated the use of objective signal evaluation metrics in the loss function, such as the perceptual evaluation of speech quality (PESQ) [6], [7], [8], [9], short-time objective intelligibility (STOI) [1], [7], [10], [11], [12], or scale-invariant signal-to-distortion ratio (SI-SDR) [4]. These studies show the potential in using perceptually motivated loss functions to improve the quality of enhanced speech. Whilst perceptually motivated speech enhancement techniques have been proposed previously, especially for traditional speech enhancement methods, using objective quality metrics in the loss function have recently shown to be effective. However, such as in the case of PESQ, the use of objective quality metrics for a loss function has its challenges, such as the algorithms being complex or non-differentiable, the latter of which making it impossible for back-propagation to occur [6].

As such, there is a motive to investigate simpler and more transparent designs for a perceptually motivated loss function in speech enhancement. In order to do so, it is essential to know which perceptual ”features” can reflect speech quality. Previous studies have aimed to find important perceptual features that can be extracted from the clean and degraded/enhanced waveforms that in turn can be used to objectively evaluate the quality of speech [13], [14], [15], [16]. These features were then used to create a speech quality objective metric that aims to provide diagnostic information on a specific speech quality score. Through a series of perception tests, these studies have identified perceptual quality features that make a significant contribution in both the judging of speech quality and are related to how a listener can perceive quality using multidimensional analyses. Resulting from this research, four perceptual quality dimensions have been identified - noisiness, discontinuity, colouration and loudness. Each of these quality dimensions is comprised of further sub-dimensions, which in turn are functions of several perceptual features that can be extracted from the speech. Although this research was intended for speech degraded by transmission over communication channels, many objective quality metrics have shown to correlate well to speech quality in both the transmitted speech and speech enhancement domains [17], [18].

Although these perceptual features were originally combined with one another to judge speech quality, it is yet to be discovered if they can work alongside typical loss functions used for speech enhancement to further improve the quality of enhanced speech. In this study, we investigate the use of quality features that were identified to be used as part of the perceptual quality dimensions, as part of the training target to train a speech enhancement model. In particular, we investigate statistical features that were identified to be related to the colouration of speech (defined in [14]), namely the cepstral

978-616-590-477-3 ©2022 APSIPA

statistics and the standard deviation of the Mel-Frequency Cepstral Coefficients (MFCCs). We use DEMUCS [19] as the test case of speech enhancement algorithm, and add the features to the loss function to determine if the quality of enhanced speech can be further improved.

II. Perceptual Quality Features

This section describes the perceptual quality features utilised in this study and the loss functions derived from these features.

A. Cepstrum Statistics

Traditionally, the complex cepstrum has been used in many speech applications such as echo detection [20], speech synthesis [21], and glottal source estimation [22]. However, statistics related to the complex cepstrum have also been used in speech quality estimation [13], [23]. Following [24], the complex cepstrum C of an arbitrary signal $x(t)$ can be calculated by the inverse Fourier transform of the complex logarithm of the Fourier transform of $x(t)$, i.e,

$$C(x(t)) = \frac{1}{2\pi}\int_{\pi}^{-\pi} \log[X(e^{jw})]e^{j\omega t}\, d\omega, \tag{1}$$

where $X(e^{j\omega})$ denotes the Fourier transform of $x(t)$ at time t.

To calculate the cepstrum statistics from the cepstrum, firstly the cepstrum for each 10 ms frame of the speech signal is obtained, then the standard deviation (σ) and kurtosis (α^4) are calculated from the cepstrum of each frame. However, in creating the loss function, the aim is to minimise the statistical difference between the cepstrums of the source and enhanced signals. As such, the cepstrum standard deviation loss (CEP-STD) and cepstrum kurtosis loss (CEP-KRT) are defined as the standard deviation and kurtosis of the difference of the cepstrums, as follows,

$$L_{\text{CEP}-\text{STD}}(y,\hat{y}) = \sigma(C(y) - C(\hat{y})) \tag{2}$$

$$L_{\text{CEP}-\text{KRT}}(y,\hat{y}) = \alpha^4(C(y) - C(\hat{y})), \tag{3}$$

where y denotes the clean signal in the time domain, $\hat{y}$ denotes the enhanced signal in the time domain and C represents the complex cepstrum described by (1). For brevity, y and $\hat{y}$ are assumed to be equivalent to $y(t)$ and $\hat{y}(t)$, respectively.

B. MFCC Standard Deviation

Mel Frequency Cepstral Coefficients (MFCCs) have also been used in speech processing applications, such as speaker recognition [25], emotion recognition [26], and speech synthesis [27]. MFCCs are extracted from a speech signal $x(t)$ by first passing Fourier transformed speech frames through a mel-filter bank containing M triangular mel weighting filters to obtain the mel spectrum $s_x(m)$, for each mel weighting filter m. The MFCCs are then calculated by applying the discrete cosine transform (DCT) to the mel spectrum as follows,

$$F_x(n) = \sum_{m=0}^{M-1} \log(s_x(m)) \cos\left(\frac{\pi n(m-0.5)}{M}\right) \tag{4}$$
$$n = 0,1,2...P-1,$$

where $F_x(n)$ is the MFCCs of order n, and P is the number of MFCCs. MFCCs are calculated for each 30 ms frame with a 50% overlap. The standard deviations of each coefficient are calculated for the entire utterance, then the mean of all the standard deviations are calculated. In the case of the loss function, as both the source and enhanced signals need to be considered, the MFCC standard deviation loss (MFCC-STD) is defined as follows,

$$L_{\text{MFCC}-\text{STD}}(y,\hat{y}) = \frac{\sum_{n=1}^{P}\sigma(F_y(n) - F_{\hat{y}}(n))}{P}, \tag{5}$$

where y denotes the clean signal in the time domain, $\hat{y}$ denotes the enhanced signal in the time domain and F represents the MFCCs of the n-th coefficient described by (4).

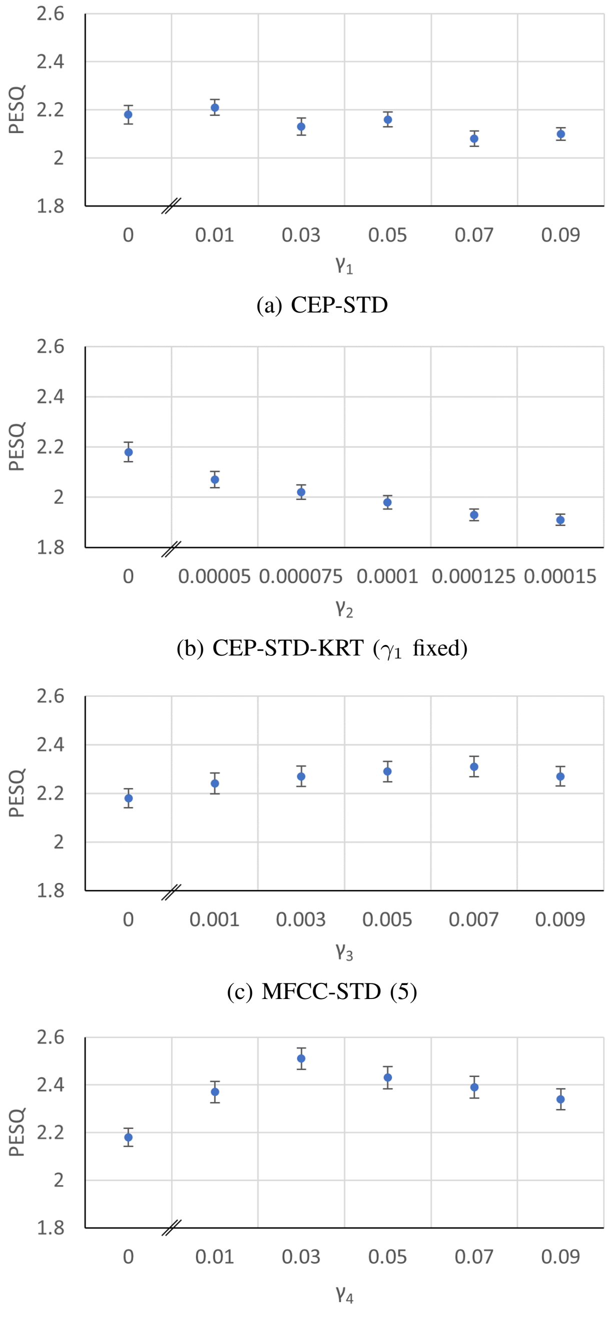
We investigate both 5th order and 20th order MFCCs, hence two loss functions are investigated, MFCC-STD (5) and MFCC-STD (20). Additionally, we also investigate calculating the standard deviations of the MFCCs acquired only from active frames (MFCC-STDa) against just taking the standard deviations of the MFCCs acquired from the entire utterance.

III. Experiments

A. Experimental design

The proposed loss functions are not tied to any specific speech enhancement algorithm, but as a test case, we use DEMUCS [19], a music source separation architecture that was adopted for speech enhancement that can work in real-time. It consists of a multi-layered convolutional encoder network followed by a multi-layered convolutional decoder network with U-Net [28] skip connections. The encoder network operates on the raw waveform as an input, and each i-th layer in the encoder network outputs a latent representation to its corresponding i-th layer in the decoder network. In the case of the causal model, a uni-directional long short-term memory (LSTM) is used for sequence modelling and connects the encoder and decoder network. Finally the decoder network decodes the latent representation to the estimated waveform.

All models were trained on the Valentini dataset [29] consisting of 28 speakers with 11,572 utterances, with a test set consisting of two speakers with 814 utterances. Speakers 286 and 287 from the training set were used to form the validation set, and thus were excluded from the training set. All utterances were downsampled to 16 kHz from 48 kHz. The hyperparameters were set as, H (number of hidden channels) = 48, K (layer kernel size) = 8, S (stride) = 4, and U (resampling factor) = 4. As in the original study [19], the Adam optimiser [30] was used with a step size of 3e-4, β_1 = 0.9 and β_2 = 0.999.



(a) CEP-STD

(b) CEP-STD-KRT (γ_1 fixed)

(c) MFCC-STD (5)

(d) MFCC-STD (20)

Fig. 1: Optimisation of weights $\gamma_1 - \gamma_4$ on the validation set for the models (a) CEP-STD, (b) CEP-KRT (with CEP-STD fixed a 0.01), (c) MFCC-STD, 5 coefficients, (d) MFCC-STD, 20 coefficients. The scores at a weight of 0 represent the baseline (L1 Waveform) model.

A baseline model was created using the L1 loss of the waveform, defined in [19] as

$$L_{\text{waveform}}(y, \hat{y}) = \frac{1}{T}||y - \hat{y}||_1, \tag{6}$$

where y and $\hat{y}$ represent the reference and enhanced speech waveforms, respectively, T represents the duration of the speech waveforms in samples, and $||\cdot||_1$ represents the L_1 norm. The test models used the losses described in Section 2 along with the L1 loss of the waveform at added specific weights γ. As a result, four models were created using the following loss functions:

$$\text{CEP-STD: } L_{\text{waveform}} + \gamma_1 \cdot L_{\text{CEP-STD}} \tag{7}$$

$$\text{CEP-STD-KRT: } L_{\text{waveform}} + \gamma_1 \cdot L_{\text{CEP-STD}} + \gamma_2 \cdot L_{\text{CEP-KRT}} \tag{8}$$

$$\text{MFCC-STD (5): } L_{\text{waveform}} + \gamma_3 \cdot L_{\text{MFCC-STD}};\ P = 5 \tag{9}$$

$$\text{MFCC-STD (20): } L_{\text{waveform}} + \gamma_4 \cdot L_{\text{MFCC-STD}};\ P = 20 \tag{10}$$

where P is the MFCC order. The models were trained to 400 epochs, with validation testing occurring every 10 epochs, in which PESQ scores were also obtained on the validation test. The final model after the 400 epochs is the model which obtained the lowest validation loss.

B. Active frame detection

For the CEP-STD and CEP-STD-KRT models, the cepstrum standard deviation and cepstrum kurtosis were calculated from the frames where speech was deemed active using a voice activity detector (VAD). To distinguish active and inactive frames, the power spectral density of the clean signal was calculated for each frame, and frames above a threshold of 0.0002 were classified as active frames. The MFCC-STD (5) and MFCC-STD (20) models used MFCCs calculated from all frames, however we also investigated the effect of creating the MFCC-STD (5) and MFCC-STD (20) models using MFCCs calculated from only the active frames. These models are referred to as MFCC-STDa (5) and MFCC-STDa (20).

C. Loss-function weight optimisation

For loss functions consisting of multiple components in the loss, it is important to assign a weighting factor to each component of the loss function to optimise results. Therefore a grid-search was used to obtain optimised weights for the models. For the grid search, models were trained on a smaller database containing the first eight speakers of the Valentini database. The models were trained to 50 epochs, with validation testing every 10 epochs. The weight for the L1 Waveform loss was fixed to one, then the additional loss component was added to the L1 Waveform loss at a differing weight γ. The weight which obtained the best PESQ result after the grid search was used to train the network on the full dataset. In the case of the CEP-STD + CEP-KRT, the best performing weight was chosen for the CEP-STD component first, and subsequently the CEP-KRT weight was iterated with the grid search. For each model, two grid-searches were run, the first search providing a rough estimate of the weights, and the second for a more precise estimate of the weights.

The average PESQ scores after the second (precise) grid search is shown in Fig. 1, alongside their 95% confidence

TABLE I: Mean objective results on the test set for each model. Bold signifies a significant difference compared to the baseline model (L1 Waveform). For the MFCC-STD (20) and MFCC-STDa (20) models, underline represents a significant difference against their respective MFCC-STD (5) models. A higher score represents higher objective quality.

	WB-PESQ	ViSQOL	NISQA	CSIG	CBAK	COVL
L1 Waveform (Baseline)	2.47	3.19	3.66	3.77	3.18	3.12
CEP-STD	2.44	3.15	**3.59**	**3.62**	3.16	**3.01**
CEP-STD-KRT	**2.27**	3.16	**3.44**	**3.35**	**3.10**	**2.80**
MFCC-STD (5)	2.52	3.19	**3.52**	3.76	3.15	3.13
MFCC-STD (20)	**2.63**	3.20	3.71	**3.89**	3.21	**3.25**
MFCC-STDa (5)	2.51	3.20	**3.56**	3.74	3.18	3.12
MFCC-STDa (20)	**2.57**	3.20	3.74	3.84	3.20	3.19

intervals. Also included in each plot is the mean PESQ score for the model trained in the same manner with just the baseline L1 Waveform loss, along with its corresponding 95% confidence interval. It should be noted that unlike the result from both MFCC-STD (5) and MFCC-STD (20) searches, which shows a local maximum PESQ value in the search, the CEP-STD and CEP-STD + CEP-KRT searches resulted in a higher PESQ scores as the weights approach zero. The highest scoring weight for the CEP-STD model obtained a mean PESQ score of 2.21 with a 95% confidence interval of ±0.033, which is only marginally higher than the baseline PESQ score of 2.18 with a 95% confidence interval of ±0.032 and thus is not significantly different. Additionally, the highest scoring CEP-STD-KRT model obtained a mean PESQ score of 2.07 with a 95% confidence interval of ±0.032, which is significantly lower than the baseline PESQ score suggesting the best weights for the CEP-STD-KRT model is 0. However, we decided to use the best performing weight from the grid-search to investigate the effect of the CEP-KRT loss on the overall quality. From this search, weights γ_1 = 0.01, γ_2 = 0.00005, γ_3 = 0.007 and γ_4 = 0.03 were selected for the CEP-STD, CEP-KRT, MFCC-STD (5), and MFCC-STD (20) weights, respectively. For the MFCC-STDa models, the same weights that were used for the MFCC-STD models were used, i.e., γ_3 for the MFCC-STDa (5) and γ_4 for the MFCC-STDa (20) models.

D. Objective evaluation

The quality of the enhanced speech was evaluated using several objective metrics. We utilised:

(i) WB-PESQ: Wide-band Perceptual Evaluation of Speech Quality [31]

(ii) ViSQOL: Virtual Signal Quality Objective Listener [32]

(iii) NISQA: Non-Intrusive Speech Quality Assessment [33]

(iv) CSIG, CBAK, COVL: Predicted measures of the signal distortion of the speech signal, intrusiveness of the background noise and overall MOS prediction, respectively [17].

IV. Results

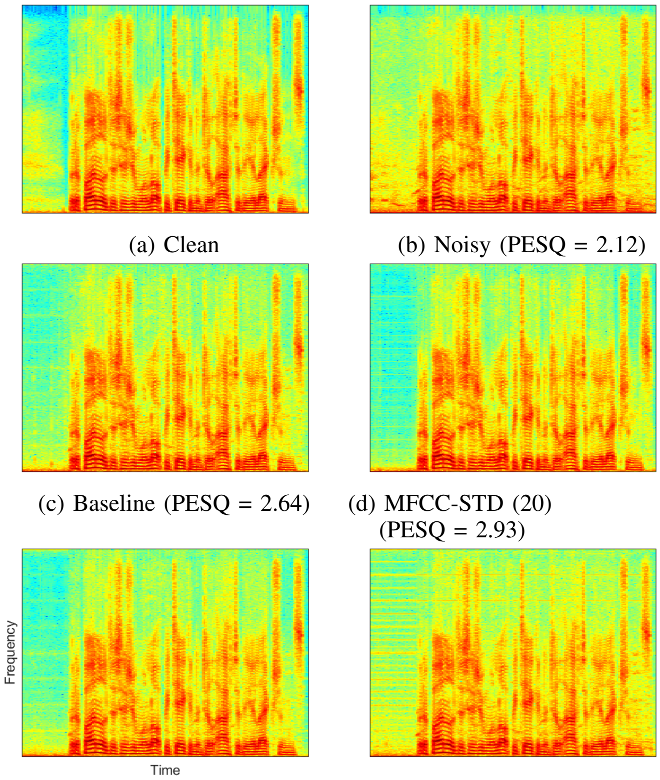
Table I shows the results of the objective metrics for the models tested, where bold numbers represent cases which are significantly different to the baseline (L1 Waveform), and underlined numbers for the MFCC-STD (20) and MFCC-STDa (20) models represent cases which are significantly different to their respective MFCC-STD (5) and MFCC-STDa (5) models. In both cases, significance was calculated through paired t-tests in R, with p-values < 0.05 considered significant, and with a Bonferroni correction due to the number of pairs being compared.

The model created with MFCC-STD (20) loss showed improvements over the baseline in all objective metrics, with significantly higher PESQ, CSIG and COVL scores. Additionally, the model created with MFCC-STD (5) loss showed no significant difference compared to the baseline apart from NISQA, in which it performed lower than the baseline. Comparing 5 and 20 coefficient models, we can observe that objective scores increase as the number of coefficients increase for all objective metrics and paired t-tests between these models show the differences are significant for all metrics except ViSQOL for the MFCC-STD models, and differences are significant for NISQA, CSIG and COVL for the MFCC-STDa models. Paired t-tests were also conducted to compare MFCC-STD models created with all frames to just active frames. However, there was no significant difference between the MFCC-STD and MFCC-STDa models in terms of all the objective metrics for both the 5 and 20 coefficient models.

For the cepstrum statistics losses, adding CEP-STD to the baseline loss resulted in little difference compared to the baseline in all objective metrics apart from NISQA, CSIG and COVL, in which it obtained lower average scores compared to the baseline, and the addition of both CEP-STD and CEP-KRT losses to the baseline loss performs significantly worse than the baseline in all metrics.

Following the grid search results shown in Fig. 1, which shows that PESQ scores increase as the CEP-STD and CEP-KRT weights decrease, these results confirm that the addition of CEP-KRT losses at any weight performs lower than the baseline is therefore detrimental to the speech quality, with quality decreasing with increasing weight.

Spectrograms of a sentence which was typical of the test set are shown in Fig. 2, along with their corresponding PESQ scores. We can observe that the MFCC-STD model is more effective in reducing noise than the baseline, and the CEP-STD model is similar in reducing noise than the baseline. However, artefacts appear shown by the horizontal lines in the spectrum in the case of CEP-STD, and this effect is worse for the CEP-STD-KRT model. This suggests that the added artefacts may



(a) Clean

(b) Noisy (PESQ = 2.12)

(c) Baseline (PESQ = 2.64)

(d) MFCC-STD (20) (PESQ = 2.93)

(e) CEP-STD (PESQ = 2.73)

(f) CEP-STD-KRT (PESQ = 2.20)

Fig. 2: Spectrograms of a utterance from the test set for: (a) clean speech, (b) noisy speech, (c) L1 Waveform baseline model, (d) MFCC-STD, 20 coefficients, (e) CEP-STD, (f) CEP-STD + CEP-KRT.

be a major cause for the lower objective scores observed in the CEP-STD and CEP-STD-KRT models.

V. Conclusion

This study investigated the use of two identified perceptual quality features, cepstrum statistics and MFCC standard deviation, as part of a loss function in a speech enhancement algorithm to improve the quality of resultant enhanced speech. The experimental results indicate that the addition of MFCC standard deviation in the loss function is able to significantly improve speech quality compared to just L1 waveform loss, scoring significantly higher than the baseline in three of the objective metrics and scoring comparably to the baseline in the other three objective metrics. However, the addition of cepstrum statistics were detrimental to speech quality, with the CEP-STD model scoring significantly worse in three of the objective metrics, and the CEP-STD-KRT model scoring significantly worse in five of the objective metrics. As there are many perceptual quality features that have been identified in the literature, future work will include running a subjective listening test, the testing of other perceptual quality features, as well as testing on other speech enhancement architectures.

References

[1] Yuma Koizumi, Kenta Niwa, Yusuke Hioka, Kazunori Kobayashi, and Yoichi Haneda, “DNN-based source enhancement to increase objective sound quality assessment score,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1780–1792, 2018.

[2] Xiaofeng Shu, Yi Zhou, Hongqing Liu, and Trieu-Kien Truong, “A human auditory perception loss function using modified bark spectral distortion for speech enhancement,” *Neural Processing Letters*, vol. 51, no. 3, pp. 2945–2957, 2020.

[3] Tsun-An Hsieh, Cheng Yu, Szu-Wei Fu, Xugang Lu, and Yu Tsao, “Improving perceptual quality by phone-fortified perceptual loss using wasserstein distance for speech enhancement,” *arXiv preprint arXiv:2010.15174*, 2020.

[4] Morten Kolbæk, Zheng-Hua Tan, Søren Holdt Jensen, and Jesper Jensen, “On loss functions for supervised monaural time-domain speech enhancement,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 825–838, 2020.

[5] Ziyue Zhao, Samy Elshamy, and Tim Fingscheidt, “A perceptual weighting filter loss for DNN training in speech enhancement,” in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 229–233.

[6] Szu-Wei Fu, Chien-Feng Liao, and Yu 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, 2019.

[7] Hui Zhang, Xueliang Zhang, and Guanglai Gao, “Training supervised speech separation system to improve STOI and PESQ directly,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5374–5378.

[8] Yuma Koizumi, Kenta Niwa, Yusuke Hioka, Kazunori Kobayashi, and Yoichi Haneda, “DNN-based source enhancement self-optimized by reinforcement learning using sound quality measurements,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 81–85.

[9] Juan Manuel Martin-Donas, Angel Manuel Gomez, Jose A Gonzalez, and Antonio 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.

[10] Szu-Wei Fu, Tao-Wei Wang, Yu Tsao, Xugang Lu, and Hisashi 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.

[11] Yan Zhao, Buye Xu, Ritwik Giri, and Tao Zhang, “Perceptually guided speech enhancement using deep neural networks,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5074–5078.

[12] Shrikant Venkataramani, Ryley Higa, and Paris Smaragdis, “Performance based cost functions for end-to-end speech separation,” in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 350–355.

[13] Friedemann Köster, Gabriel Mittag, Tim Polzehl, and Sebastian Möller, “Non-intrusive estimation of noisiness as a perceptual quality dimension of transmitted speech,” in *5th ISCA/DEGA Workshop on Perceptual Quality of Systems (PQS 2016)*, 2016, pp. 74–78.

[14] Gabriel Mittag, Friedemann Köster, and Sebastian Möller, “Non-intrusive estimation of the perceptual dimension coloration,” *Fortschritte der Akustik, DAGA*, 2016.

[15] Lu Huo, Marcel Waltermann, Ulrich Heute, and Sebastian Moller, “Estimation model for the speech-quality dimension” noisiness.”,” *Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 3864–3864, 2008.

[16] Lu Huo, Marcel Waeltermann, Ulrich Heute, and Sebastian Moeller, “Estimation of the speech quality dimension’discontinuity’,” in *ITG Conference on Voice Communication [8. ITG-Fachtagung]*. VDE, 2008, pp. 1–4.

[17] Yi Hu and Philipos C Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2007.

[18] Andrew Hines, Jan Skoglund, Anil Kokaram, and Naomi Harte, “Robustness of speech quality metrics to background noise and network degradations: Comparing ViSQOL, PESQ and POLQA,” in *2013 IEEE*

International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013, pp. 3697–3701.

[19] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi, "Real time speech enhancement in the waveform domain," *arXiv preprint arXiv:2006.12847*, 2020.

[20] Alan V Oppenheim and Ronald W Schafer, "From frequency to quefrency: A history of the cepstrum," *IEEE signal processing Magazine*, vol. 21, no. 5, pp. 95–106, 2004.

[21] Ranniery Maia, Masami Akamine, and Mark JF Gales, "Complex cepstrum for statistical parametric speech synthesis," *Speech Communication*, vol. 55, no. 5, pp. 606–618, 2013.

[22] Thomas Drugman, Baris Bozkurt, and Thierry Dutoit, "Complex cepstrum-based decomposition of speech for glottal source estimation," *arXiv preprint arXiv:1912.12602*, 2019.

[23] T Falk and W Chan, "Single ended method for objective speech quality assessment in narrowband telephony applications," *ITU-T*, p. 563, 2004.

[24] Alan V Oppenheim, John R Buck, and Ronald W Schafer, *Discrete-time signal processing. Vol. 2*, Upper Saddle River, NJ: Prentice Hall, 2001.

[25] Vibha Tiwari, "MFCC and its applications in speaker recognition," *International journal on emerging technologies*, vol. 1, no. 1, pp. 19–22, 2010.

[26] A Milton, S Sharmy Roy, and S Tamil Selvi, "SVM scheme for speech emotion recognition using MFCC feature," *International Journal of Computer Applications*, vol. 69, no. 9, 2013.

[27] Lauri Juvela, Bajibabu Bollepalli, Xin Wang, Hirokazu Kameoka, Manu Airaksinen, Junichi Yamagishi, and Paavo Alku, "Speech waveform synthesis from MFCC sequences with generative adversarial networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5679–5683.

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[29] Cassia Valentini-Botinhao et al., "Noisy speech database for training speech enhancement algorithms and TTS models," 2017.

[30] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[31] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*. IEEE, 2001, vol. 2, pp. 749–752.

[32] Andrew Hines, Jan Skoglund, Anil C Kokaram, and Naomi Harte, "ViSQOL: an objective speech quality model," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–18, 2015.

[33] Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller, "NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," *arXiv preprint arXiv:2104.09494*, 2021.