INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

A Cross-entropy-guided (CEG) Measure for Speech Enhancement Front-end Assessing Performances of Back-end Automatic Speech Recognition

Li Chai[1], *Jun Du*[2], *and Chin-Hui Lee*[3]

[1]School of Data Science, University of Science and technology of China, Hefei, Anhui, P. R. China
[2]University of Science and technology of China, Hefei, Anhui, P. R. China
[3]Georgia Institute of Technology, Atlanta, GA. USA

cl122@mail.ustc.edu.cn, jundu@ustc.edu.cn, chl@ece.gatech.edu

Abstract

One challenging problem of robust automatic speech recognition (ASR) is how to measure the goodness of a speech enhancement algorithm without calculating word error rate (WER) due to the high costs of manual transcriptions, language modeling and decoding process. In this study, a novel cross-entropy-guided (CEG) measure is proposed for assessing if enhanced speech predicted by a speech enhancement algorithm would produce a good performance for robust ASR. CEG consists of three consecutive steps, namely the low-level representations via the feature extraction, high-level representations via the nonlinear mapping with the acoustic model, and the final CEG calculation between the high-level representations of clean and enhanced speech. Specifically, state posterior probabilities from the output of the neural network for the acoustic model are adopted as the high-level representations and a cross-entropy criterion is used to calculate CEG. Experimental results show that CEG could consistently yield the highest correlations with WER and achieve the most accurate assessment of the ASR performance when compared to distortion measures based on human auditory perception and an acoustic confidence measure. Potentially, CEG could be adopted to guide the parameter optimization of deep learning based speech enhancement algorithms to further improve the ASR performance.

Index Terms: acoustic model, state posterior probabilities, cross entropy, speech enhancement, robust speech recognition

1. Introduction

Noise robustness is one of the critical issues to make automatic speech recognition (ASR) system widely used in real world today [1]. Various approaches have been proposed to tackle the problem to make ASR systems robust against environmental distortions. One approach is to use speech enhancement as a pre-processor to ASR, whose objective is to convert an observed speech signal to a set of input features of the ASR system that are insensitive to environmental distortion while simultaneously containing a sufficient amount of discriminant information [2].

Quality evaluation of the resulting enhanced signals is a very complex problem that depends on the application field. In many cases, the main objective of speech enhancement is preserving some characteristics that are required for the task concerned [3]. For example, a good listening quality and intelligibility in terms of human perception is a highest priority for speech communication systems while for ASR systems human auditory perception is not as important as preserving some acoustic cues that are used by the ASR systems to perform the recognition. Quality evaluation for speech enhancement is a very important step in the development of advanced algorithms. When speech enhancement is used as a pre-processing stage for ASR, the quality measure of speech enhancement applied for assessment should be related to the word error rate (WER) of the ASR system.

So far, few works have been presented with specific proposals for quality evaluation of the resulting enhanced signals. Particularly, in the context of ASR, one direct way to evaluate the performance of speech enhancement algorithms is to calculate the WER. Nevertheless, it requires a large amount of computation and manual transcription costs. Therefore, more easily computed instrumental measures are desired for assessing performance of speech enhancement algorithms for robust ASR without using reference transcriptions. They are beneficial to the research and development of speech enhancement algorithms for robust ASR.

In the past few decades, several methods have been proposed to assess ASR performance without using reference transcriptions. The idea has emerged, of using a distortion measure based on human auditory perception that was originally developed for the objective evaluation of perceptual performance and represents the difference between degraded speech and its original clean version, to assess ASR performance of speech enhancement algorithms for robust ASR. In [4–7], a very good correlation between WER and the perceptual evaluation of speech quality (PESQ) [8] for measuring speech quality has been verified. Recently, it has been shown that the correlation coefficient between WER and the short-time objective intelligibility (STOI) [9] for measuring speech intelligibility is higher than other distortion measures [10, 11]. Besides, in [12, 13], an acoustic confidence measure usually defined as the entropy of the posterior distribution from the output of the artificial neural network (ANN) for the acoustic model was proposed and shown a high degree of correlation with WER, where the discriminatory power of the ANN decreases and the posterior probabilities tend to become more uniform with a higher entropy.

In this paper, we propose a cross-entropy-guided (CEG) measure defined as the cross entropy of the state posterior probabilities between the enhanced speech and clean speech from the ANN output for the ANN-HMM (hidden Markov model) based acoustic model. Experiments demonstrate a consistently highest degree of correlation between the CEG and WER compared with the acoustic confidence measure, STOI and PESQ from different aspects including acoustic models, language models, speech enhancement algorithms, signal-to-noise-ratio (SNR) levels and noise types. Furthermore, CEG achieves the most accurate assessment of recognition performance of speech enhancement algorithms for noise-

http://dx.doi.org/10.21437/Interspeech.2019-2511

robust ASR. Accordingly, it provides a more accuracy guide at the time of choosing a suitable speech enhancement algorithm as a mean to introduce robustness into the recognizer.

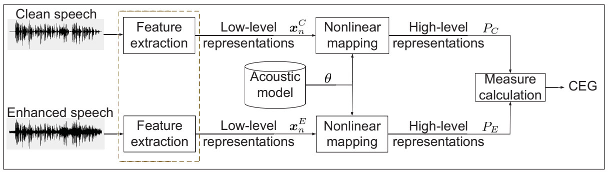


Figure 1: *The overall framework of CEG.*

2. CEG

The overall framework of CEG is illustrated in Fig. 1. Generally, it is a measure function of high-level representations of the clean speech and enhanced speech calculated by the nonlinear operations between the acoustic model of the ASR system and the raw time signals or input features of the ASR system. The CEG calculation process mainly includes three steps. The first step is the extraction of low-level representations which could be from the raw time signals or input features of the ASR system such as mel-frequency cepstral coefficients (MFCCs), log-mel-filterbank (FBANK) features and feature-space maximum likelihood linear regression (fMLLR) [14]. Due to the lack of ASR acoustic information in low-level representations, in the second step, the acoustic model of the ASR system is adopted to map the low-level representations to high-level representations which could provide useful acoustic knowledge for better assessment of ASR performance. Specifically, state posterior probabilities from the ANN output for the ANN-HMM based acoustic model are adopted as the high-level representations. Therefore, CEG currently aims to work with the ANN-HMM based ASR system which is also one of the main streams. In future, we will explore other high-level representations learned from low-level representations for generic acoustic models including both the mainstream ANN-HMM based acoustic model and the acoustic models with end-to-end optimization [15]. The last step is the calculation of CEG which measures the difference between the high-level representations of the clean speech and enhanced speech via a criterion, e.g. cross entropy, Kullback-Leibler divergence and minimum mean squared error (MMSE). Motivated by the training criterion of ANN-HMM based acoustic models, cross entropy is adopted in this study. Accordingly, CEG is defined as follows:

$$m = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{I} P_C(s_i|\boldsymbol{x}_n^C,\theta)\log P_E(s_i|\boldsymbol{x}_n^E,\theta) \qquad (1)$$

where P_C and P_E are the state posterior probabilities between the clean speech and enhanced speech from the ANN output for the ANN-HMM based acoustic model, respectively, I is the number of output classes or phonemes, N is the number of frames, $\boldsymbol{x}_n^C$ and $\boldsymbol{x}_n^E$ are the raw time signal or input feature vector of the clean speech and enhanced speech for the n-th frame, respectively and θ is a set of parameters of the ANN-HMM based acoustic model of ASR system.

In many cases, speech enhancement algorithms for noisy ASR is tuned according to the distortion measures based on human auditory perception such as PESQ and STOI. However, some researches have shown that speech enhancement algorithms which achieve a better distortion value may result in worse ASR performance especially for multi-condition training [16–18] because they inevitably introduce distortions that could not be captured by the distortion measures. Distortion measures were originally developed to assess the perceptual performance and not directly correlated with ASR performance. In contrast, CEG considers the ASR acoustic information by the introduction of high-level representations, which is directly correlated with ASR system. Furthermore, the difference of the acoustic characteristics between the clean speech and enhanced speech is accurately measured by the cross entropy. CEG can capture the changes of the ASR performance caused by different acoustic models while distortion measures are invariant to the back-end. Distortion measures which are usually calculated by waveforms of the enhanced speech and clean speech are not easily applicable to the speech enhancement algorithms which directly output the speech features for ASR [19, 20], while CEG can be used to evaluate the ASR performance on both the signal level and feature level. Although CEG additionally requires the acoustic model of the ASR system compared to the distortion measures, the acoustic model is available because the speech enhancement algorithms is usually designed for increasing the robustness of a existing recognizer.

In comparison to the acoustic confidence measure using the entropy, although CEG requires parallel speech pairs, it achieves a more accurate assessment of the ASR performance. Moreover the original clean speech is available since an assumption is made in the process of research and development of speech enhancement algorithms that the noisy speech is generated by recording the noise in different noisy environments and artificially adding it to the clean speech. This assumption is reasonable from the viewpoint of reducing the recording cost.

It is worth mentioning that CEG is differentiable and thus can replace the MMSE as the optimization criterion for ANN-based speech enhancement algorithms aiming at better ASR performance. Although many advanced objective functions have been investigated recently [21, 22], they are not correlated well with the WER and thus can not guarantee better ASR performance. We will disclose more details in our future work due to the space limitation here.

3. Evaluation procedure

3.1. Correlation coefficients

There are three kinds of commonly used correlation coefficients , namely the Pearson correlation coefficient, the Spearman rank correlation coefficient and the Kendall Tau rank correlation coefficient, while the Pearson correlation coefficient is the most common one [23]. Hence Pearson correlation coefficient is adopted in our study, which is a measure of the linear correlation between two data sets and can be calculated as follows.

$$\rho_{xy} = \frac{\sum_{n=1}^{N}(x_n-\overline{x})(y_n-\overline{y})}{\sqrt{\sum_{n=1}^{N}(x_n-\overline{x})^2}\sqrt{\sum_{n=1}^{N}(y_n-\overline{y})^2}} \qquad (2)$$

This equation can be considered as an expression of a ratio of how much the two data sets $\boldsymbol{x} = [x_1x_2...x_N]^\top$ and $\boldsymbol{y} = [y_1y_2...y_N]^\top$ vary together compared to how much they vary separately [23]. The magnitude of the correlation coefficient indicates the strength of the correlation and the sign indicates if the correlation is positive or negative.

3.2. Mapping

We are interested in measuring the monotonic relation between the CEG and WER. Accordingly, first a mapping is used in order to account for a nonlinear relation between the CEG and WER. The main reason for this mapping procedure is to linearize the data such that we can use the Pearson correlation coefficient. Motivated by [6,9,11,24], a logistic function is used here:

$$f(m) = \frac{100}{1 + \exp(am + b)} \tag{3}$$

where a and b are constants to be determined by a data-fitting using the least-squares method, m represents the CEG score and $f(m)$ could be considered as an estimator of the WER which is between 0 and 100. Please note that a logistic function is also monotonic and will therefore not influence the monotonicity between the CEG and WER. Then the performance of CEG is evaluated by means of the Pearson correlation coefficient (ρ), which is applied on the mapped objective scores, i.e., $f(m)$. Please note that the same evaluation procedure is used as with CEG for the acoustic confidence measure and distortion measures. Since we are only interested in the strength of the correlation, the results of the magnitude of the correlation coefficient denoted as ρ is shown in the following experiments.

4. Experiments

4.1. Experimental setting

Experiments were conducted on the 1-channel CHiME-4 task [25]. To make the calculation of the distortion measures (e.g. PESQ and STOI) feasible, we evaluated the correlations of WER with CEG, the acoustic confidence measure and the distortion measures by simulated data that were generated by artificially mixing background noises including cafe (CAF), street junction (STR), public transport (BUS), and pedestrian area (PED) with clean speech data from the development and test data consisting of 410 and 330 utterances respectively. The correlations were investigated for five situations, namely different speech enhancement algorithms as a pre-processing stage of ASR, different acoustic models, different language models, different noise types, and different SNRs. Experiments of the top four situations were conducted on the simulated data from the official development and test sets consisting of 1640 and 1320 utterances respectively which had been generated by mixing the abovementioned clean speech data with background noises. Experiments of the last situation were conducted on the simulated data constructed by mixing the abovementioned clean speech data with background noises from all six channels at six levels of SNRs (-5dB, 0dB, 5dB, 10dB, 15dB and 20dB) to form 75480 utterances, respectively. Please note that WER, STOI, PESQ and the acoustic confidence measure were all computed per utterance. All the correlation coefficients in this paper were computed by the procedure in Sec. 3.2. In this study, the adopted acoustic models in multi-condition training mode are from a deep neural network (DNN)-based acoustic model trained on fMLLR transformed features by cross entropy minimization, a DNNsMBR-based acoustic model trained on fMLLR features by cross entropy minimization followed by state-level minimum Bayes risk (sMBR) optimization, and a deep convolutional neural network (DCNN)-based [26] acoustic model trained on the FBANK features by cross entropy minimization, where the first two acoustic models are official baselines provided by Kaldi [14]. All of them were trained on the same training set, namely, the real and simulated noisy training set from channel 5 consisting of 1600 and 7138 utterances respectively. Besides, the adopted acoustic model in clean-condition training mode is DNNsMBR-based model trained on the 7138 utterances of the clean WSJ0 training set [27]. 3-gram, 5-gram and recurrent neural networks (RNN) based official language models provided by Kaldi were adopted.

4.2. Correlation comparison in different situations

Fig. 2 shows a consistently highest degree of correlation between CEG and WER compared with the acoustic confidence measure denoted as Entropy and distortion measures (PESQ and STOI) in acoustic models using different input features, ANN structures and optimization criterions and in language models including 3-gram, 5-gram and RNN for multi-condition training. Moreover, the correlation is robust to both acoustic models and language models. The DNNsMBR-3gram based recognition system was adopted in the following experiments.

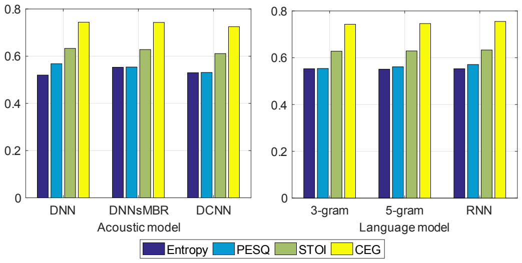
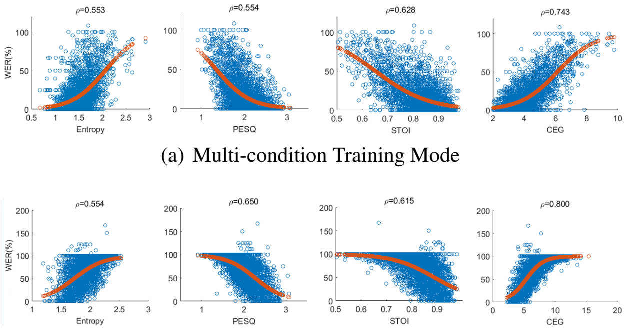


Figure 2: *Pearson correlation coefficients in multi-condition training mode.*

Fig. 3 shows the scatter plots of the relation between WER and the Entropy, PESQ, STOI and CEG, respectively, where the red lines represent the applied mapping functions which clearly show good performance by means of a strong monotonic relation with WER. Obviously, CEG has the highest degree of correlation with WER for both clean-condition and multi-condition training. Please note that the correlation score between the CEG and WER for multi-condition training tends to be smaller than that for clean-condition training because the number of the same scores of WER corresponding to different scores of CEG for multi-condition training is larger, which leads to worse correlation statistics. Unlike the conclusions in [10, 11], we observe that the correlation scores between



(a) Multi-condition Training Mode

(b) Clean-condition Training Mode

Figure 3: *Scatter plots between the WER and instrumental measures.*

the STOI and WER tend to be larger than those between the PESQ and WER for multi-condition training and the contrary conclusion could be drawn for clean-condition training, which may indicate that the recognition performance depends more largely on speech quality for clean-condition training and on speech intelligibility for multi-condition training. Besides, it is noted that the acoustic confidence measure and the proposed CEG have positive correlations with WER, on the contrary, PESQ and STOI have negative correlations with WER.

Recognition performance varies with different speech enhancement algorithms. Therefore, we investigated the relation with WER by two representative noise reduction algorithms in addition to the reference case of unprocessed noisy speech, namely, an optimally-modified log-spectral amplitude (OM-LSA) speech estimator [28] and a masking-based DNN enhancement algorithm [29]. Table 1 shows the consistently strongest correlations between CEG and WER compared with the Entropy, PESQ and STOI in both multi-condition training and clean-condition training modes for different enhancement algorithms.

Table 1: *Pearson correlation coefficients for different speech enhancement algorithms.*

Training mode	Algorithms	Entropy	PESQ	STOI	CEG
Multi-condition	Noisy	0.553	0.554	0.628	0.743
	OM-LSA	0.589	0.582	0.624	0.726
	DNN	0.676	0.600	0.639	0.744
Clean-condition	Noisy	0.554	0.650	0.615	0.800
	OM-LSA	0.689	0.691	0.645	0.794
	DNN	0.670	0.664	0.625	0.778

The relationship is also investigated for different noise conditions and different SNR levels, where CEG still consistently shows strongest monotonic relationship to the WER compared with Entropy, PESQ and STOI shown in Fig. 4. The correlation becomes weak in very high SNR levels where most values of WERs are 0% or in very low SNR levels where most values of WERs are 100% due to the negative influence of these points which have the same WERs but different values of the evaluation measures.

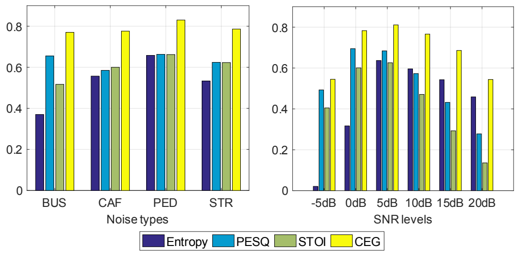


Figure 4: *Pearson correlation coefficients in clean-condition training mode for different noise types and SNR levels.*

4.3. Comparison of evaluation accuracy

From Table 2, we can see that the multi-condition training mode could bring better recognition performance compared with the clean-condition training mode. However, the improvements could not be evaluated by distortion measures due to their invariance to the backend. In contrast, the acoustic confidence measure and CEG both related to the backend could evaluate it, where the smaller values of them are, the lower WER is. Besides, there are many cases where the recognition performance could not be accurately evaluated by both the distortion measures and acoustic confidence measure, e.g., the smaller Entropy of OM-LSA does not bring a decline in WER compared with that of DNN in both multi-condition training and clean-condition training modes, better PESQ or STOI of enhanced speech does not bring improvements of recognition performance in multi-condition training mode, the same Entropy of DNN as that of the unprocessed noisy speech does not bring the same WER, and worse STOI of OM-LSA brings a decline in WER for clean-condition training. In contrast, CEG could accurately evaluate the recognition performance of different speech enhancement algorithms for both multi-condition training and clean-condition training.

Table 2: *The average evaluation measure scores and WERs of different speech enhancement algorithms for ASR on the official simulated development and test sets (2960 utterances).*

Training mode		Noisy	OM-LSA	DNN
Multi-condition	Entropy	1.52	1.65	1.71
	PESQ	2.00	2.25	2.30
	STOI	0.819	0.808	0.846
	CEG	4.23	4.45	4.39
	WER(%)	19.39	25.70	24.46
Clean-condition	Entropy	1.92	1.89	1.92
	PESQ	2.00	2.25	2.30
	STOI	0.819	0.808	0.846
	CEG	6.81	5.99	5.68
	WER(%)	65.52	57.4	51.59

Many existing speech enhancement algorithms can improve speech quality but not speech intelligibility [30, 31], such as OM-LSA in Table 2. However, these speech enhancement algorithms may improve recognition accuracy especially for clean-condition training. Their recognition performance cannot be accurately evaluated by STOI which measures the speech intelligibility. For example, resulting speech enhanced by OM-LSA could improve recognition accuracy for clean-condition training regardless of its worse STOI shown in Table 2. Accordingly, the conclusion in [10, 11] that the correlation coefficient between the WER and STOI is higher than other distortion measures (e.g., PESQ) is not accurate and reliable enough. Some researches [22, 32] suggested by the conclusion in [10, 11] designed a speech enhancement front-end to especially improve STOI and thus achieve better ASR performance. Our findings reveal the unreasonableness of this design of speech enhancement front-end as a pre-processor to ASR.

5. Conclusion

In this study, we propose a measure, i.e., CEG to evaluate the performance of the speech enhancement algorithms for noise-robust ASR without using reference transcriptions, language models and recognition process. Compared with the acoustic confidence measure, PESQ and STOI, CEG shows the highest correlation with WER and achieves the most accurate evaluation of recognition performance. Moreover, CEG could be directly adopted as the optimization criterion of the ANN-based speech enhancement algorithms for improving ASR performance instead of the conventional MMSE criterion, which will be explored in our another work.

6. 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] T. Yoshioka and M. J. Gales, "Environmentally robust ASR front-end for deep neural network acoustic models," *Computer Speech & Language*, vol. 31, no. 1, pp. 65–86, 2015.

[3] L. Di Persia, M. Yanagida, H. L. Rufiner, and D. Milone, "Objective quality evaluation in blind source separation for speech recognition in a real room," *Signal Processing*, vol. 87, no. 8, pp. 1951–1965, 2007.

[4] H. Sun, L. Shue, and J. Chen, "Investigations into the relationship between measurable speech quality and speech recognition rate for telephony speech," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1. IEEE, 2004, pp. I–865.

[5] T. Yamada and N. Kitawaki, "A PESQ-based performance prediction method for noisy speech recognition," in *International Congress on Acoustics, ICA2004, Proc. pp. Tu. P*, vol. 2, 2004.

[6] T. Yamada, M. Kumakura, and N. Kitawaki, "Performance estimation of speech recognition system under noise conditions using objective quality measures and artificial voice," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, 2006.

[7] L. Di Persia, D. Milone, H. L. Rufiner, and M. Yanagida, "Perceptual evaluation of blind source separation for robust speech recognition," *Signal Processing*, vol. 88, no. 10, pp. 2578–2583, 2008.

[8] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 749–752.

[9] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[10] A. H. Moore, P. P. Parada, and P. A. Naylor, "Speech enhancement for robust automatic speech recognition: Evaluation using a baseline system and instrumental measures," *Computer Speech & Language*, vol. 46, pp. 574–584, 2017.

[11] D. Thomsen and C. E. Andersen, "Speech enhancement and noise-robust automatic speech recognition," *Aalborg University*, 2015.

[12] H. Misra, H. Bourlard, and V. Tyagi, "New entropy based combination rules in HMM/ANN multi-stream ASR," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 2. IEEE, 2003, pp. II–741.

[13] J. Barker, G. Williams, and S. Renals, "Acoustic confidence measures for segmenting broadcast news," in *Fifth International Conference on Spoken Language Processing*, 1998.

[14] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[15] D. Yu and J. Li, "Recent progresses in deep learning based acoustic models," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 3, pp. 396–409, 2017.

[16] J. Du, Q. Wang, T. Gao, Y. Xu, L.-R. Dai, and C.-H. Lee, "Robust speech recognition with speech enhanced deep neural networks," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[17] S.-J. Chen, A. S. Subramanian, H. Xu, and S. Watanabe, "Building state-of-the-art distant speech recognition using the CHiME-4 challenge with a setup of speech enhancement baseline," *arXiv preprint arXiv:1803.10109*, 2018.

[18] A. S. Subramanian, S.-J. Chen, and S. Watanabe, "Student-teacher learning for BLSTM mask-based speech enhancement," *arXiv preprint arXiv:1803.10013*, 2018.

[19] K. Han, Y. He, D. Bagchi, E. Fosler-Lussier, and D. Wang, "Deep neural network based spectral feature mapping for robust speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[20] D. Bagchi, M. I. Mandel, Z. Wang, Y. He, A. Plummer, and E. Fosler-Lussier, "Combining spectral feature mapping and multi-channel model-based source separation for noise-robust automatic speech recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 496–503.

[21] L. Chai, J. Du, and C.-H. Lee, "Error modeling via asymmetric laplace distribution for deep neural network based single-channel speech enhancement," *Proc. Interspeech 2018*, pp. 3269–3273, 2018.

[22] 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 (TASLP)*, vol. 26, no. 9, pp. 1570–1584, 2018.

[23] P. Y. Chen and P. M. Popovich, *Correlation: Parametric and nonparametric measures*. Sage, 2002, no. 137-139.

[24] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *IEEE signal processing magazine*, vol. 32, no. 2, pp. 114–124, 2015.

[25] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.

[26] D. Yu, W. Xiong, J. Droppo, A. Stolcke, G. Ye, J. Li, and G. Zweig, "Deep convolutional neural networks with layer-wise context expansion and attention." in *Interspeech*, 2016, pp. 17–21.

[27] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "Csr-i (wsj0) complete," *Linguistic Data Consortium, Philadelphia*, 2007.

[28] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on speech and audio processing*, vol. 11, no. 5, pp. 466–475, 2003.

[29] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7092–7096.

[30] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 1, pp. 47–56, 2011.

[31] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2013.

[32] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 006–012.