Interspeech 2024
1-5 September 2024, Kos, Greece

Are Recent Deep Learning-Based Speech Enhancement Methods Ready to Confront *Real-World* Noisy Environments?

Candy Olivia Mawalim, Shogo Okada, Masashi Unoki

Japan Advanced Institute of Science and Technology, Japan
candylim@jaist.ac.jp, okada-s@jaist.ac.jp, unoki@jaist.ac.jp

Abstract

Recent advancements in speech enhancement techniques have ignited interest in improving speech quality and intelligibility. However, the effectiveness of recently proposed methods is unclear. In this paper, a comprehensive analysis of modern deep learning-based speech enhancement approaches is presented. Through evaluations using the Deep Suppression Noise and Clarity Enhancement Challenge datasets, we assess the performances of three methods: Denoiser, DeepFilterNet3, and FullSubNet+. Our findings reveal nuanced performance differences among these methods, with varying efficacy across datasets. While objective metrics offer valuable insights, they struggle to represent complex scenarios with multiple noise sources. Leveraging ASR-based methods for these scenarios shows promise but may induce critical hallucination effects. Our study emphasizes the need for ongoing research to refine techniques for diverse real-world environments.

Index Terms: speech enhancement, Whisper, objective metrics, deep noise suppression, Clarity challenge

1. Introduction

The development of machine learning approaches and the availability of large-scale datasets enable more sophisticated modelling of speech and noise characteristics, leading to improved noise reduction and speech intelligibility. As a result, contemporary speech enhancement systems exhibit unprecedented levels of performance and adaptability to diverse application domains, such as telecommunications, voice assistants, hearing aids, and audio-visual communication [1].

Recent studies on objective evaluation are pivotal in advancing speech enhancement by quantitatively measuring improvements in speech quality [2], noise reduction [3], and intelligibility [4, 5], facilitating the development of more effective speech enhancement techniques. Three categories of objective evaluation methods: conventional feature-based methods, data-driven methods, and neurophysiological methods, can be used to compare different enhancement performances to a certain degree [6]. However, research has shown that these metrics are insufficient and sometimes inconsistent, especially when handling noisy environments [7].

In recent years, automatic speech recognition (ASR)-based methods have demonstrated superior accuracy in speech intelligibility prediction compared to other approaches, particularly in noisy conditions [5, 6, 8]. Numerous studies employing various deep learning techniques and datasets have consistently highlighted the enhanced performance of ASR-based methods [5, 9]. Notably, approaches leveraging state-of-the-art Whisper models [10] trained on extensive multilingual and multitask datasets, totalling 680k hours of training, have exhibited promising accuracy and efficacy in speech intelligibility prediction, even in challenging acoustic environments [11, 12]. Recent research on listening tests with English speakers evaluating non-native talkers also found that Whisper can accurately mirror human listener intelligibility prediction [8].

However, the inherent black-box nature of deep learning-based methodologies limits our understanding to mere accuracy improvements. Recognizing the limitations of ASR-based techniques is crucial, as is devising strategies to optimize their utility. Karbasi noted ASR's challenges in comprehending speech in reverberant environments, especially amidst background noise from concurrent conversations [5]. Additionally, Koenecke et al. reported harms caused by hallucination phenomena associated with the Whisper model [13]. These challenges must be addressed to harness the full potential of ASR-based methodologies.

This paper provides a comprehensive assessment of recent speech enhancement methods, focusing on their advantages and constraints in mismatches across tasks. To accomplish this goal, we introduce Beep-PER, which assesses phoneme-level errors between predicted transcripts of both original and enhanced speech using an ASR system. In our experiments, we utilize established datasets for speech enhancement tasks and employ cutting-edge speech enhancement methods. Furthermore, we discuss the potential and limitations of ASR-based objective evaluation methods, informed by our experimental results.

Our paper contributes to the understanding of the following research questions: (1) How do recent deep learning-based speech enhancement methods perform in cross-challenge evaluations? (2) What are the strengths and limitations of objective metrics in assessing speech enhancement methods in challenging environments? (3) How applicable are these techniques in real-world scenarios, specifically domains such as hearing aids? This comprehensive exploration aims to advance our knowledge in the field and guide future research endeavours.

2. ASR-based Objective Evaluation

The word error rate (WER) is commonly employed to assess automatic speech recognition (ASR) systems, yet it is not directly utilized for evaluating speech enhancement methods. The WER quantifies the discrepancy between the recognized text and the reference transcript and is calculated as follows.

$$\mathrm{WER} = \frac{(I + D + S)}{N} \times 100 \tag{1}$$

where I denotes the insertions, D denotes the deletions, S denotes substitutions, and N denotes the total number of words in the reference speech.

The WER has several limitations. For example, it requires a reference transcript, considers only words, so the noise impact

10.21437/Interspeech.2024-129

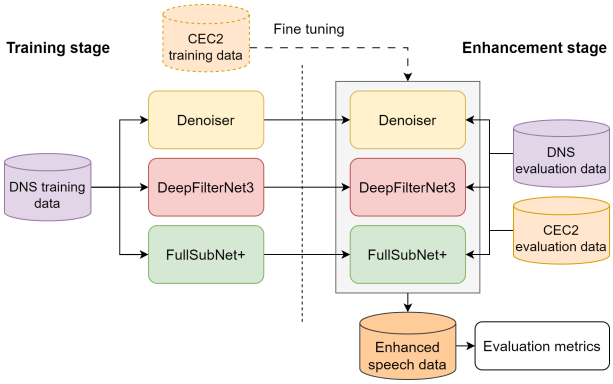

Figure 1: *Experimental pipeline for evaluating real-time deep learning-based speech enhancement. Three methods were employed: Denoiser, DeepFilterNet3, and FullSubNet+. A cross-corpus evaluation was conducted using the CEC2 and DNS datasets, both with and without fine-tuning. We analysed the enhanced speech signal outputs using the PESQ, STOI, WER, and Beep-PER metrics.*

on recognition is unclear, and cannot gauge the naturalness of enhanced speech. To address the first limitation, various methods have been proposed to estimate the WER [14] or introduce alternative metrics within the ASR framework [5]. Although WER may not be the perfect gauge for speech enhancement, recent studies have shown that ASR-based methods are more effective for predicting speech intelligibility than are traditional metrics such as short-time objective intelligibility (STOI) [15].

In this study, we focus on speech recognition in noisy environments. We specifically investigate phoneme-level recognition errors using Whisper, a cutting-edge ASR system. We assess recent speech enhancement techniques using objective metrics and introduce a new metric, the Beep dictionary Phoneme-level Error Rate (Beep-PER). The phoneme-level error rate (PER) offers clear advantages over WER for evaluating ASR systems [16]. PER is especially valuable in early development stages or under challenging conditions, and it helps pinpoint pronunciation issues with greater precision. By assessing systems at the phoneme level, we gain insights into their performance, particularly in noisy environments where word boundaries may be unclear.

The Beep dictionary[1], which primarily serves as a pronunciation resource, incorporates data from the Oxford Text Archive releases 710 and 1054, copyrighted by Oxford University Press and the Medical Research Council. We employed the Beep dictionary to parse phoneme-level tokens of recognized words after conducting some standard transformations, such as lower-casing words and removing exceeding white spaces & punctuation. We then calculated the WER based on these phoneme-level tokens. The WER can surpass 100% [17], suggesting a higher error count than the total number of words in the reference. For simplicity, we streamline all the Beep-PER results presented in subsequent sections, ensuring that the values remain within the 0 to 100% range in both tables and graphs. The code for Beep-PER and the enhancement demonstration will be publicly available[2].

[1] https://www.openslr.org/14/

[2] https://github.com/candyolivia/is2024_deep_enhancement

[3] https://github.com/microsoft/DNS-Challenge

[4] https://clarity challenge.org/docs/cec2/data/cec2_data

3. Experiment

3.1. Dataset

We utilized subsets from two speech enhancement challenges: the 3rd Deep Noise Suppression (DNS) Challenge[3][18] and the 2nd Clarity Enhancement Challenge (CEC2)[4] [19, 20]. The DNS challenge is a sequence of single-channel noise suppression challenges by Microsoft. Meanwhile, the Clarity Challenge is a sequence of machine-learning challenges for hearing aids. The CEC2 dataset was recorded using hearing aid microphones. We selected these datasets because, to our knowledge, the noise stimuli they contain best represent real-world scenarios. However, it is important to note a limitation: drawing conclusions becomes challenging in scenarios where the speech-to-noise ratio (SNR) is low, such as during the Lombard effect [21]. Nonetheless, actual noisy environments can have much lower SNRs, which may not be covered by other evaluation sets in different noise reduction datasets. Moreover, we can use these datasets to identify types of noise that are difficult to reduce and that affect speech intelligibility perception.

In addition to the original test data provided in the DNS challenge, we generated an evaluation dataset comprising 2,707 utterances by utilizing the DNS challenge dataset. This dataset was created by mixing clean speech extracted from the VCTK [22] dataset with randomly chosen noise signals with SNR values ranging from -5 to 15 dB. Consequently, this dataset contains diverse real-world test scenarios encompassing various noise types, reverberations, and even paralinguistic elements such as laughter or sighs. Our utilization of the VCTK dataset was restricted due to our reliance on the Beep dictionary, which is tailored to British accents for phoneme derivation in recognized speech. We named the evaluation datasets for the original test data (synthetic clips with reverberation) Set-1 and generated test data Set-2.

The CEC2 dataset presents a scenario where a listener occupies a room while a target speaker utters a sentence in the presence of two or three active interfering sound sources. These scenes are characterized by randomized parameters such as room size, materials, target speaker identity, specific spoken sentences, listener and target talker locations, noise interference positions, as well as details such as listener head orientation. We utilized the front signal of hearing aid channel signals from the CEC2 development set (2,500 scenes) with SNR ranging from -12.5 to 7.5 dB and its subset with SNR ranging from 0 to 7.5 dB with their corresponding anechoic clean speech as evaluation data. For simplicity, we named these evaluation datasets Set-1 and Set-2, respectively.

3.2. Speech Enhancement Methods

As shown in Fig. 1, we selected three modern speech enhancement techniques: Denoiser[5], developed using the DEMUCS architecture [23, 26]; DeepFilterNet[6] [24]; and FullSubNet+[7] [25]. These methods were selected for their outstanding performance in previous DNS challenges, their efficiency in terms of parameter count, and their ability to run in real time on a notebook CPU, making them suitable for many applications.

Denoiser, a real-time speech enhancement model, employs a unique encoder-decoder architecture with skip connections, allowing it to process audio directly from its raw waveform. Denoiser can address both stationary noise and the more challenging nonstationary noise, effectively restoring clean speech even in reverberant environments [23]. Moreover, achieves state-of-the-art speech enhancement results while maintaining

Table 1: *Objective evaluation results of three speech enhancement methods using DNS and CEC2 datasets: Denoiser (initial hidden channels H of 48 and 64) [23], DeepFilterNet3 [24], and FullSubNet+ [25]. The mean values of PESQ, STOI, WER (%), and Beep-PER (%) are displayed, with arrows indicating the superior performance direction for each metric.*

Dataset	Subset	SNR (dB)	Method	RTF (↓)	Causal	PESQ (↑)	STOI (↑)	WER (↓)	Beep-PER (↓)
DNS	Set-1	[0, 20]	Unprocessed	-	-	1.822	0.866	9.728	9.234
			Denoiser (H=48)	0.80	Yes	2.932	0.921	5.740	5.333
			Denoiser (H=64)	1.05	Yes	2.938	0.926	5.673	5.132
			DeepFilterNet3	0.19	No	3.168	**0.942**	**4.353**	**4.232**
			FullSubNet+	0.21	No	**3.218**	0.938	4.666	4.581
DNS	Set-2	[-5, 15]	Unprocessed	-	-	2.287	0.219	**20.747**	19.071
			Denoiser (H=48)	0.80	Yes	2.770	0.047	27.988	25.094
			Denoiser (H=64)	1.05	Yes	2.844	0.048	26.325	23.593
			DeepFilterNet3	0.19	No	**3.055**	0.128	21.268	**18.918**
			FullSubNet+	0.21	No	1.966	**0.640**	33.440	29.954
CEC2	Set-1	[-12.5, 7.5]	Unprocessed	-	-	1.421	0.620	**60.980**	**57.523**
			Denoiser (H=48)	0.80	Yes	1.168	0.062	79.059	74.315
			Denoiser (H=64)	1.05	Yes	1.246	0.062	78.527	73.953
			DeepFilterNet3	0.19	No	**1.438**	**0.625**	71.142	66.844
			FullSubNet+	0.21	No	1.354	0.529	76.218	71.187
CEC2	Set-2	[0, 7.5]	Unprocessed	-	-	1.782	0.771	**27.174**	**24.482**
			Denoiser (H=48)	0.80	Yes	1.655	0.062	50.298	55.082
			Denoiser (H=64)	1.05	Yes	1.722	0.060	51.647	56.017
			DeepFilterNet3	0.19	No	**1.941**	**0.790**	43.418	39.204
			FullSubNet+	0.21	No	1.636	0.651	53.772	48.277

impressive efficiency. This efficiency is attributed to its ability to leverage the knowledge of speech production and psychoacoustic perception while having a real-time performance [24]. FullSubNet+ was built upon its successor, FullSubNet [27]. Unlike FullSubNet, FullSubNet+ employs a novel "multiscale time sensitive channel attention" module that pinpoints crucial frequency regions for noise reduction and replaces computationally expensive LSTM layers with stacked temporal convolutional network [28] blocks, making it lighter and faster than FullSubNet while maintaining performance. This method was reported to surpass other state-of-the-art methods in the DNS challenge [25].

3.3. Evaluation

For our evaluation, we either retrained existing speech enhancement models with default parameters or utilized pretrained models available online, all of which were trained on the DNS training dataset. We confirmed that our results closely align with those reported in the original papers. To simulate real-world conditions, we also calculated the real-time factor (RTF) during inference on a quad-core Intel i5 CPU notebook computer using a single thread. Additionally, we incorporated information about the causality of speech enhancement methods, which is very important for the development of several real-time speech-based technologies. We follow the definition of a causal system outlined in the CEC2, which dictates that no look-ahead information from input samples beyond 5 ms is included.

Two commonly used objective metrics for evaluating noise reduction effectiveness include the perceptual evaluation of speech quality (PESQ) [29] and short-term objective intelligibility (STOI) [30]. We also incorporated wide-band PESQ and STOI metrics to analyse the speech enhancement quality, along with WER and Beep-PER. Importantly, since we lack transcripts for all reference signals, we computed WER and Beep-PER using recognized speech from pairs of clean-noisy or clean-enhanced signals employing Whisper [10] with a medium scale trained on English-only data[8]. We chose a medium-scale model[8] based on initial findings, showing it achieves about 2% higher accuracy in WER with clean data and about 5% higher accuracy with noisy data in CEC2 Set-1 compared to a large-scale model[9].

3.4. Results

Table 1 displays the comprehensive results of speech enhancement without fine-tuning. Across all methods, the quality of the enhanced speech signals significantly improves for DNS Set-1, as validated by metrics including PESQ, STOI, WER, and Beep-PER. However, despite advancements in speech enhancement techniques, maintaining robust performance is still challenging, particularly when dealing with mismatches in SNRs between training and evaluation data, and during cross-dataset evaluations. We obtained intriguing results indicating that the STOI metric displayed a different trend compared to other metrics, and employing methods that utilized STOI-related features resulted in improved STOI scores.

Through fine-tuning, speech enhancement methods can be closely adapted to specific datasets and environmental conditions, thereby enhancing quality and intelligibility. Consequently, we conducted experiments involving the fine-tuning of our model using the CEC2 training dataset. To streamline the task, we exclusively utilized signals from one channel microphone simulating the Behind-The-Ear form of a hearing aid. Our training set comprised a total of 12,000 scenes of training data[3], encompassing both left and right signals. For illustrative purposes, we plotted the results obtained using Denoiser (H=48), which was chosen for its causality and low RTF, as well as DeepFilterNet3, which performed the best on the CEC2-Set1, as depicted in Fig. 2 and Fig. 3.

The example in Fig. 2 illustrates that the noisy signal exhibits greater enhancement at higher SNRs than at lower SNRs. While the spectrogram suggests that the fine-tuned Denoiser model closely resembles clean speech in the first row, the actual voice appears corrupted and less smooth, making it more challenging to recognize than the results obtained with the DeepFilterNet3 model. Additionally, the DeepFilterNet3 model struggles to effectively reduce noise from nontarget speakers. Figure 3 displays the distribution of the enhanced noisy speech of these fine-tuned models. Although the Denoiser exhibits noise reduction in some signals at higher SNRs, on average, its per-

[5] https://github.com/facebookresearch/denoiser
[6] https://github.com/Rikorose/DeepFilterNet
[7] https://github.com/RookieJunChen/FullSubNet-plus
[8] https://huggingface.co/openai/whisper-medium.en
[9] https://huggingface.co/openai/whisper-large-v2

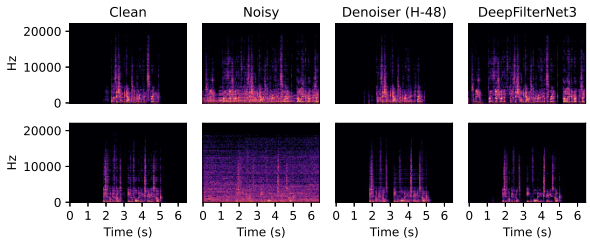

Figure 2: *Spectrograms depicting enhanced speech signals processed using the fine-tuned Denoiser (H=48) and DeepFilterNet3 models (300 epochs). The spectrograms are divided into two rows: the first row corresponds to signals with SNR < 0 dB (another speaker's voice as noise), while the second row corresponds to signals with SNR ≥ 0 dB (music as noise).*

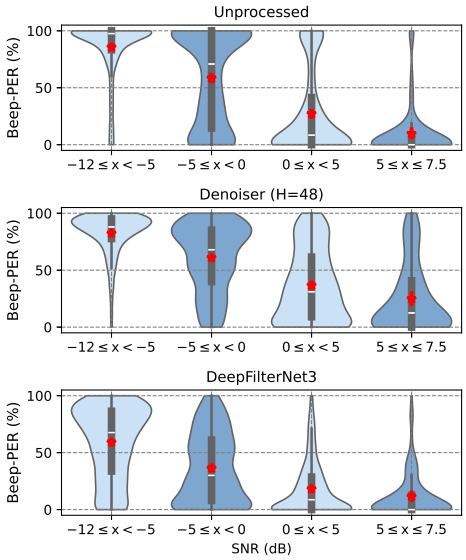

Figure 3: *Violin plots illustrate noisy speech alongside enhanced speech signals processed with the fine-tuned Denoiser (H=48) and DeepFilterNet3 models. We also performed t tests on all feasible pairs of noisy and enhanced signals in the same SNR group. Our analysis confirmed statistically significant differences ($p < 0.001$) between these methods. The red dots indicate the mean value of each group.*

formance does not surpass that of noisy signals. Conversely, the DeepFilterNet3 model performs notably well, particularly at higher SNRs, yet it still encounters difficulty in enhancing noisy signals at lower SNRs, with a mean Beep-PER of approximately 50%.

Furthermore, we analyse objective evaluation issues, particularly speech intelligibility, which we measure using the STOI and Beep-PER scores. Table 1 indicates that Beep-PER demonstrates a lower error rate compared to WER and exhibits a similar trend. Thus, to focus on further analysis, we opt not to consider the latter. Our analysis focuses on the output of the fine-tuned DeepFilterNet3 model, revealing that over 70% (1,783/2,500) of scenes exhibit STOI scores exceeding 0.5, yet less than half exhibit Beep-PER scores below 30%. Interestingly, some scenes with nearly perfect STOI scores exhibit significantly high Beep-PER values, and vice versa. In extreme cases, 53 out of 2,500 scenes achieve Beep-PER scores of zero (perfect recognition) but STOI scores of nearly zero. Despite this discrepancy, our inspection confirms that noise is reduced, as shown by PESQ scores surpassing 3.5.

We also identify recognition errors attributed to hallucinations induced by Whisper, albeit to a lesser extent than other recognition errors. Preprocessing, notably conversion to phoneme-level tokens, mitigates some issues, reducing instances of unknown words such as screaming tokens ('AHHH-HHHH') and excessive punctuation. However, critical hallucination effects persist in Whisper, notably redundancy, violent words, and expressions of gratitude. Using a parser, we automatically label these prevalent hallucination effects, identifying approximately 23 scenes (approximately 0.9%) in recognized text derived from enhanced speech.

4. Discussion

Our experiment yielded three key findings:

(1) While these methods generally performed better under similar conditions and high SNRs, their robustness across different datasets, even after fine-tuning, was notably lacking.

(2) While common objective metrics offer insights for assessing speech enhancement methods under conditions similar to that of the training data and high SNRs, they struggle to accurately evaluate methods in challenging settings characterized by low SNRs and noise from multiple speakers.

(3) The practical applicability of speech enhancement techniques in real-world scenarios is limited, particularly in domains such as hearing aids, necessitating further investigation. Despite advancements, adapting these techniques to diverse and dynamic real-life environments is still challenging, underscoring the importance of ongoing research and refinement efforts.

We recognize the limitations of our experiment, primarily stemming from the inability of the dataset to fully represent all real-world noisy environments. Furthermore, our comparison of speech enhancement methods may not encompass all available advanced techniques, and ensuring a fair review under identical conditions is challenging due to accessibility issues. Additionally, our study did not incorporate the latest advancements in objective measurements or consider potentially superior ASR systems. These constraints underscore the necessity for further research to navigate these complexities and deepen our comprehension of speech enhancement across diverse real-world scenarios and application domains.

5. Conclusion

In conclusion, our study underscores the ongoing challenges faced by recent speech enhancement methods, particularly in noisy environments with low SNRs and additional noise from other human voices. While fine-tuning approaches show potential for improving enhancement results, their efficacy remains contingent upon the dataset used for training, limiting their applicability in real-world scenarios with unpredictable noise. Additionally, existing metrics cannot effectively assess speech enhancement in noisy environments, and ASR-based methods are hindered by issues such as text hallucination. Therefore, extensive testing across diverse datasets and real-world scenarios are required to improve the effectiveness of recent speech enhancement methods.

6. Acknowledgements

This work was supported by Japan Society for the Promotion of Science (JSPS) KAKENHI (22H04860, 22H00536, and 23K18491), JST AIP Trilateral AI Research (JPMJCR20G6), and the SCAT Foundation. We are particularly grateful to Jon Barker for his invaluable feedback throughout the manuscript preparation process. His insights significantly improved the clarity and focus of our work.

7. References

[1] H. Dubey, V. Gopal, R. Cutler, A. Aazami, S. Matusevych, S. Braun, S. E. Eskimez, M. Thakker, T. Yoshioka, H. Gamper, and R. Aichner, "ICASSP 2022 Deep Noise Suppression Challenge," in *Proc. of ICASSP*. IEEE, 2022, pp. 9271–9275.

[2] A. R. Avila, H. Gamper, C. K. A. Reddy, R. Cutler, I. Tashev, and J. Gehrke, "Non-intrusive speech quality assessment using neural networks," in *Proc. of ICASSP*. IEEE, 2019, pp. 631–635.

[3] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR - Half-baked or Well Done?" in *Proc. of ICASSP*. IEEE, 2019, pp. 626–630.

[4] Y. Tang, M. Cooke, and C. Valentini-Botinhao, "Evaluating the predictions of objective intelligibility metrics for modified and synthetic speech," *Computer Speech and Language*, vol. 35, pp. 73–92, 7 2016.

[5] M. Karbasi and D. Kolossa, "ASR-based speech intelligibility prediction: A review," *Hearing Research*, vol. 426, 12 2022.

[6] Y. Feng and F. Chen, "Nonintrusive objective measurement of speech intelligibility: A review of methodology," *Biomedical Signal Processing and Control*, vol. 71, p. 103204, 2022.

[7] I. López-Espejo, A. Edraki, W.-Y. Chan, Z.-H. Tan, and J. Jensen, "On the deficiency of intelligibility metrics as proxies for subjective intelligibility," *Speech Communication*, vol. 150, pp. 9–22, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016763932300050X

[8] Seung-Eun Kim and Bronya R. Chernyak and Olga Seleznova and Joseph Keshet and Matthew Goldrick and Ann R. Bradlow , "Automatic recognition of second language speech-in-noise," *JASA Express Letters*, 2024.

[9] R. E. Zezario, S. Fu, F. Chen, C. Fuh, H. Wang, and Y. Tsao, "Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 54–70, 2023.

[10] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. of ICML*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 2023, pp. 28 492–28 518.

[11] R. E. Zezario, F. Chen, C. Fuh, H. Wang, and Y. Tsao, "Utilizing whisper to enhance multi-branched speech intelligibility prediction model for hearing aids," *CoRR*, vol. abs/2309.09548, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2309.09548

[12] F. S. Oliveira, E. Casanova, A. C. Júnior, L. R. S. Gris, A. da Silva Soares, and A. R. G. Filho, "Evaluation of speech representations for MOS prediction," in *Proc. of Text, Speech, and Dialogue*, ser. Lecture Notes in Computer Science, vol. 14102. Springer, 2023, pp. 270–282.

[13] A. Koenecke, A. S. G. Choi, K. Mei, H. Schellmann, and M. Sloane, "Careless whisper: Speech-to-text hallucination harms," *ArXiv*, 2 2024. [Online]. Available: http://arxiv.org/abs/2402.08021

[14] A. Ali and S. Renals, "Word Error Rate Estimation for Speech Recognition: e-WER," in *Proc. of ACL (Vol. 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 20–24.

[15] K. Arai, A. Ogawa, S. Araki, K. Kinoshita, T. Nakatani, N. Kamo, and T. Irino, "Intelligibility prediction of enhanced speech using recognition accuracy of end-to-end asr systems," in *Proc. of APSIPA ASC*, 2022, pp. 1583–1589.

[16] A. Fang, S. Filice, N. Limsopatham, and O. Rokhlenko, "Using phoneme representations to build predictive models robust to asr errors," in *Proc. of ACM SIGIR*, ser. SIGIR '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 699–708.

[17] R. Errattahi, A. El Hannani, and H. Ouahmane, "Automatic Speech Recognition Errors Detection and Correction: A Review," *Procedia Computer Science*, vol. 128, pp. 32–37, 2018, 1st International Conference on Natural Language and Speech Processing.

[18] C. K. A. Reddy, H. Dubey, K. Koishida, A. A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "INTERSPEECH 2021 Deep Noise Suppression Challenge," in *Proc. of Interspeech*. ISCA, 2021, pp. 2796–2800.

[19] M. A. Akeroyd, W. Bailey, J. Barker, T. J. Cox, J. F. Culling, S. Graetzer, G. Naylor, Z. Podwińska, and Z. Tu, "The 2nd Clarity Enhancement Challenge for Hearing Aid Speech Intelligibility Enhancement: Overview and Outcomes," in *Proc. of ICASSP*, 2023, pp. 1–5.

[20] S. Graetzer, M. A. Akeroyd, J. Barker, T. J. Cox, J. F. Culling, G. Naylor, E. Porter, and R. Viveros-Muñoz, "Dataset of british english speech recordings for psychoacoustics and speech processing research: The clarity speech corpus," *Data in Brief*, vol. 41, p. 107951, 2022.

[21] S. A. Zollinger and H. Brumm, "The lombard effect," *Current Biology*, vol. 21, no. 16, pp. R614–R615, 2011.

[22] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92)," 2019.

[23] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," *ArXiv*, 6 2020. [Online]. Available: http://arxiv.org/abs/2006.12847

[24] H. Schröter, T. Rosenkranz, A. N. Escalante-B., and A. Maier, "DeepFilterNet: Perceptually motivated real-time speech enhancement," in *Proc. of Interspeech*, 2023.

[25] J. Chen, Z. Wang, D. Tuo, Z. Wu, S. Kang, and H. Meng, "FullSubNet+: Channel Attention FullSubNet with Complex Spectrograms for Speech Enhancement," in *Proc. of ICASSP*. IEEE, 2022, pp. 7857–7861.

[26] A. Défossez, "Hybrid spectrogram and waveform source separation," in *Proc. of ISMIR*, 2021.

[27] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A Full-Band and Sub-Band Fusion Model for Real-Time Single-Channel Speech Enhancement," in *Proc. of ICASSP*. IEEE, 2021, pp. 6633–6637.

[28] C. S. Lea, R. Vidal, A. Reiter, and G. Hager, "Temporal Convolutional Networks: A Unified Approach to Action Segmentation," in *ECCV Workshops*, 2016.

[29] "ITU-T recommendation P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2001.

[30] 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.