

SPEECH ENHANCEMENT WITH ZERO-SHOT MODEL SELECTION

Ryandhimas E. Zezario¹², Chiou-Shann Fuh¹, Hsin-Min Wang³, Yu Tsao²

¹Department of Computer Science and Information Engineering, National Taiwan University, Taiwan

²Research Center for Information Technology Innovation, Academia Sinica, Taiwan

³Institute of Information Science, Academia Sinica, Taiwan

ABSTRACT

Recent research on speech enhancement (SE) has seen the emergence of deep learning-based methods. It is still a challenging task to determine effective ways to increase the generalizability of SE under diverse test conditions. In this paper, we combine zero-shot learning and ensemble learning to propose a zero-shot model selection (ZMOS) approach to increase the generalization of SE performance. The proposed approach is realized in two phases, namely offline and online phases. The offline phase clusters the entire set of training data into multiple subsets, and trains a specialized SE model (termed component SE model) with each subset. The online phase selects the most suitable component SE model to carry out enhancement. Two selection strategies are developed: selection based on quality score (QS) and selection based on quality embedding (QE). Both QS and QE are obtained by a Quality-Net, a non-intrusive quality assessment network. In the offline phase, the QS or QE of a training utterance is used to group the training data into clusters. In the online phase, the QS or QE of the test utterance is used to identify the appropriate component SE model to perform enhancement on the test utterance. Experimental results have confirmed that the proposed ZMOS approach can achieve better performance in both seen and unseen noise types compared to the baseline systems, which indicates the effectiveness of the proposed approach to provide robust SE performance.

Index Terms— *speech enhancement, deep learning, zero-shot learning, model selection.*

1. INTRODUCTION

Speech enhancement (SE) is an important front-end module for various speech-related applications, such as automatic speech recognition (ASR) [1–3], assistive listening [4–8], speech coding [9–10], and speaker recognition [11–12] systems. The primary aim of SE is to retrieve clean speech signals from noisy ones. With the emergence of deep learning algorithms, notable improvements for SE have been made over the traditional SE methods. Well-known examples include the fully connected neural network [13–14], deep denoising auto-encoder (DDAE) [15–16], convolutional neural network (CNN) [17–19], and long-short-term memory (LSTM) [20–21]. Despite promising improvements have been made, increasing the generalizability of these deep learning-based SE methods to unseen environments remains a critical research topic.

Zero-shot learning is one of the machine learning algorithms and has been proven to be capable of improving generalizability to unseen environments. This learning criterion has been successfully implemented in the field of image processing to recognize unseen objects with satisfactory performance [22–26]. So far, zero-shot learning has not been fully explored in the field of speech processing [27–

28]. Particularly, no attempt has been made to apply zero-shot learning to the SE task.

In this paper, we propose a novel zero-shot model selection (ZMOS) approach for SE. The proposed approach combines zero-shot learning and ensemble learning to improve SE performance under any specific test conditions. The proposed approach is implemented in two phases, namely offline and online phases. In the offline phase, we prepare multiple specialized SE models (termed component SE models). Each component SE model is trained to match a specific noisy condition. In the online phase, we select the most suitable component SE model to enhance a test utterance. For the proposed approach, how to effectively cluster the training data to train the multiple component SE models in the offline phase and how to select the most suitable component SE model for a test utterance in the online phase are critical points. We propose to perform data clustering and model selection by using a pre-trained Quality-Net [30, 31]. Quality-Net is a deep learning-based non-intrusive quality assessment model. Given an utterance, Quality-Net will output a quality assessment score. In previous studies, it has been shown that Quality-Net can accurately predict the quality assessment score of an utterance.

Two types of data clustering and model selection strategies are developed: one is based on the quality score (QS), and the other is based on the quality embedding (QE); the corresponding approaches are termed ZMOS-QS and ZMOS-QE, respectively. Both QS and QE are estimated by Quality-Net. Given an utterance, QS is based on the output score of Quality-Net, and QE is based on the embedding vector of Quality-Net. In the offline phase, the QS or QE is used to group the training data into several clusters. Each cluster is used to train a corresponding specialized SE model. A centroid vector is computed to represent each specialized SE model. In the online phase, the QS or QE of the test utterance is used to identify one cluster of training data, i.e., the corresponding component SE model. Finally, the selected SE model is used to perform enhancement. It is worth noting that there are other reference neural network models that can be used to prepare features for data clustering and model selection. The reason for choosing Quality-Net is that the model is trained to predict the quality score, so it should possess useful speech information.

To evaluate the proposed zero-shot model selection approach, we adopted the perceptual evaluation of speech quality (PESQ) [32] and short-time objective intelligibility (STOI) [33] objective evaluation metrics. Experimental results under both seen and unseen noisy conditions show that the proposed approach can achieve notable improvements compared with the baselines, thereby confirming the effectiveness of the proposed SE approach to provide robust enhancement performance.

This paper is organized as follows. The proposed systems are presented in Section 2. The experiments and results are discussed in Section 3. Finally, the conclusions and future work are presented in Section 4.

2. THE PROPOSED SYSTEMS

In this study, we propose two types of zero-shot model selection strategies, based on quality score (QS), and quality embedding (QE). Both strategies share the similar concept by incorporating Quality-Net as a reference model to extract quality features for performing data clustering and model selection processes. In this section, we will first review the Quality-Net model, and then introduce how to extract QS and QE features with Quality-Net and how to establish the ZMOS-QS and ZMOS-QE systems.

2.1 Quality-Net

Quality-Net is a non-intrusive quality assessment neural network model trained with the aim to predict utterance-level PESQ scores. Since the length of the utterance varies, a bidirectional LSTM (BLSTM) is used to model the longer temporal information. In addition, to achieve a more accurate prediction score and mimic the human perceptive system, a conditional frame-wise constraint is introduced to train the model. Accordingly, the objective function of Quality-Net is derived as follows:

$$O = \frac{1}{n} \sum_{n=1}^N [(Q_n - \hat{Q}_n)^2 + \frac{\alpha(Q_n)}{L(U_n)} \sum_{l=1}^{L(U_n)} (Q_n - q_{n,l})^2] \quad (1)$$

where N and $L(u_n)$ indicate the number of training utterances and the number of frames of the n -th utterance, respectively; Q_n and \hat{Q}_n indicate the true and predicted PESQ scores, respectively; and $q_{n,l}$ and $\alpha(Q_n)$ indicate the estimated frame-level quality of the l -th frame of utterance n and weighting factor, respectively. Finally, given a noisy input y_n , the quality-net equation can be derived as follow:

$$\hat{Q}_n = \text{QualityNet}(y_n), \quad (2)$$

where $\text{QualityNet}(\cdot)$ denotes the PESQ prediction function.

In our previous studies [30, 31], we have confirmed the high prediction capability of the Quality-Net. We believed that both the output scores and latent representations of Quality-Net carry useful information to determine the quality of given speech. This is the main motivation of this study.

2.2. The Proposed System I: ZMOS-QS

The overall system architecture of ZMOS-QS is shown in Fig. 1. In the training stage, we first apply the short time Fourier transform (STFT) to convert speech waveforms to spectral features. With the paired spectral features, $\mathbf{Z}=[\mathbf{X}, \mathbf{Y}]$, formed by noisy spectral features \mathbf{Y} and clean spectral features, \mathbf{X} , PESQ scores are computed and used as a reference to cluster the entire set of training data into several subsets: $\{\mathbf{Z}_1, \dots, \mathbf{Z}_t, \dots, \mathbf{Z}_T\}$, where \mathbf{Z}_t is the t -th subset of paired training data, and T is the total number of subsets. Based on the T subsets of training data, we then estimate T component SE models:

$$\begin{aligned} \mathbf{X}_1 &= F_1(\mathbf{Y}_1), \\ &\dots \\ \mathbf{X}_t &= F_t(\mathbf{Y}_t), \\ &\dots \\ \mathbf{X}_T &= F_T(\mathbf{Y}_T), \end{aligned} \quad (3)$$

where \mathbf{Y}_t , \mathbf{X}_t and F_t are the input, output, and transformation, respectively, of the t -th SE model.

In ZMOS-QS, the training data are clustered based on their PESQ scores predicted by Quality-Net. Specifically, their PESQ scores are ranked. The training utterances with similar PESQ scores are grouped into a subset for training the corresponding component SE model. The average PESQ score of each subset is computed.

In the testing phase, given a noisy speech with the spectral feature, $\tilde{\mathbf{y}}$, its PESQ is estimated by Quality-Net first. Then, the enhancement is carried out by $\tilde{\mathbf{x}} = F_t(\tilde{\mathbf{y}})$, when $\text{QualityNet}(\tilde{\mathbf{y}})$ is closest to the average PESQ score of the t -th component SE model. Finally, an inverse STFT is applied to reconstruct the enhanced speech waveform by the enhanced spectral features, where the phase from the noisy speech is used.

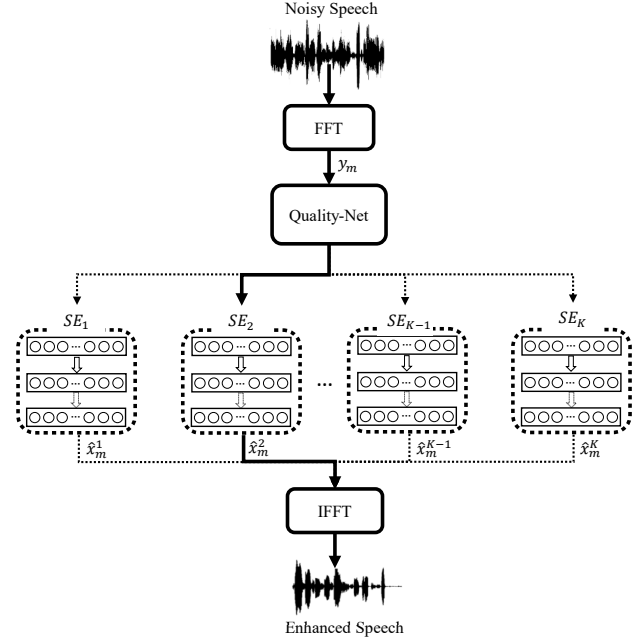


Fig. 1: The architecture of the ZMOS-QS approach.

2.3 The Proposed System II: ZMOS-QE

ZMOS-QE adopts a similar idea of ZMOS-QS. Instead of QS, ZMOS-QE uses the latent representations of Quality-Net to perform data clustering and model selection, as shown in Fig. 2. In the training phase, given noisy spectral features, $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_n, \dots, \mathbf{y}_N]$, where N is the total number of frames, a set of QE features, $\mathbf{Q} = [\mathbf{q}_1, \dots, \mathbf{q}_n, \dots, \mathbf{q}_N]$, is extracted. Next, by applying the K-means algorithm on the entire set of QE features, we can cluster the QE features into T clusters. Accordingly, the training data can be divided into T subsets, $\{\mathbf{Z}_1, \dots, \mathbf{Z}_t, \dots, \mathbf{Z}_T\}$, represented by T centroid QE vectors, $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_t, \dots, \mathbf{v}_T]$, respectively. Then, we prepare T component SE models by:

$$\begin{aligned} \mathbf{X}_1 &= F_1(\mathbf{Y}_1), \\ &\dots \\ \mathbf{X}_t &= F_t(\mathbf{Y}_t), \\ &\dots \\ \mathbf{X}_T &= F_T(\mathbf{Y}_T), \end{aligned} \quad (4)$$

In the testing stage, given a noisy speech with spectral features, $\hat{\mathbf{y}}$, we first compute the QE feature, $\hat{\mathbf{q}}$, by using Quality-Net. Then, we calculate the distance between $\hat{\mathbf{q}}$ and each of the centroid QE features in $[\mathbf{v}_1, \dots, \mathbf{v}_t, \dots, \mathbf{v}_T]$. Next, we perform SE by $\hat{\mathbf{x}} = F_t(\hat{\mathbf{y}})$ if \mathbf{v}_t is closest to $\hat{\mathbf{q}}$. With the enhanced spectral feature, $\hat{\mathbf{x}}$, we can obtain the enhanced waveform by applying ISTFT along with the phase from the noisy speech.

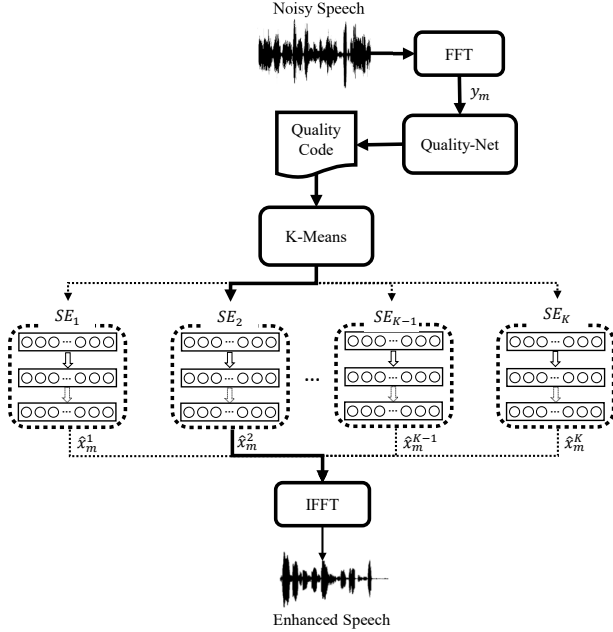


Fig. 2: The architecture of the ZMOS-QE approach.

3. EXPERIMENTS

In this section, we will first present the experimental setup, including dataset preparation and neural-network model architectures. Next, we will present the experimental results of ZMOS-QS and ZMOS-QE and provide discussions on our findings.

3.1 Experimental setup

We adopted the Wall Street Journal (WSJ) [34] dataset to evaluate the proposed ZMOS-QS and ZMOS-QE approaches. The WSJ dataset consists of 37,416 training and 330 test utterances recorded at a 16-kHz sampling rate. We prepared the noisy training utterances by injecting 100 types of stationary and non-stationary noises [35] into the WSJ training utterances at 31 SNR levels ranging from 20 to -10 dB with a step of 1 dB. For the test data, we prepared the noisy utterances by injecting two seen (white and engine noises) and two unseen (car and street noises) noise types at 6 SNR levels (-10, -5, 0, 5, 10, and 15 dB). With a Hamming window of 32 ms and a hop size of 16 ms, a 512-point STFT was performed on the training and test utterances to extract 257-dimensional log-power-spectra (LPS) features.

We compared the proposed approaches with a CNN-based baseline system. The CNN model consists of 12 convolutional layers, followed by a fully connected layer consisting of 128 neurons. Each convolutional layer contains four channels $\{16, 32, 64, 128\}$. Each channel is with three types of strides $\{1, 1, 3\}$. The entire set of training utterances was used to train this CNN-based baseline. The component SE models in ZMOS-QS and ZMOS-QE were implemented

based on the same CNN architecture for a fair comparison. The training data were first divided into several subsets, with each subset used to train a component SE model. In this study, we divided the entire set of training data into 4 clusters. Therefore, there are 4 component SE models.

We used the standardized PESQ and STOI scores to evaluate the proposed ZMOS-QS and ZMOS-QE approaches. PESQ was used to evaluate the quality of speech, with a score ranging from -0.5 to 4.5. STOI was designed to evaluate the intelligibility of speech, with a score ranging from 0 to 1. Higher PESQ and STOI scores indicate that the enhanced speech has better speech quality and intelligibility, respectively.

3.2 Objective evaluation results

The PESQ and STOI scores of unprocessed noisy speech, enhanced speech by CNN baseline, ZMOS-QS, and ZMOS-QE under white and engine noise types are shown in Tables 1 and 2, respectively. These two noise types were seen in the training.

Table 1. PESQ comparison of Noisy, CNN, ZMOS-QS, ZMOS-QE systems under seen noise conditions.

	15	10	5	0	-5	-10	Ave
Stationary (White)							
Noisy	3.55	3.13	2.72	2.34	1.97	1.64	2.56
CNN	3.31	3.22	3.06	2.84	2.60	2.37	2.90
ZMOS-QS	3.38	3.28	3.13	2.88	2.62	2.39	2.95
ZMOS-QE	3.40	3.29	3.11	2.88	2.66	2.47	2.97
Non-Stationary (Engine)							
Noisy	2.61	2.17	1.82	1.57	1.42	1.34	1.82
CNN	2.99	2.76	2.47	2.12	1.74	1.43	2.25
ZMOS-QS	3.09	2.78	2.49	2.13	1.74	1.41	2.28
ZMOS-QE	2.99	2.77	2.50	2.16	1.78	1.45	2.28

Table 2. STOI comparison of Noisy, CNN, ZMOS-QS, ZMOS-QE systems under seen noise conditions.

	15	10	5	0	-5	-10	Ave
Stationary (White)							
Noisy	0.97	0.94	0.89	0.84	0.80	0.76	0.87
CNN	0.91	0.90	0.89	0.87	0.85	0.82	0.87
ZMOS-QS	0.92	0.91	0.90	0.87	0.84	0.82	0.88
ZMOS-QE	0.92	0.91	0.89	0.87	0.85	0.83	0.88
Non-Stationary (Engine)							
Noisy	0.95	0.90	0.82	0.71	0.58	0.46	0.74
CNN	0.91	0.89	0.86	0.81	0.72	0.57	0.79
ZMOS-QS	0.92	0.90	0.86	0.81	0.71	0.57	0.80
ZMOS-QE	0.91	0.89	0.87	0.82	0.73	0.60	0.80

From Tables 1 and 2, we can note that both ZMOS-QS and ZMOS-QE achieve notably better PESQ and STOI scores than the unprocessed noisy speech and the baseline system under all SNR level conditions. Moreover, the proposed approaches can maintain consistent performance improvements in both stationary and non-stationary noisy environments.

Tables 3 and 4 show the PESQ and STOI scores of unprocessed noisy speech, enhanced speech by CNN baseline, ZMOS-QS, and

ZMOS-QE under car and street noise types. These two noise types were not seen in the training. From Tables 3 and 4, we can note again that ZMOS-QS, and ZMOS-QE notably outperform unprocessed noisy speech and the baseline system under all SNR level conditions. The trends in Tables 3 and 4 are similar to those in Tables 1 and 2. In addition, the experimental results show that ZMOS-QS slightly outperforms ZMOS-QE under higher SNR level conditions, while ZMOS-QE is better than ZMOS-QS under lower SNR level conditions. Overall, the results confirmed the effectiveness of the proposed approaches for robust speech enhancement performance.

Table 3. PESQ comparison of Noisy, CNN, ZMOS-QS, ZMOS-QE systems under unseen noise conditions.

	15	10	5	0	-5	-10	Ave
Stationary (Car)							
Noisy	3.36	2.93	2.51	2.12	1.79	1.56	2.38
CNN	3.25	3.12	2.92	2.61	2.27	1.95	2.69
ZMOS-QS	3.34	3.24	2.97	2.61	2.28	1.96	2.73
ZMOS-QE	3.33	3.17	2.94	2.66	2.36	2.04	2.75
Non-Stationary (Street)							
Noisy	2.92	2.48	2.09	1.78	1.55	1.46	2.05
CNN	3.15	2.98	2.71	2.31	1.88	1.56	2.43
ZMOS-QS	3.31	3.10	2.75	2.31	1.87	1.52	2.48
ZMOS-QE	3.16	2.99	2.75	2.40	1.97	1.60	2.48

Table 4. STOI comparison of Noisy, CNN, ZMOS-QS, ZMOS-QE systems under unseen noise conditions.

	15	10	5	0	-5	-10	Ave
Stationary (Car)							
Noisy	0.97	0.93	0.88	0.82	0.75	0.69	0.84
CNN	0.91	0.90	0.88	0.86	0.82	0.77	0.86
ZMOS-QS	0.92	0.91	0.89	0.85	0.81	0.76	0.86
ZMOS-QE	0.92	0.91	0.89	0.86	0.83	0.78	0.86
Non-Stationary (Street)							
Noisy	0.97	0.93	0.86	0.77	0.66	0.56	0.79
CNN	0.91	0.90	0.88	0.84	0.76	0.65	0.82
ZMOS-QS	0.93	0.92	0.89	0.83	0.75	0.62	0.82
ZMOS-QE	0.92	0.90	0.88	0.84	0.77	0.66	0.83

3.3 Spectrogram analysis

In addition to the objective evaluations, we present the spectrograms to visualize the processed speech. Fig. 3 shows the spectrograms of a clean utterance (top left), the corresponding noisy utterance at 0 dB SNR under car-noise (top right), the enhanced speech by the CNN baseline (bottom left), and the enhanced speech by ZMOS-QE (bottom right). We only present the resulting spectrogram of ZMOS-QE because both ZMOS-QS and ZMOS-QE output similar results. From Fig. 3, we can first confirm the effectiveness of the CNN baseline for SE. Next, the proposed ZMOS-QE model can yield even better noise reduction results and more accurately recover the speech, compared with the CNN baseline, as can be seen in the red box that the speech processed by ZMOS-QE retains more detailed speech information than the CNN baseline.

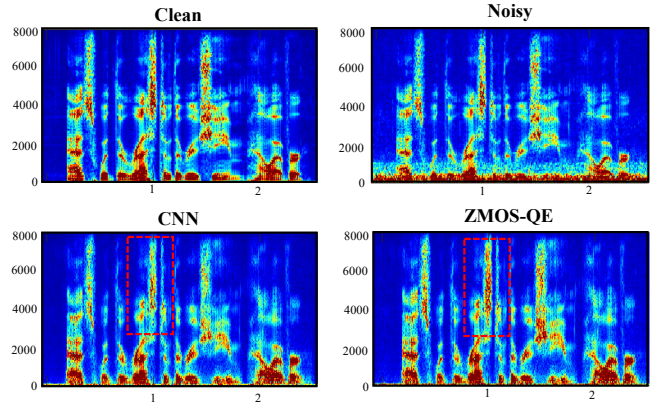


Fig. 3: Spectrograms of a clean utterance (Clean), along with its noisy version (car noise at 0 dB SNR) (Noisy), and the CNN baseline and ZMOS-QE enhanced ones.

4. CONCLUSIONS

In this paper, we proposed two zero-shot model selection approaches for SE, namely ZMOS-QS and ZMOS-QE. The proposed approaches are derived based on zero-shot learning and ensemble learning. The quality score and embedding from the Quality-Net were used to perform data clustering and model selection. Experimental results have confirmed that the proposed approaches effectively improve the SE performance of the baseline system, based on which the proposed approaches are built. To the best of our knowledge, this work is the first attempt to perform zero-shot learning on SE and has successfully improved performance. In the future, we will explore the applicability of the proposed zero-shot model selection approaches in other speech-processing tasks, such as dereverberation or cross-corpus speech enhancement tasks.

5. REFERENCES

- [1] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.
- [2] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, "Robust automatic speech recognition: a bridge to practical applications." *Academic Press*, 2015.
- [3] Z.-Q. Wang and D. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 796–804, 2016.
- [4] P. C. Loizou, "Speech enhancement: theory and practice," *CRC Press*, 2007.
- [5] D. Wang, "Deep learning reinvents the hearing aid," *IEEE Spectrum*, vol. 54, no. 3, pp. 32–37, 2017.
- [6] G. S. Bhat and C. K. Reddy, "Smartphone-based real-time super gaussian single microphone speech enhancement to improve intelligibility for hearing aid users using formant information," in *Proc. EMBC*, pp. 5503–5506, 2018.
- [7] H. Levitt, "Noise reduction in hearing aids: an overview," *Journal of Rehabilitation Research and Development*, vol. 38, pp. 111–121, 2001.
- [8] F. Chen, Y. Hu, and M. Yuan, "Evaluation of noise reduction methods for sentence recognition by Mandarin-speaking cochlear implant listeners," *Ear and Hearing*, vol. 36, no. 1, pp. 61–71, 2015.
- [9] R. Martin and R. V. Cox, "New speech enhancement techniques for low bit rate speech coding," in *Proc. IEEE Workshop on Speech Coding*, pp. 165–167, 1999.

- [10] Z. Zhao, H. Liu, and T. Fingscheidt, "Convolutional neural networks to enhance coded speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 663–678, 2019.
- [11] S. Shon, H. Tang, and J. Glass, "Voiceid loss: speech enhancement for speaker verification," *arXiv preprint arXiv:1904.03601*, 2019.
- [12] M. Kolbk, Z.-H. Tan, and J. Jensen, "Speech enhancement using long short-term memory based recurrent neural networks for noise-robust speaker verification," in *Proc. SLT*, pp. 305–311, 2016.
- [13] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.
- [14] D. Liu, P. Smaragdis, and M. Kim, "Experiments on deep learning for speech denoising," in *Proc. INTERSPEECH*, pp.2685-2689, 2014.
- [15] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. INTERSPEECH*, pp.436-440, 2013.
- [16] P. G. Shivakumar and P. G. Georgiou, "Perception optimized deep denoising autoencoders for speech enhancement," in *Proc. INTERSPEECH*, 2016, pp. 3743–3747.
- [17] S.-W. Fu, Y. Tsao, and X. Lu, "Snr-aware convolutional neural network modelling for speech enhancement," in *Proc. INTERSPEECH*, pp. 3768-3772, 2016.
- [18] S.-W. Fu, T.-W. 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.
- [19] A. Pandey and D. Wang, "A new framework for CNN-based speech enhancement in the time domain," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 27, no. 7, pp. 1179–1188, 2019.
- [20] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *Proc. HSCMA*, 2017.
- [21] Z. Chen, S. Watanabe, H. Erdogan, and J. R. Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," in *Proc. INTERSPEECH*, pp. 3274-3278, 2015.
- [22] C.H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. IEEE CVPR*, pp. 951–958, 2009.
- [23] J. Liu, B. Kuipers, and S. Savarese, "Recognizing human actions by attributes," in *Proc. IEEE CVPR*, pp. 3337–3344, 2011.
- [24] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no.3, pp.453–465, 2014.
- [25] D. Jayaraman and K. Grauman, "Zero-shot recognition with unreliable attributes," in *Proc. NIPS*, pp. 3464–3472, 2014.
- [26] Z. Al-Halah, M. Tapaswi, and R. Stiefelhagen, "Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learnin," in *Proc. IEEE CVPR*, pp. 5975–5984, 2016.
- [27] X. Li, S. Dalmia, D. R. Mortensen, F. Metze, and A. W. Black, "Zero-shot learning for speech recognition with universal phonetic model," 2019. [Online]. Available: <https://openreview.net/forum?id=BkfhZnC9>
- [28] E. Cooper, C.-I. Lai, Y. Yasuda, F. Fang, X. Wang, N. Chen, and J. Yamagishi, "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," *arXiv preprint arXiv:1910.10838*, 2019.
- [29] A. Khoreva, R. Benenson, E. Ilg, T. Brox, and B. Schiele, "Lucid data dreaming for object tracking," *arXiv preprint arXiv: 1703.09554*, 2017.
- [30] S.-W. Fu., Y. Tsao, H.-T. Hwang, and H.-W. Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on BLSTM," in *Proc. INTERSPEECH*, pp. 1873–1877, 2018.
- [31] R. E. Zezario, S.-W. Fu, X. Lu, H.-M. Wang, and Y. Tsao, "Specialized Speech Enhancement Model Selection Based on Learned Non-Intrusive Quality Assessment Metric," in *Proc. INTERSPEECH*, pp.3168-3172, 2019.
- [32] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, pp. 749–752, 2001.
- [33] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.
- [34] D. Paul and J. Baker, "The design for the wall street journal-based csr corpus," in *Proc. ICSLP*, pp. 899–902, 1992.
- [35] D. Hu, "100nonspeechenvironmentalsounds2004[online]," <http://www.cse.ohio-state.edu/pnl/corpus/HuCorpus.html>, 2004