Interspeech 2024
1-5 September 2024, Kos, Greece

Non-Intrusive Speech Intelligibility Prediction for Hearing Aids using Whisper and Metadata

Ryandhimas E. Zezario[1][2], *Fei Chen*[3], *Chiou-Shann Fuh*[1], *Hsin-Min Wang*[2], *Yu Tsao*[2]

[1]National Taiwan University, [2]Academia Sinica,
[3]Southern University of Science and Technology

{ryandhimas, yu.tsao}@citi.sinica.edu.tw

Abstract

Automated speech intelligibility assessment is pivotal for hearing aid (HA) development. In this paper, we present three novel methods to improve intelligibility prediction accuracy and introduce MBI-Net+, an enhanced version of MBI-Net, the top-performing system in the 1st Clarity Prediction Challenge. MBI-Net+ leverages Whisper's embeddings to create cross-domain acoustic features and includes metadata from speech signals by using a classifier that distinguishes different enhancement methods. Furthermore, MBI-Net+ integrates the hearing-aid speech perception index (HASPI) as a supplementary metric into the objective function to further boost prediction performance. Experimental results demonstrate that MBI-Net+ surpasses several intrusive baseline systems and MBI-Net on the Clarity Prediction Challenge 2023 dataset, validating the effectiveness of incorporating Whisper embeddings, speech metadata, and related complementary metrics to improve prediction performance for HA.

Index Terms: speech intelligibility, hearing aid, hearing loss, weak supervision, cross-domain features

1. Introduction

Accurate metrics for predicting speech intelligibility are crucial for optimizing the efficacy of various speech-related applications, such as speech enhancement [1, 2], hearing aid (HA) devices [3, 4], and telecommunications [5, 6]. The most direct method of evaluation is to conduct a human listening test. However, for a fair evaluation of results, a large-scale listening test is typically required, which can be costly and less practical. Therefore, various objective speech intelligibility measures have been proposed. Traditionally, these measures are derived based on signal processing and psychoacoustic knowledge, such as speech intelligibility index (SII) [7], extended SII (ESII) [8], speech transmission index (STI) [9], short-time objective intelligibility (STOI) [10], modified binaural short-time objective intelligibility (MBSTOI) [11], gammachirp envelope similarity index (GESI) [12], and the hearing aid speech perception index (HASPI)[13]. However, these traditional measures have limited applicability, as they generally require a clean reference, which may not always be available in real-world scenarios.

With the emergence of deep learning models, numerous non-intrusive automatic speech assessment models have been developed, such as [14, 15, 16, 17, 18, 19, 20, 21]. Meanwhile, many studies have focused on developing speech intelligibility prediction models for HA [22, 23, 24, 25, 26]. For example, [22] encodes the hearing loss pattern into a vector, merges it with speech signals, and feeds it into a deep learning model to predict two hearing aid evaluation metrics: HASPI and the hearing aid speech quality index (HASQI) [23]. Besides, [24] utilizes hidden layer representations of an automatic speech recognition (ASR) model as acoustic features for predicting speech intelligibility scores. Furthermore, [25] introduces a multi-branched speech intelligibility prediction model, namely MBI-Net, consisting of two branch modules to estimate the frame-level scores of the left and right channel inputs; the outputs of the two branches are then concatenated and fused in a linear layer to produce the final intelligibility prediction score for HA.

While MBI-Net [25] demonstrates strong performance and achieves top results in the 1st Clarity Prediction Challenge [3], there are several options available to further enhance its performance. In this study, we explore three methods and introduce an advanced version of MBI-Net, termed MBI-Net+. First, MBI-Net+ incorporates Whisper [27] embeddings to generate cross-domain features. Considering Whisper underwent training using a comprehensive dataset comprising 680,000 utterances along with their respective transcripts, we assume that the extracted embedding features possess the capability to encapsulate rich phonetic information. Second, MBI-Net+ includes metadata from speech signals to mitigate potential prediction bias. This addition involves incorporating an extra classifier to distinguish different types of speech enhancement methods (front-end processors in HAs). We assume that if the model can optimally distinguish speech signals processed by different enhancement methods (as metadata), the model has a better understanding of differentiating the front-end processing of HA, potentially improving the prediction performance. In the following discussion, we refer to this classifier as the system classifier (SC). Third, MBI-Net+ utilizes the HASPI measure as a complementary metric, employing a multi-task learning criterion for model training. This method builds upon the previous study by [28], which integrates additional related metrics correlated with the main metric to enhance prediction performance.

To implement MBI-Net+, the input speech data is initially divided into two channels, with one channel dedicated to the left ear and the other to the right ear of an HA. Each channel is processed by a unified module that combines feature extraction and neural network modeling. These two branches of modules simultaneously predict frame-level scores of intelligibility and HASPI, while also extracting embedding representations for the SC module. The predicted frame-level scores from both branch modules are then combined and fused via two task-specific modules to obtain final utterance-level scores for intelligibility and HASPI. Each task-specific module consists of a linear layer followed by a global average pooling layer. The embedding representations from each task-independent module

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62371217).

10.21437/Interspeech.2024-716

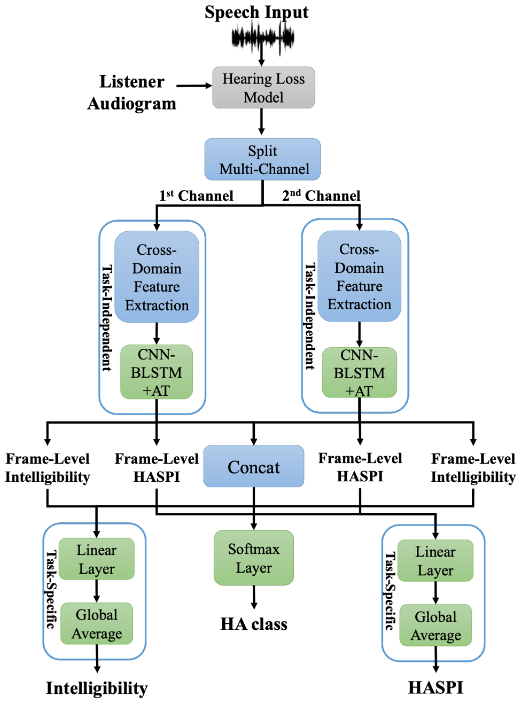

Figure 1: *Architecture of the MBI-Net+ model.*

are concatenated and processed by a dense layer with softmax to form the SC module that predicts the HA system.

Experimental results demonstrate that MBI-Net+ achieves a 7.11% improvement in prediction performance, measured by root mean square error (RMSE), compared to the original MBI-Net model. This confirms the advantages of Whisper in deploying cross-domain features over the self-supervised learning (SSL) speech model, WavLM [29], as well as the benefits of incorporating additional metadata. In the Clarity Prediction Challenge 2023 dataset, MBI-Net+ ranked third in overall performance compared to other non-intrusive systems, with comparable correlation values (0.76 (MBI-Net+) vs 0.78 (1st place [30]) vs 0.77 (2nd place [31])). It's noteworthy that MBI-Net+ requires less GPU memory compared to [30, 31]. This reduction can be attributed to MBI-Net+'s utilization of the last layer of Whisper to extract embedding representations. Unlike other approaches, MBI-Net+ doesn't require the extraction of each individual transformer output of Whisper and additional processing before forwarding it to the subsequent stage of training.

The remainder of this paper is organized as follows. Section II presents the proposed MBI-Net+. Section III describes the experimental setup and results. Finally, Section IV provides conclusions and future work.

2. MBI-Net+

The proposed MBI-Net+ consists of multi-branched task-independent modules characterizing each speech signal channel in a binaural HA. Then, task-specific modules are used to predict intelligibility and HASPI scores, respectively. The overall architecture of MBI-Net+ and its task-independent module are illustrated in Figs. 1 and 2, respectively. Firstly, we utilize the pre-trained Whisper model instead of the pre-trained SSL models, HuBERT [32] and WavLM [29], employed in MBI-Net, to incorporate cross-domain features. We assume that Whisper features provide better phonetic information than SSL features, considering their larger training size and access to transcripts during deployment. Secondly, we include an SC module to categorize enhancement systems that were used to process speech signals. Considering their distinct characteristics, we assume that if MBI-Net+ can distinguish the different properties of processed speech signals, it can achieve higher prediction capabilities. This improvement is due to MBI-Net+ incorporating additional information when capturing acoustic information, potentially reducing overfitting during training. Thirdly, we introduce HASPI as a complementary metric. We assume HASPI correlates with subjective intelligibility score and, therefore, can potentially reduce overfitting and improve the prediction performance in multi-task learning scenarios [28]. Furthermore, The loss function, O, for training MBI-Net+ is as follows:

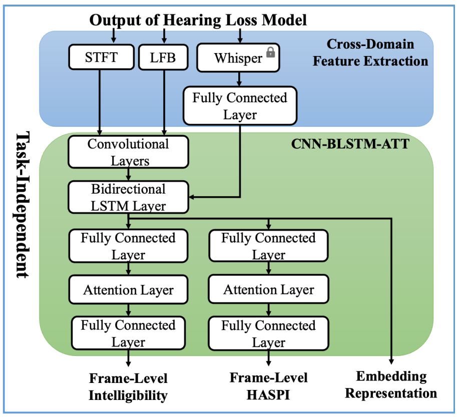

Figure 2: *Illustration of extracting cross-domain features and estimating frame-level intelligibility scores using the CNN-BLSTM+AT architecture.*

$$O = \gamma_1 \mathcal{L}_{Int} + \gamma_2 \mathcal{L}_{HASPI} + \gamma_3 \mathcal{L}_{CE}, \tag{1}$$

where $\mathcal{L}_{CE}$ represents the cross-entropy calculation from the SC module between the estimated and reference types of enhancement methods (front-end processors in HA systems). Specifically, we employ ten classes to account for the ten different enhancement systems used as input in our MBI-Net+. γ_1, γ_2, γ_2 are the weights between losses. $\mathcal{L}_{Int}$ is calculated as

$$\begin{aligned}\mathcal{L}_{Int} = \frac{1}{U}\sum_{u=1}^{U}[(I_u - \hat{I}_u)^2 + \frac{\alpha}{F_u}\sum_{f=1}^{F_u}(I_u - \hat{i}_{u,f})^2] + \\ \mathcal{L}_{left-int} + \mathcal{L}_{right-int},\end{aligned} \tag{2}$$

where I_u, $\hat{I}_u$, and $\hat{i}_{u,f}$ denote the true utterance-level score, predicted utterance-level score, and predicted frame-level score (merged from the first and second channels by the linear layer in Figure 1) of intelligibility, respectively; U denotes the total number of training utterances; F_u denotes the number of frames in the u-th training utterance; α is a weight between utterance-level and frame-level losses; $\mathcal{L}_{left-int}$ and $\mathcal{L}_{right-int}$ are the frame-level losses of the left (l) branch (i.e., the first channel) and right (r) branch (i.e., the second channel) in frame-level intelligibility estimation (as shown in Figure 2), which are calculated as

$$\begin{aligned}\mathcal{L}_{left-int} = \frac{\alpha^l}{F_u}\sum_{f=1}^{F_u}(I_u - \hat{i}^l_{u,f})^2 \\ \mathcal{L}_{right-int} = \frac{\alpha^r}{F_u}\sum_{f=1}^{F_u}(I_u - \hat{i}^r_{u,f})^2,\end{aligned} \tag{3}$$

where α^l and α^r are weights of *l* and *r* branches, respectively, and $\hat{i}^l_{u,f}$ and $\hat{i}^r_{u,f}$ denote the predicted frame-level scores of *l* and *r* branches, respectively. $\mathcal{L}_{HASPI}$ is calculated as

$$\mathcal{L}_{HASPI} = \frac{1}{U}\sum_{u=1}^{U}[(H_u - \hat{H}_u)^2 + \frac{\beta}{F_u}\sum_{f=1}^{F_u}(H_u - \hat{h}_{u,f})^2] + \mathcal{L}_{left-haspi} + \mathcal{L}_{right-haspi}, \tag{4}$$

where H_u, $\hat{H}_u$, and $\hat{h}_{u,f}$ denote the true utterance-level score, predicted utterance-level score, and predicted frame-level score (merged from the first and second channels by the linear layer in Figure 1) of HASPI, respectively; β is a weight between utterance-level and frame-level losses, and $\mathcal{L}_{left-haspi}$ and $\mathcal{L}_{right-haspi}$ are the frame-level losses of *l* branch and *r* branch in frame-level HASPI estimation calculated as

$$\begin{aligned}\mathcal{L}_{left-haspi} &= \frac{\beta^l}{F_u}\sum_{f=1}^{F_u}(H_u - \hat{h}^l_{u,f})^2 \\ \mathcal{L}_{right-haspi} &= \frac{\beta^r}{F_u}\sum_{f=1}^{F_u}(H_u - \hat{h}^r_{u,f})^2,\end{aligned} \tag{5}$$

where β^l and β^r are weights of *l* and *r* branches, and $\hat{h}^l_{u,f}$ and $\hat{h}^r_{u,f}$ denote the predicted frame-level scores of *l* and *r* branches, respectively. Furthermore, we hypothesize that the supplementary information from HASPI and the SC module can improve overall prediction performances. Subsequently, the corresponding frame-level scores are combined and integrated through two linear layers. This combination produces the final predictive scores for intelligibility and HASPI scores.

3. Experiments

3.1. Experimental Setup

The 2023 Clarity Prediction Challenge (CPC) dataset [33] comprises scenes associated with 6 talkers, 10 enhancement methods (accordingly 10 HA systems) from the 2022 Clarity Enhancement Challenge [34], and 25 listeners who rated the intelligibility scores. To elaborate, the dataset is divided into three tracks. The first track consists of 2779 utterances, the second track consists of 2796 utterances, and the third track consists of 2772 utterances. For each track, we selected 90% of the utterances as the training set and the rest as the development set. Additionally, the three tracks have 305, 294, and 298 test utterances, respectively. Please note that the test involves unseen listeners and unseen HA systems. Specifically, the model is trained and tested on three tracks. In our experiment, the overall performance across all tracks is denoted as All. Three evaluation metrics, namely root mean square error (RMSE), linear correlation coefficient (LCC), and Spearman's rank correlation coefficient (SRCC) [35] were used to evaluate the prediction performance. A lower RMSE value means that the predicted scores are closer to the ground-truth scores (lower is better). In contrast, higher LCC and SRCC values indicate a higher correlation between the predicted scores and the ground-truth scores (higher is better). All models compared in this paper were trained and evaluated on the CPC 2023 dataset.

3.2. Correlation of Intelligibility, HASPI, and Distance of Whisper Embeddings

In the first experiment, we aim to analyze the correlation between subjective intelligibility, HASPI, and the distance of Whisper embeddings, d_{ws}. A higher correlation indicates a stronger relationship. We aim to investigate whether considering these attributes, MBI-Net+ can achieve better predictions. We estimate d_{ws} using the following equations

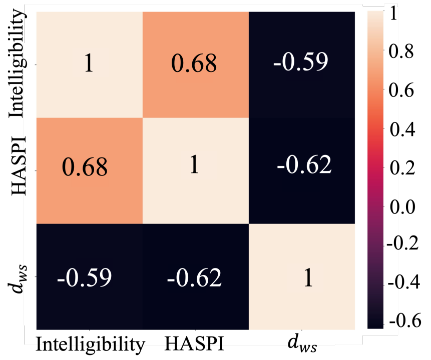

Figure 3: *Correlation analysis between Intelligibility, HASPI, and Whisper.*

Table 1: *LCC, SRCC, and MSE results of MBI-Net, MBI-Net+(w/o SC) and MBI-Net+ on the development set.*

Model	Feature	LCC	SRCC	RMSE
Track 1				
MBI-Net [25]	WavLM	0.724	0.719	28.707
MBI-Net+(*w/o SC-H*)	Whisper	0.754	0.738	27.221
MBI-Net+(*w/o SC*)	Whisper	0.758	0.748	26.857
MBI-Net+	Whisper	**0.773**	**0.763**	**26.081**
Track 2				
MBI-Net [25]	WavLM	0.742	0.749	29.328
MBI-Net+(*w/o SC-H*)	Whisper	0.794	0.788	26.011
MBI-Net+(*w/o SC*)	Whisper	0.799	0.797	25.844
MBI-Net+	Whisper	**0.804**	**0.799**	**25.392**
Track 3				
MBI-Net [25]	WavLM	0.795	0.772	27.155
MBI-Net+(*w/o SC-H*)	Whisper	0.797	0.764	24.706
MBI-Net+(*w/o SC*)	Whisper	0.801	0.765	24.800
MBI-Net+	Whisper	**0.817**	**0.773**	**23.575**

$$d_{ws} = \sum_{t,f}^{T,F}(\mathbf{X}_{ws}[t,f] - \mathbf{Y}_{ws}[t,f])^2 \tag{6}$$

where $\mathbf{X}_{ws}[t,f]$ and $\mathbf{Y}_{ws}[t,f]$ represent the last encoder output of Whisper at the t-th time index and f-th element from the clean waveform $\mathbf{X}$ and enhanced waveform $\mathbf{Y}$, respectively. T and F denote the time length and the feature dimension of the embedding representation.

Figure 3 illustrates a moderate correlation score of 0.68 between HASPI and intelligibility, suggesting HASPI's ability to surrogate subjective intelligibility. Interestingly, the distance of Whisper embedding, namely d_{ws}, demonstrates correlation scores of -0.59 and -0.62 with intelligibility and HASPI, respectively. This suggests a moderate degree of correlation between Whisper's representation and both intelligibility and HASPI. Therefore, we assume that employing Whisper embedding could be beneficial for estimating subjective intelligibility and HASPI scores.

Table 2: *LCC, SRCC, and MSE results of MBI-Net, MBI-Net+(w/o SC), MBI-Net+(w/o SC-H), and MBI-Net+ on the test set.*

Model	Feature	LCC	SRCC	RMSE
Track 1				
MBI-Net [25]	WavLM	0.669	0.665	30.260
MBI-Net+(*w/o SC-H*)	Whisper	0.703	0.690	29.270
MBI-Net+(*w/o SC*)	Whisper	0.711	0.692	29.010
MBI-Net+	Whisper	**0.721**	**0.714**	**28.370**
Track 2				
MBI-Net [25]	WavLM	0.710	0.710	29.660
MBI-Net+(*w/o SC-H*)	Whisper	0.751	0.737	26.730
MBI-Net+(*w/o SC*)	Whisper	**0.765**	**0.764**	26.440
MBI-Net+	Whisper	0.754	0.737	**25.920**
Track 3				
MBI-Net [25]	WavLM	0.810	0.834	23.920
MBI-Net+(*w/o SC-H*)	Whisper	0.808	**0.840**	24.040
MBI-Net+(*w/o SC*)	Whisper	**0.826**	0.834	**23.340**
MBI-Net+	Whisper	0.813	0.814	23.740
Track All				
MBI-Net [25]	WavLM	0.724	0.729	28.100
MBI-Net+(*w/o SC-H*)	Whisper	0.754	0.752	26.790
MBI-Net+(*w/o SC*)	Whisper	**0.767**	0.765	26.390
MBI-Net+	Whisper	0.764	**0.767**	**26.100**

3.3. Comparing MBI-Net+ with Original MBI-Net

In the first experiment, we aim to compare the performance of MBI-Net+ and MBI-Net. The MBI-Net+ entails: 1) employing a 257-dimensional spectral feature; 2) leveraging Whisper medium as the pre-trained model, with the final layer's output serving as input for our MBI-Net+; 3) utilizing four convolutional layers each comprising channels with sizes of 16, 32, 64, and 128, one-layered Bidirectional Long Short-Term Memory (BLSTM) with 128 nodes, fully connected layer containing 128 neurons, attention mechanism, and one neuron of frame level score; 5) leveraging two task-specific modules and SC module (with ten classes); 6) setting a learning rate of 0.0001 with Adam optimizer [36]. Note that we developed three different versions of MBI+, namely MBI-Net+(*w/o SC*), MBI-Net+(*w/o SC-H*), and MBI-Net+. MBI-Net+ represents the complete configuration in Figure 1, and MBI-Net+(*w/o SC*) follows the same model architecture as shown in Figure 1, except not including the SC module. On the other hand, MBI-Net+(*w/o SC-H*) does not include HASPI and SC modules.

As shown in Tables 1 and 2, all versions of MBI-Net+ can outperform MBI-Net. Please note that MBI-Net+(*w/o SC-H*) shares the same model architecture as MBI-Net but uses different speech representations (SSL for MBI-Net and Whisper for MBI-Net+(*w/o SC-H*)). From Tables 1 and 2, the comparison between MBI-Net+(*w/o SC-H*) and MBI-Net first confirms the advantage of incorporating Whisper embeddings for deploying cross-domain features. Next, the comparison between MBI-Net+(*w/o SC*) and MBI-Net+(*w/o SC-H*) confirms the effectiveness of utilizing supplementary HASPI metric which has moderate correlation with the main objective metric in preventing overfitting in multi-task learning scenario. Finally, the comparison between MBI-Net+ and MBI-Net+(*w/o SC*) shows that the SC module contributes to better prediction performance, confirming the advantages of incorporating additional metadata of the speech signals.

Table 3: *RMSE and LCC scores of all compared systems from the First and Second Clarity Challenge on the test set.*

System	Non-Intrusive	RMSE	LCC
E011 [30]	Yes	**25.1**	**0.78**
E002 [31]	Yes	25.3	0.77
MBI-Net+	Yes	26.1	0.76
MBI-Net+(*w/o SC*)	Yes	26.4	0.76
MBI-Net+(*w/o SC-H*)	Yes	26.8	0.75
E025 [33]	Yes	27.9	0.72
MBI-Net [25]	Yes	28.1	0.72
Baseline [33]	No	28.7	0.70
E003 [33]	Yes	31.1	0.64
E024 [33]	Yes	31.7	0.62
E015 [33]	Yes	35.0	0.60
E020 [33]	Yes	39.8	0.33
Prior [3]	No	40	-

3.4. Comparison with Other Prediction Models for HA

In the third experiment, we compare MBI-Net+ models with other speech intelligibility prediction models for HA. Four systems (E011 [30], E002 [31], MBI-Net [25], E025 [33]) utilized acoustic features from large pre-trained models, while the other four [33] (E003, E024, E015, E020) utilized signal processing techniques for feature extraction. The overall RMSE and LCC scores from the three tracks are shown in Table 3. We first confirm that all versions of MBI-Net+ can perform better than the baseline [33] model that adopts an intrusive-based method. Furthermore, our models achieve good performance, ranking third (MBI-Net+) overall among non-intrusive speech intelligibility prediction models. The LCC score of MBI-Net+ is 0.76, which is very close to the best score (0.78) and the second-best score (0.77). It's noteworthy that MBI-Net+ requires less GPU memory compared to [30, 31]. Considering MBI-Net+ doesn't require the extraction of each individual transformer output of Whisper and additional processing before forwarding it to the subsequent stage of training. Finally, this result demonstrates the potential benefits of adopting a multi-task, multi-branched model architecture and incorporating Whisper-based cross-domain features along with adding supplementary metrics and the SC module into better prediction capabilities.

4. Conclusion

In this study, we have proposed MBI-Net+, a novel speech intelligibility prediction model for HA. MBI-Net+ implements a multi-branched architecture that incorporates Whisper-based cross-domain features and includes the supplementary HASPI metric and the SC module. Experimental results first confirm a moderate correlation between subjective intelligibility, HASPI, and distance of Whisper embeddings, d_{ws}, suggesting potential advantages for their joint combinations. Next, we confirm the effectiveness of MBI-Net+ by achieving a notable improvement over MBI-Net and yielding a top-3 performance among non-intrusive systems in the Clarity Prediction Challenge 2023. Since MBI-Net+ only utilizes the final layer of the pre-trained Whisper model, this approach avoids the heavy computational requirements of other top-performing systems. In future work, we plan to investigate the incorporation of other types of metadata from speech signals and model architectures when deploying speech intelligibility prediction models for HA.

5. References

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC press, 2007.

[2] G. Close, T. Hain, and S. Goetze, "The Effect of Spoken Language on Speech Enhancement Using Self-Supervised Speech Representation Loss Functions," in *Proc. WASPAA*, 2023, pp. 1–5.

[3] J. Barker, M. Akeroyd, J. Trevor, J. Culling, J. Firth, S. Graetzer, H. Griffiths, L. Harris, G. Naylor, Z. Podwinska, E. Porter, and R. Munoz, "The 1st Clarity Prediction Challenge: A Machine Learning Challenge for Hearing Aid Intelligibility Prediction," in *Proc. INTERSPEECH*, 2022, pp. 3508–3512.

[4] B. Schuh, W. Wardah, B. Naderi, T. Michal, and S. Moeller, "Hearing Impairment in Crowdsourced Speech Quality Assessments: Its Effect and Screening with Digit Triplet Hearing Test," in *Speech Communication; 15th ITG Conference*, 2023, pp. 21–25.

[5] G. Yi, W. Xiao, Y. Xiao, B. Naderi, S. Möller, W. Wardah, G. Mittag, R. Culter, Z. Zhang, D. S. Williamson, F. Chen, F. Yang, and S. Shang, "ConferencingSpeech 2022 Challenge: Non-Intrusive Objective Speech Quality Assessment (NISQA) Challenge for Online Conferencing Applications," in *Proc. INTERSPEECH*, 2022, pp. 3308–3312.

[6] G. Mittag and S. Möller, "Non-intrusive Speech Quality Assessment for Super-wideband Speech Communication Networks," in *Proc. ICASSP*, 2019, pp. 7125–7129.

[7] ANSI Std. S3.5 1997, "Methods for Calculation of The Speech Intelligibility Index," in *Acoustical Society of America*, 1997.

[8] T. Houtgast and H. . M. Steeneken, "Evaluation of Speech Transmission Channels by Using Artificial Signals," *Acustica*, vol. 25, no. 6, pp. 355–367, 1971.

[9] H. J. M. Steeneken and T. Houtgast, "A Physical Method for Measuring Speech-Transmission Quality," *Journal of the Acoustical Society of America*, vol. 67, no. 1, pp. 318–326, 1980.

[10] 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.

[11] A. H. Andersenan, J. M. Haan, Z.-H. Tan, and J.Jensen, "Refinement and Validation of The Binaural Short Time Objective Intelligibility Measure for Spatially Diverse Conditions," *Speech Communication*, vol. 102, pp. 1–13, 2018.

[12] T. Irino, H. Tamaru, and A. Yamamoto, "Speech Intelligibility of Simulated Hearing Loss Sounds and its Prediction using the Gammachirp Envelope Similarity Index (GESI)," in *Proc. INTERSPEECH*, 2022, pp. 3929–3933.

[13] J. M. Kates and K. H. Arehart, "The Hearing-Aid Speech Perception Index (HASPI) Version 2," *Speech Communication*, vol. 131, pp. 35–46, 2021.

[14] X. Dong and D. S. Williamson, "An Attention Enhanced Multi-Task Model for Objective Speech Assessment in Real-World Environments," in *Proc. ICASSP*, 2020, pp. 911–915.

[15] Y. Leng, X. Tan, S. Zhao, F. Soong, X.-Y. Li, and T. Qin, "MBNet: MOS Prediction for Synthesized Speech with Mean-Bias Network," in *Proc. ICASSP*, 2021, pp. 391–395.

[16] E. Cooper, W.-H. Huang, T. Toda, and J. Yamagishi, "Generalization Ability of MOS Prediction Networks," in *Proc. ICASSP*, 2022, pp. 8442–8446.

[17] Z. Yang, W. Zhou, C. Chu, S. Li, R. Dabre, R. Rubino, and Y. Zhao, "Fusion of Self-Supervised Learned Models for MOS Prediction," in *Proc. INTERSPEECH*, 2022, pp. 5443–5447.

[18] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022," in *Proc. INTERSPEECH*, 2022, pp. 4521–4525.

[19] R. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wan, and Y. Tsao, "Deep Learning-Based Non-Intrusive Multi-Objective Speech Assessment Model With Cross-Domain Features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 54–70, 2023.

[20] R. Zezario, S.-W. Fu, F. Chen, C. S. Fuh, H.-M. Wang, and Y. Tsao, "MTI-Net: A Multi-Target Speech Intelligibility Prediction Model," in *Proc. INTERSPEECH*, 2022, pp. 5463–5467.

[21] E. Cooper, W.-C. Huang, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "A Review on Subjective and Objective Evaluation of Synthetic Speech," *Acoustical Science and Technology*, vol. advpub, p. e24.12, 2024.

[22] H.-T. Chiang, S.-W. Fu, H.-M. Wang, Y. Tsao, and J. H. L. Hansen, "Multi-objective Non-intrusive Hearing-aid Speech Assessment Model," *arXiv:2311.08878*, 2023.

[23] J. M. Kates and K. H. Arehart, "The Hearing-Aid Speech Quality Index (HASQI) Version 2," *Journal of the Audio Engineering Society*, vol. 62, no. 3, pp. 99–117, 2014.

[24] Z. Tu, N. Ma, and J. Barker, "Exploiting Hidden Representations from a DNN-based Speech Recogniser for Speech Intelligibility Prediction in Hearing-Impaired Listeners," in *Proc. INTERSPEECH*, 2022, pp. 3488–3492.

[25] R. Zezario, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "MBI-Net: A Non-intrusive Multi-Branched Speech Intelligibility Prediction Model for Hearing Aids," in *Proc. INTERSPEECH*, 2022, pp. 3944–3948.

[26] C. O. Mawalim, B. A. Titalim, S. Okada, and M. Unoki, "Non-Intrusive Speech Intelligibility Prediction Using an Auditory Periphery Model with Hearing Loss," *Applied Acoustics*, vol. 214, p. 109663, 2023.

[27] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in *Proc. ICML*, 2023, pp. 28 492–28 518.

[28] R. E. Zezario, B.-R. B. Bai, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Multi-Task Pseudo-Label Learning for Non-Intrusive Speech Quality Assessment Model," in *Proc. ICASSP*, 2024, pp. 831–835.

[29] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[30] S. Cuervo and R. Marxer, "Speech Foundation Models on Intelligibility Prediction for Hearing-Impaired Listeners," in *Proc. ICASSP*, 2024, pp. 1421–1425.

[31] R. Mogridge, G. Close, R. Sutherland, T. Hain, J. Barker, S. Goetze, and A. Ragni, "Non-Intrusive Speech Intelligibility Prediction for Hearing-Impaired Users Using Intermediate ASR Features and Human Memory Models," in *Proc. ICASSP*, 2024, pp. 306–310.

[32] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 3451–3460, 2021.

[33] J. P. Barker, M. A. Akeroyd, W. Bailey, T.J.Cox, J. F. Culling, J. Firth, S. Graetzer, and G. Naylor, "The 2nd Clarity Prediction Challenge: A Machine Learning Challenge for Hearing Aid Intelligibility Prediction," in *Proc. ICASSP*, 2024, pp. 11 551–11 555.

[34] 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. ICASSP*, 2023, pp. 1–5.

[35] C. Spearman, "The Proof and Measurement of Association between Two Things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.

[36] D. Kingma and J.Ba, "Adam: A Method for Stochastic Optimization," in *Proc. ICLR*, 2015, pp. 1–13.