Speech Enhancement Based on Masking Approach Considering Speech Quality and Acoustic Confidence for Noisy Speech Recognition

Shih-Chuan Chu, Chung-Hsien Wu, Yun-Wen Lin
Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan

***Abstract*—In recent years, voice-operated applications have been widely accepted by the public, while the background noise is still a challenging issue for automatic speech recognition (ASR). This paper proposes a mask-based speech enhancement front-end approach taking into account speech quality score as well as acoustic confidence for speech enhancement to reduce the word error rate (WER) for noisy speech recognition. First, the features of speakers, phones, and noises are extracted and considered in the loss function to improve the speech quality. In addition to speech quality, this study also considers the phone confidence from the Kaldi-based ASR into the loss function for training the mask generation model to improve speech quality as well as improve noisy speech recognition performance. Compared with the baseline model, the proposed model not only improved STOI by 2.14% and PESQ by 7.22%, but also reduced 12.13% in WER.**

Index Terms: Noisy speech recognition, speech enhancement, ideal ratio mask, phone confidence, STOI, PESQ.

I. INTRODUCTION

Currently, automatic speech recognition (ASR) for voice assistants in mobile phones, smart speakers, and auto-pilot functions in cars or a navigation system have been popularly used to help users to execute instructions or tasks through speech. But this is only at the tip of the iceberg in many applications. In real applications, many devices should work in an open environment, and there are many noise sources in the environment at the same time. We can imagine that if we input a large number of types of noise and speech at the same time, it will be very challenging for speech recognition. For humans, neural adaptation may ultimately help enhance the robustness of speech in noise, but for machines, there is still a lot of room for performance improvement in noisy speech recognition.

In recent years, in the field of speech processing, most people regard speech separation and speech enhancement [1-5] as classification tasks, that is, to classify speech and noise in audio, so a lot of the methods based on data-driven and larger model have been widely studied [6, 24-26]. While most training models do classification tasks, they basically infer a part of each frame as noise, and treat the rest as speech, and then further remove or reduce the part that is classified as noise. But if we group all types of noise (such as factory, car, and street noise) into one group, and treat the rest of the audio as the desired speech, it is actually difficult for the trained model to find the actual type of noise, or even the noise is not in the training set. Therefore, when the target is regarded as a classification task, in order to improve the classification accuracy and provide more audio-related and useful information, the characteristics of speaker [7-9], noise [10, 11] and phone [8, 12, 13] in the nosy speech should be considered. Many previous studies [14-16] used mean square error (MSE) as the loss function for training the mask-based speech enhancement model. Although the MSE-based approach improved the speech quality, it may not match the human auditory perception as well as the ASR system. However, when the devices cannot adapt to voice signals like humans, they need to rely on classification technology to provide external information to help machine learning and understanding. Therefore, we need to consider the feature of phone and speaker in speech and the function of noise in the environment at the same time to obtain more information to enhance speech.

According to the above discussion, we can find some problems. First, some approaches used either the feature of speaker or phone, instead of both. In [9], Shon et al. observed that because each speaker has different speaking and pronunciation features, the model is more likely to extract higher-level abstract features in the process of feature extraction, such as a wide variety of phone category features. It also addressed the characteristics for different phone features of the same speaker and different speaker features of the same phone. When performing speech enhancement, speaker features and phone features should be considered at the same time.

On the other hand, Fu et al. [17] mentioned that the MSE loss function has a certain gap with the current actual evaluation methods, such as PESQ and STOI. Therefore, they proposed the loss function that cooperates with the two evaluation metrics to narrow the gap. But in fact, the difference between MSE, PESQ/STOI and the recognition rate of the ASR system has not been solved. Having a high PESQ/STOI does not necessarily mean that a high recognition accuracy from the ASR system can be obtained. Part of the reason for this comes from the fact that the ASR is another system. During the training process of the ASR system, there is no consideration of PESQ or STOI, so the trained model is not directly related to PESQ and STOI. The other part is that, whether you use the MSE, PESQ loss or STOI loss, the mask generated by the model may mask the spectrogram that is important to ASR. There is no way to directly judge whether the current loss is helpful to improve the recognition rate. Therefore, there is a gap between the result of speech enhancement and ASR.

II. Proposed Method

In order to solve the above-mentioned problems, the mask-based speech enhancement system that uses the features of speaker and phone as shown Fig. 1 is proposed. The proposed mechanism is trained with PESQ, STOI and phone confidence as the loss function to improve both speech quality and speech recognition rate. Phone confidence is the probability of all phones in the ASR acoustic model classification result. The training phase includes three parts: feature extraction, mask generation modeling, and judgment model construction. Phone feature extraction model, speaker feature extraction model, and noise estimation model are constructed to extract the features of phone, speaker, and noises from the spectrogram of the noisy speech signal, respectively. The features are then used as the input of the noise mask generation model to generate a mask, and the spectrogram of the noisy speech is masked to obtain an enhanced spectrogram. The masked spectrogram is then fed to the judgment model for evaluation, and further the loss is back propagated to train the mask generation model. In the test phase, the speech input is fed to the trained speech enhancement model to obtain the enhanced speech, and finally sent to the ASR for recognition.

2.1. Phone Feature Extraction Model

As described in [12], it is difficult to classify phones in noisy speech. An encoder and a decoder are adopted by using two fully connected (FC) layers in front of the CNN to project information to another domain, so that the model can classify phones more effectively. At the same time, the model does not use any pooling, but uses all convolutional layers to reduce the loss of information due to the pooling operation [18].

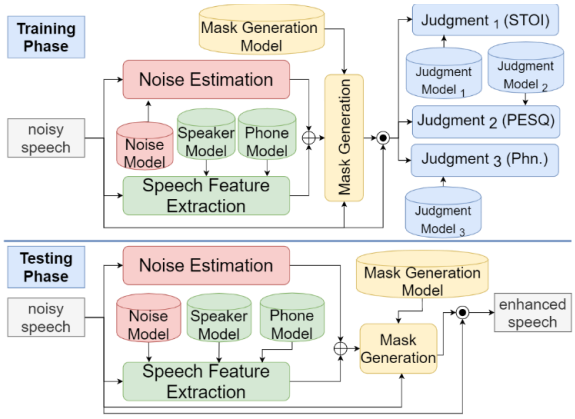

Fig. 1 System framework

The audio spectrogram of the noisy speech segment with a window of five frames is first passed through two FC layers and then six two-dimensional convolutional layers. In the convolutional layer, each layer uses the ReLU as the activation function, and the kernel size is 3*5. Each layer is padded with zero at the boundary, and the stride is 1 to ensure a fixed dimensionality. Then, the two FC layers are used to reduce the output dimension to 128. Finally, the outputs of the last layer conforming to 40 phone classes provide the phone class output with respect to the current frame, as shown in Fig. 2.

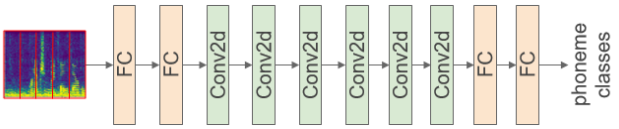

Fig. 2 Architecture of phone feature extraction model

2.2. Speaker Feature Extraction Model

In addition to phone features, speaker features from noisy speech are extracted based on the speaker embedding model with time-delay neural networks (TDNN) presented in the study of Hong et al. [19].

2.3. Noise Estimation Model

As well as the features of phone and speaker, we also consider the noise features which include the SNR level and the noise type. The multi-tasking learning approach proposed by Fu et al. [20] is employed to establish a noise feature model, as shown in Fig. 3. During the training process, two CNNs are used to extract the features for the noisy speech spectrogram as well as the time-domain signal segment of the current frame, respectively. When the training of this model is completed, the output in the last hidden layer is used as the noise features.

After that, we concatenate the output obtained on both sides, and then process it with two FC layers. The first layer outputs a 128-dimensional feature vector, and the last FC layer outputs the classification results of the SNR level and noise type. The training target is a multi-label target, and we will also use entropy to deduce the probability of true label difference, and the binary cross entropy is used as the loss function.

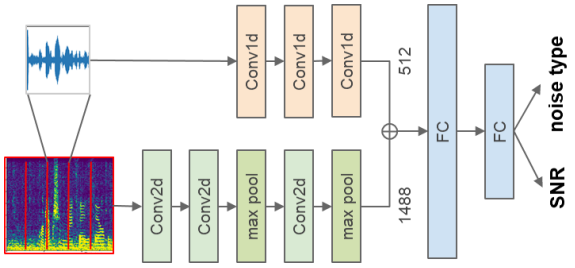

Fig. 3 Architecture of noise estimation model

2.4. Judgment Model

In order to judge whether the quality of the enhanced speech has been improved or the phones can be correctly recognized, an objective evaluation on the speech quality model and a phone-based judgment model are used in the training stage. First, two models estimating the values of PESQ and STOI are constructed, respectively. The two architectures are the same. Both are six-layer DNN with noisy speech spectrogram of 5 frames as input, and the final output is the value of normalized PESQ or STOI predicted by the model.

Second, the phone-based judgment model architecture is the same as the afore mentioned phone feature extraction model

architecture. The difference is that when this model is trained, the outputs of the last layer of the model are regarded as the recognition probabilities of the phone classes.

Finally, in order to make the enhanced speech have a better recognition rate and avoid the excessive difference between the speech quality and the recognition rate, here we further directly use the result of the Kaldi ASR system as the weight of the loss function. Because the ASR system is another independent system, the returned text cannot be directly used as a loss to train the mask generation model. So, we calculate the MSE between the enhanced spectrogram and clean spectrogram. The phone confidence obtained from the ASR system, subtracted by 1, is used as the weight which is then multiplied by the difference of each frame as the loss.

2.5. Mask Generation Model

The purpose of the mask generation model is to generate the ideal ratio mask (IRM) of 5 frames. The features of phones, speakers and noises are concatenated with noisy speech spectrogram as the input. The dimensions for the features of phones, speakers, and noise are 128, 512, and 128, respectively. Concatenating 1285 dimensions of speech segment of five frames, the features with a total of 2053 dimensions is obtained.

The structure of the mask generation model is a 6-layer DNN; each layer has an output with 1024 dimensions, and the last layer applies the sigmoid function to convert the output value to a value between 0 and 1 to meet the boundary of IRM. And the output dimension of the mask is 1285 covering five frames.

In the training phase, the loss function except the traditional MSE can be divided into three parts: L_1, L_2 and L_3. First, L_1 is speech quality average loss obtained from the objective judgment model defined in (1).

$$L_1 = \frac{|1-M_{stoi}(\hat{y}_i)|+|1-M_{pesq}(\hat{y}_i)|}{2} \quad (1)$$

where $\hat{y}_i$ means the enhanced spectrogram, M_{stoi} and M_{pesq} are the normalized scores estimate by judgement_1 and judgement_2. L_2 is the acoustic phone judgment loss which uses the enhanced results for phone identification defined in (2) and (3)

$$p_i = softmax(M_{phn_{jud}}(\hat{y}_i)) \quad (2)$$

$$L_2 = |1 - {p_i}^{gt}| \quad (3)$$

p_i is the probability of correct phone category gt, M_{phn_jud} is the result of the phone-based judgment model, judgement_3 shown in Fig. 1. Acoustic phone confidence loss L_3 refers to the use of the ASR system and obtains the confidence level of the acoustic phone with the help of the language model, as shown in (4). We also hope that the confidence level is close to 1, so L_3 is rewritten as (5). $\hat{y}_i'$ is the time-domain signal after noise reduction.

$$phone\ confidence_i = ASR(\hat{y}_i') \quad (4)$$

$$L_3 = |1 - phone\ confidence_i| \quad (5)$$

Finally, in order to find the best loss function, we consider the combination of attention weight and multi-task method. Among them, the value of L multiplied by the MSE is based on an attention-like approach. After calculating the squared error for a set of batches, multiplication is performed before average. In addition, (10)-(12) adopt the multi-task proportional addition method.

$$J_1 = MSE * L_1 \quad (6)$$

$$J_2 = MSE * L_3 \quad (7)$$

$$J_3 = MSE * L_1 * L_2 \quad (8)$$

$$J_4 = MSE * L_1 * L_3 \quad (9)$$

$$J_5 = \lambda_1 MSE * L_2 + (1-\lambda_1)L_1 \quad (10)$$

$$J_6 = \lambda_1 MSE * L_3 + (1-\lambda_1)L_1 \quad (11)$$

$$J_7 = \lambda_1 MSE * \lambda_2 L_2 + (1-\lambda_1-\lambda_2)L_1 \quad (12)$$

III. Experimental Results

3.1. Datasets

There were two data sets used in this study; one was the speech data set, and the other was the noise data set. The speech data set was the DARPA-TIMIT [21], an English data set with a sampling rate of 16 kHz. It consisted of 630 speakers, each speaking ten sentences. And each audio file had strong labels of words and phones. In the noise data set, the NoiseX-92 [22] data set was used. There were 15 audio files, each with different type of noises. The sampling rate was 19.8 kHz, and the total length of each audio was 235 seconds.

The training data used in this study were the first 250 sentences from the TIMIT training data subset. Five types of noise in NoiseX-92, including volvo, pink, f16, babble and destroyerops, were used. The SNR with five levels: -10dB, -5dB, 0dB, 5dB, and 10dB, were adopted. In the test data, the first 25 sentences from the TIMIT test subset combining with the two noise types of pink and destroyerops in NoiseX-92. The SNR with five levels: -10dB, -5dB, 0dB, 5dB, and 10dB, were considered to obtain a total of 250 audio test data. The audio files used in the experiment were resampled to 16 kHz. The length of the window for STFT was 512, with an overlap of 256 samples. Furthermore, for the phone label, we relabeled the data by the ASR system using CMU [23] phone set pre-training.

3.2. Experimental results

We evaluated each model in the framework separately, and compared with the baseline model without adding extra features and using the judgment models

3.2.1. Evaluation of Phone Feature Extraction Model

In order to extract the feature of the phone, we built a model using FC and convolutional layers. We also used the network output of the last hidden layer as the features of phones. The model was trained with the 40 phones represented by IPA CMU. The phone recognition results are shown in TABLE I.

TABLE I: The accuracy (%) of phone recognition

SNR / Noise type	10dB	-5dB	0dB	5dB	10dB	Aver.
destroyerops	33.53	36.48	41.60	45.25	46.06	40.44
pink	31.49	36.23	42.03	45.61	46.09	

3.2.2. Evaluation on Noise Estimation Model

The process of noise estimation is similar to that of phone feature extraction. After model training, the output of the last hidden layer was used as the features of noise. Five noise types, 5 levels of SNR, clean speech, and silence were considered for training the model with multi labels.

Besides spectrogram, the time-domain speech segment of the current frame was also used for training. The result was improved slightly, and the SNR level achieved the best results, as shown in TABLE II.

TABLE II: The results of noise estimation

(%)	Noise Type				SNR
	Acc.	Prec.	recall	F1	Acc.
noise_SNR	-	-	-	-	54.08
noise_type	92.59	89.22	81.73	85.31	-
noise_SNR_type	92.13	87.55	81.75	84.55	67.98
noise_SNR_type w/ raw wave	92.18	88.09	81.30	84.56	68.39

3.2.3. Evaluation of Phone Judgement Model

In the acoustic phone judgment model, the same architecture as the phone recognition model described above was used. The speech signals without noise were used for training, and the average phone recognition accuracies are shown in Table III.

TABLE III: Comparison of accuracies of phone recognition

model	Aver. Acc. (%)
phone_w/o_front-end_FC	36.65
phone_w/_front-end_FC	40.44
phone_w/_front-end_FC_clean	48.30

3.2.4. Evaluation on Mask Generation Model

In this experiment, the test data with 0, 5, and 10dB were evaluated. The model was trained 50 epochs and the mask of the current frame was multiplied with the noisy speech element-wise to obtain the enhanced results.

In TABLE IV, we observed that only MSE loss was used for training. In the average of PESQ or STOI, the mask generated by the model with all three features achieved the best speech quality. "spkr", "phn" and "noise" in the table means speaker feature, phone feature, and noise feature, respectively. (a) and (b) mean original ASR score for corrupted and clean speech. And (c) is the result without adding any extra features, (d)~(g) is to add specified feature and spectrum as input. According to the results of PESQ and STOI, (c) and (g) were increased by 8.43% and 2.43%, respectively. It can be seen in (e) that the features of the phone have a relatively large contribution to speech quality, and in (d) speaker feature is more helpful for speech recognition. In order to maintain the audio quality and WER, we chose (g) that obtained the best audio quality and good WER for the follow-up experiments.

Then we experimented with the loss mentioned above, and the results are shown in TABLE V. We observed that the result of adding the phone-based judgment L_2 can reduce the WER, indicating that the guidance worked. However, phone confidence L_3 decreased the WER, and speech quality was also decreased slightly. Therefore, we further cooperated with PESQ and STOI judgment to improve the quality of the enhanced speech. In the results of using J_4, we obtained the lowest WER of 21.59% with unsatisfactory PESQ and STOI score, which also proves that recognition rate is difficult to be compatible with speech quality.

TABLE IV: The speech quality and WER for different models

metric / model	PESQ	STOI	WER(%)
(a) noisy	2.32	0.9040	29.08
(b) clean	4.50	1.0000	11.27
(c) baseline	2.49	0.8961	33.72
(d) (c)+spkr	2.57	0.9140	**25.93**
(e) (c)+phn	2.63	0.9070	28.44
(f) (c)+noise	2.57	0.9088	28.50
(g) (c)+spkr+phn+noise	**2.70**	**0.9179**	**26.36**

TABLE V: The results with different loss functions

metric / loss	PESQ	STOI	WER(%)
(g) MSE	2.70	0.9179	26.36
J_1	2.66	0.9146	26.22
J_2	2.62	0.9154	24.16
J_3	2.59	0.9057	35.59
J_4	2.67	0.9153	**21.59**
J_5 $(\lambda_1 = 0.8)$	2.72	0.9128	28.84
J_6 $(\lambda_1 = 0.8)$	2.59	0.8957	34.56
J_7 $(\lambda_1 = 0.6, \lambda_2 = 0.2)$	**2.78**	**0.9152**	24.70

IV. Conclusions

In this study, we consider speech quality and acoustic confidence to enhance speech quality and reduce the WER of noisy speech recognition. In the training process, we use the phone confidence obtained from the ASR system, the pre-trained phone-based judgment model, and the loss function considering MSE, STOI, and PESQ to guide the generation of the mask model. In addition, we consider the characteristics of speaker, phone, and noise to distinguish clear speech components and observe that the speaker and the phone have different contributions to improving the speech quality and recognition accuracy, and subsequent experiments also proved that the voice quality may have limited effect on ASR recognition. Finally, compared with the baseline model, this method can use different losses for training according to the purpose. As the front-end of ASR, WER can be reduced from 33.72% to 26.36% compared with the baseline, and after adjusting the loss, it can be reduced to 21.59%, which is a total reduction of 7.49% compared to the original ASR.

Acknowledgment

This work was funded in part by Qualcomm through a Taiwan University Research Collaboration Project.

REFERENCES

[1] L. Huang, G. Cheng, P. Zhang, Y. Yang, S. Xu, and J. Sun, "Utterance-level Permutation Invariant Training with Latency-controlled BLSTM for Single-channel Multi-talker Speech Separation," in 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2019, pp. 1256-1261.

[2] G. Kim, H. Lee, B.-K. Kim, S.-H. Oh, and S.-Y. Lee, "Unpaired speech enhancement by acoustic and adversarial supervision for speech recognition," IEEE Signal Processing Letters, vol. 26, no. 1, pp. 159-163, 2018.

[3] Y. Liu, M. Delfarah, and D. Wang, "Deep Casa for Talker-independent Monaural Speech Separation," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6354-6358.

[4] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," IEEE/ACM transactions on audio, speech, and language processing, vol. 27, no. 8, pp. 1256-1266, 2019.

[5] K. Wang, F. Soong, and L. Xie, "A pitch-aware approach to single-channel speech separation," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 296-300: IEEE.

[6] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 10, pp. 1702-1726, 2018.

[7] 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 Interspeech2019, pp. 3173-3177.

[8] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, no. 11, pp. 1641-1648, 1989.

[9] S. Shon, H. Tang, and J. Glass, "Frame-level speaker embeddings for text-independent speaker recognition and analysis of end-to-end model," in 2018 IEEE Spoken Language Technology Workshop (SLT), 2018, pp. 1007-1013.

[10] C.-F. Liao, Y. Tsao, H.-Y. Lee, and H.-M. Wang, "Noise adaptive speech enhancement using domain adversarial training," arXiv preprint arXiv:1807.07501, 2018.

[11] R. Yao, Z. Zeng, and P. Zhu, "A priori SNR estimation and noise estimation for speech enhancement," EURASIP Journal on Kinoshita, K., Ochiai, T., Delcroix, M., & Nakatani, T. (2020, May). Improving noise robust automatic speech recognition with single-channel time-domain enhancement network. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7009-7013). IEEE. Advances in Signal Processing, vol. 2016, no. 1, p. 101, 2016.

[12] S. E. Chazan, S. Gannot, and J. Goldberger, "A phoneme-based pre-training approach for deep neural network with application to speech enhancement," in 2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC), 2016, pp. 1-5.

[13] P. Karjol and P. K. Ghosh, "Broad Phoneme Class Specific Deep Neural Network Based Speech Enhancement," in 2018 International Conference on Signal Processing and Communications (SPCOM), 2018, pp. 372-376.

[14] M. Ge, L. Wang, N. Li, H. Shi, J. Dang, and X. Li, "Environment-Dependent Attention-Driven Recurrent Convolutional Neural Network for Robust Speech Enhancement," in Interspech 2019, pp. 3153-3157.

[15] Y. Xu, J. Du, Z. Huang, L.-R. Dai, and C.-H. Lee, "Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement," in Proc. Interspeech, 2015, pp. 1508-1512.

[16] Y. Xu, J. Du, L.-R. Dai, C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 1, pp. 7-19, 2014.

[17] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, 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.

[18] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," arXiv preprint arXiv:1609.07132v1, 2016.

[19] Q.-B. Hong, C.-H. Wu, H.-M. Wang, and C.-L. Huang, "Statistics Pooling Time Delay Neural Network Based on X-Vector for Speaker Verification," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6849-6853.

[20] S.-W. Fu, Y. Tsao, and X. Lu, "SNR-Aware Convolutional Neural Network Modeling for Speech Enhancement," in Interspeech2016, pp. 3768-3772.

[21] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," NASA STI/Recon technical report, vol. 93, p. 27403, 1993.

[22] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech communication, vol. 12, no. 3, pp. 247-251, 1993.

[23] Carnegie Mellon University, The CMU Pronouncing Dictionary, http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[24] J.-C. Wang, Y.-S. Lee, C.-H. Lin, S.-F. Wang, C.-H. Shih, & C.-H. Wu, "Compressive Sensing-Based Speech Enhancement," IEEE/ACM Trans. Audio, Speech, and Language Processing, Vol. 24, No. 11, November 2016, pp. 2122-2131.

[25] S.-W. Fu, C.-F. Liao, Y. Tsao, & S.-D. Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement." International Conference on Machine Learning. PMLR, 2019.

[26] S.-W. Fu, C. Yu, T. A. Hsieh, Plantinga, P., M. Ravanelli, X. Lu, & Y. Tsao, "MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement." arXiv preprint arXiv:2104.03538, 2021.