Speech Enhancement with Zero-Shot Model Selection

Ryandhimas E. Zezario*
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
ryandhimas@citi.sinica.edu.tw

Chiou-Shann Fuh
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
fuh@csie.ntu.edu.tw

Hsin-Min Wang
Institute of Information Science
Academia Sinica
Taipei, Taiwan
whm@iis.sinica.edu.tw

Yu Tsao
*Research Center for Information Technology Innovation
Academia Sinica
Taipei, Taiwan
yu.tsao@citi.sinica.edu.tw

***Abstract*—Recent research on speech enhancement (SE) has seen the emergence of deep-learning-based methods. It is still a challenging task to determine the effective ways to increase the generalizability of SE under diverse test conditions. In this study, 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 the 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 perform the enhancement. Furthermore, two selection strategies were developed: selection based on the quality score (QS) and selection based on the quality embedding (QE). Both QS and QE were obtained using a Quality-Net, a non-intrusive quality assessment network. Experimental results confirmed that the proposed ZMOS approach can achieve better performance in both seen and unseen noise types compared to the baseline systems and other model selection systems, which indicates the effectiveness of the proposed approach in providing robust SE performance.**

Keywords— speech enhancement, deep learning, zero-shot learning, model selection.

I. 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 signals. With the emergence of deep learning algorithms, notable improvements in 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) [15–17], convolutional neural network (CNN) [18–19], long short-term memory (LSTM) [20–21] and their combinations [22–24]. Despite past promising improvements, increasing the generalizability of deep-learning-based SE methods to unseen environments remains a critical research topic.

Zero-shot learning is a machine learning algorithm that 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]. In the field of speech processing, several attempts have been made to incorporate a zero-shot learning algorithm for robust performance [27–30]. For instance, in [29], speaker embedding was extracted and used as additional guidance to conduct noise reduction. In [30], noise embedding, namely the dynamic noise embedding, was extracted and used to characterize background noise information to develop more optimal noise reduction performance. However, most of the current zero-shot learning strategies rely on a similar fashion, where the generated latent representation is incorporated as an additional feature into the main task. In addition, due to the notable success of model selection approaches [31–32], we aim to use zero-shot learning as a model selection approach. To the best of our knowledge, no prior work has proposed the use of latent representations for model selection in speech enhancement tasks.

In this study, 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 condition and is implemented in two phases: offline and online. In the offline phase, we prepared multiple specialized SE models (termed component SE models). Each component SE model was trained to match the specific noisy condition. In the online phase, we selected the most suitable component SE model to enhance the test utterance. For the proposed approach, the effective clustering of the training data to train the multiple-component SE models in the offline phase and selecting 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 using a pre-trained Quality-Net [33]. A Quality-Net is a deep-learning-based non-intrusive quality assessment model. Given an utterance, the Quality-Net outputs a quality assessment score. Previous studies have shown that the 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 as ZMOS-QS and ZMOS-QE, respectively. Both the QS and QE were estimated using the Quality-Net. Given an utterance, the QS is based on the output score of the Quality-Net, and the QE is based on the embedding vector of Quality-Net. In the offline phase, QS or QE was used to group the training data into several clusters. Each cluster was used to train a specialized SE model. A centroid vector was 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 was used to perform the

enhancement. Notably, the other reference neural network models can be used to prepare features for data clustering and model selection. The Quality-Net was chosen because the model was trained to predict the quality score, so it should possess useful speech information.

To evaluate the proposed zero-shot model selection (ZMOS) approach, we adopted the perceptual evaluation of speech quality (PESQ) [34] and short-time objective intelligibility (STOI) [35] 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 and other model selection approaches, thereby confirming the effectiveness of the proposed SE approach in providing robust enhancement performance.

II. The Proposed Systems

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

A. Quality-Net

Quality-Net is a non-intrusive quality assessment neural network model trained with the aim of predicting utterance-level PESQ scores. As 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 the 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 follows:

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

where $QualityNet(.)$ denotes the PESQ prediction function.

In our previous studies [32–33], we have confirmed the high prediction capability of the Quality-Net. We believe that both the output scores and latent representations of the Quality-Net provide useful information for determining the quality of given speech. This was the main motivation for this study.

B. The Proposed System I: ZMOS-QS

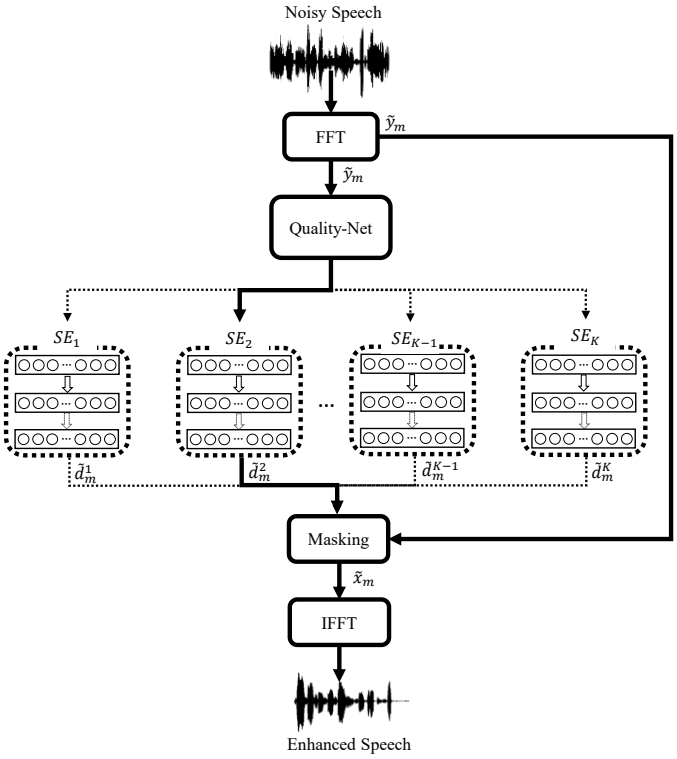

Fig. 1: The architecture of the ZMOS-QS approach.

The overall system architecture of the ZMOS-QS is shown in Fig. 1. In the training stage, we first apply the short-time Fourier transform (STFT) to convert speech waveforms into spectral features. With the paired spectral features, $\boldsymbol{Z}$=[$\boldsymbol{X}$, $\boldsymbol{Y}$], which are formed by noisy spectral features $\boldsymbol{Y}$ and clean spectral features $\boldsymbol{X}$, PESQ scores are computed. They are used as a reference to cluster the entire set of training data into several subsets: $\{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_t, \ldots \boldsymbol{Z}_T\}$, where $\boldsymbol{Z}_t$ is the t-th subset of paired training data, and T is the total number of subsets. Based on the T subsets of the training data, we then estimate the T-component SE models with an ideal ratio mask (IRM) [36] in the log domain as the training target criterion:

$$\begin{aligned} \boldsymbol{D}_1 &= F_1(\boldsymbol{Y}_1), \\ &\cdots \\ \boldsymbol{D}_t &= F_t(\boldsymbol{Y}_t), \\ &\cdots \\ \boldsymbol{D}_T &= F_T(\boldsymbol{Y}_T), \end{aligned} \quad (3)$$

where $\boldsymbol{Y}_t$, $\boldsymbol{D}_t$ 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 the Quality-Net. Specifically, the PESQ scores were ranked. The training utterances with similar PESQ scores were grouped into a subset for training the corresponding component SE model. The average PESQ score for each subset was computed.

In the testing phase, given a noisy speech with the spectral feature $\tilde{\boldsymbol{y}}$, its PESQ is first estimated by the Quality-Net. Then,

the enhancement is carried out by $\tilde{\boldsymbol{d}} = F_t(\tilde{\boldsymbol{y}})$, when $QualityNet(\tilde{\boldsymbol{y}})$ is closest to the average PESQ score of the t-th component SE model and enhanced spectral feature $\tilde{\boldsymbol{x}} = F_M(\tilde{d}, \tilde{\boldsymbol{y}})$, where F_M is the masking function. Finally, an inverse STFT is applied to reconstruct the enhanced speech waveform using enhanced spectral features, where the phase from the noisy speech is used.

C. The Proposed System I: ZMOS-QE

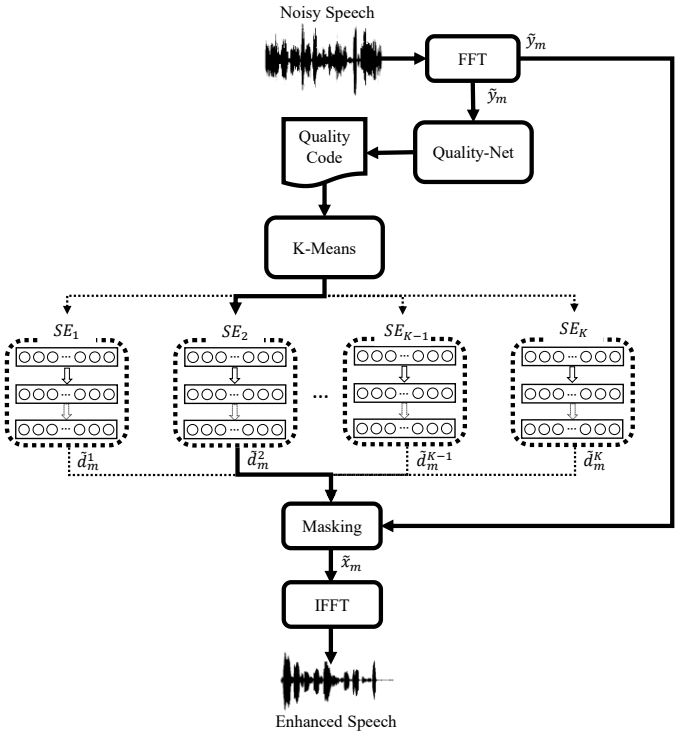

Fig. 2: The architecture of the ZMOS-QE approach.

ZMOS-QE adopts a similar idea to ZMOS-QS. Instead of QS, ZMOS-QE uses the latent representations of the Quality-Net to perform the data clustering and model selection, as shown in Fig. 2. In the training phase, given noisy spectral features, $\boldsymbol{Y} = [\boldsymbol{y}_1, \dots, \boldsymbol{y}_n \dots, \boldsymbol{y}_N]$, where N is the total number of frames, a set of QE features, $\boldsymbol{Q} = [\boldsymbol{q}_1, \dots, \boldsymbol{q}_n \dots, \boldsymbol{q}_N]$, is extracted. Next, by applying the K-means algorithm to 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, $\{\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_t, \dots \boldsymbol{Z}_T\}$, represented by T centroid QE vectors, $\boldsymbol{V} = [\boldsymbol{v}_1, \dots, \boldsymbol{v}_t, \dots \boldsymbol{v}_T]$, respectively. Then, we prepared T-component SE models, as shown in Eq. 3.

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

III. Experiments

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

A. Experimental Setup

We adopted the Wall Street Journal (WSJ) [37] 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 [38] into the WSJ training utterances at 31 signal-to-noise ratio (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 five SNR levels (-10, -5, 0, 5, and 10 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 features.

We compared the proposed approaches with a CNN-based baseline system. The CNN model consisted of 12 convolutional layers, followed by a fully connected layer consisting of 128 neurons. Each convolutional layer contains four channels {16, 32, 64, and 128}. Each channel has three types of strides: {1, 1, 3}. The entire set of training utterances was used to train the CNN-based baseline. The component SE models in the 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 four clusters. Therefore, there are four 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.

B. Objective Evaluation Results

The average 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. These two noise types were seen in the training. For comparison, we implemented and tested performance using another model selection method, specialized Speech Enhancement Model Selection (SSEMS) [32], in which the component models are trained to learn gender and signal-to-noise-ratio (SNR) information instead of the data-driven approach used in ZMOS. From Tables 1, we can note that both ZMOS-QS and ZMOS-QE achieve notably better PESQ and STOI scores than the unprocessed noisy speech, the baseline CNN system, and the SSEMS in both stationary and non-stationary noisy environments.

Table 1. PESQ and STOI comparison of Noisy, CNN, SSEMS, ZMOS-QS, ZMOS-QE systems under seen noise conditions (white and engine noises).

	PESQ	STOI
Noisy	2.01	0.77
CNN	2.42	0.78
SSEMS [32]	2.43	0.78
ZMOS-QS	2.46	0.79
ZMOS-QE	2.52	0.79

Table 2. PESQ and STOI comparison of Noisy, CNN, SSEMS, ZMOS-QS, ZMOS-QE systems under unseen noise conditions (car and street noises).

	PESQ	STOI
Noisy	1.71	0.68
CNN	2.49	0.80
SSEMS [32]	2.53	0.80
ZMOS-QS	2.51	0.81
ZMOS-QE	2.57	0.80

Table 2 shows the average PESQ and STOI scores of the unprocessed noisy speech, the enhanced speech by CNN baseline, SSEMS, ZMOS-QS, and ZMOS-QE for car and street noise types. These two noise types were not observed during the training. From Table 2, we can again note that ZMOS-QE achieves considerably better performance compared to the other systems. The ZMOS-QS achieved better performance than the baseline systems and comparable performance with the SSEMS systems. Overall, the results confirm the effectiveness of the proposed approach for robust speech enhancement (SE) performance.

C. Model Selection Analysis

In the previous section, we demonstrated the effectiveness of the proposed method for noise reduction. In particular, we demonstrated the effectiveness of using latent representations to develop the component models and perform the model selection. Based on the notable performances achieved by ZMOS-QE, we conducted additional evaluations, where the comparative systems adopted the same component models as those used in ZMOS-QE but different model selection strategies.

Two other systems, namely the auto-encoder-based approach DAE [31] and SSEMS-QE [32], were developed. The DAE selects the best candidate based on the reconstruction error of the auto-encoder. Meanwhile, SSEMS-QE selects the best candidate based on the highest PESQ score given several component models. In contrast to the original SSEMS, SSEMS-QE adopted the quality embedding-based component models as those used in the ZMOS-QE approach. As shown in Figs. 3 and. 4, the proposed ZMOS-QE consistently overcomes the other selection methods in terms of PESQ and STOI scores under seen and unseen noises. Interestingly, unlike the other selection methods that require computing all possible component models to select the most suitable model, our proposed method can use only one utterance to identify the best fit model. Therefore, it

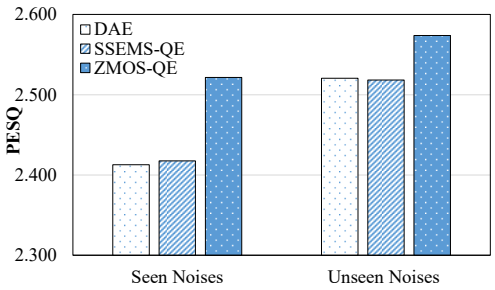

Fig.3: PESQ comparison of DAE, SSEMS-QE, and ZMOS-QE under seen and unseen conditions.

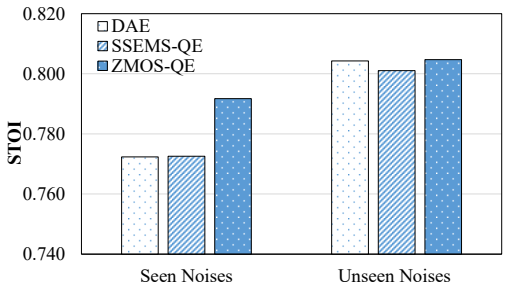

Fig.4: STOI comparison of DAE, SSEMS-QE, and ZMOS-QE under seen and unseen conditions.

can reduce the computational cost but yet still maintains better enhancement performances.

D. Spectrogram Analysis

In addition to the objective evaluations, we present the spectrograms to visualize the processed speech. Fig. 5 shows the spectrograms of a clean utterance (top left), corresponding noisy utterance at 0 dB SNR under car-noise (top right), enhanced speech by the CNN baseline (bottom left), and enhanced speech by ZMOS-QE (bottom right). We present the resulting spectrogram of ZMOS-QE only because ZMOS-QE has consistently achieved more effective reduction performance. From Fig. 5, we can confirm the effectiveness of the CNN baseline for SE. The proposed ZMOS-QE model can yield even better noise reduction results and recover the speech more accurately, compared with the CNN baseline, as seen in the red box; the speech processed by ZMOS-QE retains more detailed speech information than the CNN baseline.

IV. Conclusions

In this study, we proposed two zero-shot model selection approaches for SE: ZMOS-QS and ZMOS-QE. The proposed approaches were 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 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 as a model selection for SE and has improved the performance. In the future, we will explore the applicability of the proposed ZMOS approaches in other speech-processing tasks, such as dereverberation or cross-corpus SE tasks.

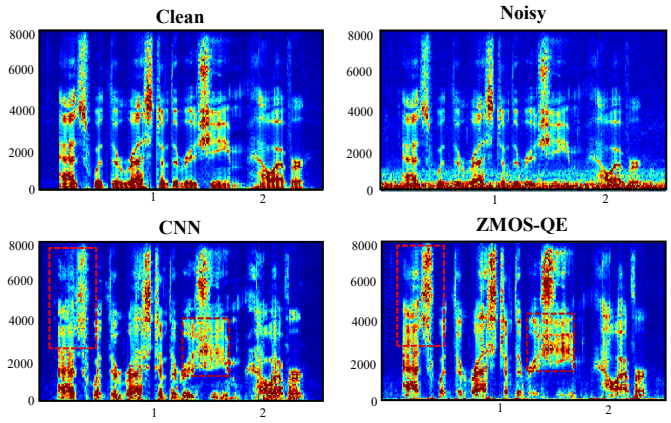

Fig.5: Spectrograms of a clean utterance (Clean), along with its noisy version (car noise at 0 dB SNR) (Noisy), and the CNN baseline and ZMOS-QE enhanced ones.

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] R. E. Zezario, T. Hussain, X. Lu, H. Wang, and Y. Tsao, "Self-supervised denoising autoencoder with linear regression decoder for speech enhancement," *in Proc. ICASSP*, pp. 6669–6673, 2020.

[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] F.-K. Chuang, S.-S. Wang, J.-w. Hung, Y. Tsao, and S.-H. Fang, "Speaker-aware deep denoising autoencoder with embedded speaker identity for speech enhancement," *in Proc. INTERSPEECH*, pp. 3173–3177, 2019

[30] J. Lee, Y. Jung, M. Jung, and H. Kim, "Dynamic noise embedding: Noise aware training and adaptation for speech enhancement," *arXiv preprint arXiv:2008.11920,* 2020.

[31] M. Kim, "Collaborative deep learning for speech enhancement: A run-time model selection method using autoencoders," *in Proc. ICASSP*, pp. 76–80, 2017.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," *In Proc. ICASSP*, pp. 7092–7096, 2013

[37] D. Paul and J. Baker, "The design for the wall street journal-based csr corpus," *in Proc. ICSLP*, pp. 899–902, 1992.

[38] D. Hu, "100nonspeechenvironmentalsounds2004[online]," http://www.cse.ohio-state.edu/pnl/corpus/HuCorpus.html, 2004