Feature Importance across Domains for Improving Non-Intrusive Speech Intelligibility Prediction in Hearing Aids

Ryandhimas E. Zezario[1], *Sabato M. Siniscalchi*[2], *Fei Chen*[3], *Hsin-Min Wang*[1], *Yu Tsao*[1]

[1]Academia Sinica, Taiwan
[2]University of Palermo, Italy
[3]Southern University of Science and Technology, China

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

Abstract

Given the critical role of non-intrusive speech intelligibility assessment in hearing aids (HA), this paper enhances its performance by introducing Feature Importance across Domains (FiDo). We estimate feature importance on spectral and time-domain acoustic features as well as latent representations of Whisper. Importance weights are calculated per frame, and based on these weights, features are projected into new spaces, allowing the model to focus on important areas early. Next, feature concatenation is performed to combine the features before the assessment module processes them. Experimental results show that when FiDo is incorporated into the improved multi-branched speech intelligibility model MBI-Net+, RMSE can be reduced by 7.62% (from 26.10 to 24.11). MBI-Net+ with FiDo also achieves a relative RMSE reduction of 3.98% compared to the best system in the 2023 Clarity Prediction Challenge. These results validate FiDo's effectiveness in enhancing neural speech assessment in HA.

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

1. Introduction

Assessing speech intelligibility is an important factor in evaluating the performance of various speech processing applications, such as speech enhancement [1, 2] and hearing aid (HA) devices [3, 4]. The most direct and reliable assessment method is for human listeners to assess how many words they can correctly recognize from a given audio sample. However, despite the reliability of human-based assessment scores, cost and practicality remain significant challenges in obtaining unbiased assessment scores. Therefore, more practical assessment methods that mimic human-based assessments are needed.

Previously, most traditional methods relied on signal processing and psychoacoustic principles to replicate human auditory perception for intelligibility assessment. Well-known methods include the speech intelligibility index (SII) [5], extended SII (ESII) [6], speech transmission index (STI) [7], short-time objective intelligibility (STOI) [8], modified binaural short-time objective intelligibility (MBSTOI) [9], and hearing aid speech perception index (HASPI) [10]. Despite the notable assessment performance of these methods, the availability of clean references is crucial to obtain more accurate assessment scores, which limits their practicality in situations where clean audio is not available as a ground-truth reference.

Recently, with the emergence of deep learning models and the availability of large-scale human assessment labels, researchers have begun to use deep learning for speech assessment [11, 12, 13, 14, 15, 16], with the goal of replicating human assessment behavior and improving the practicality of assessment methods. Interestingly, deep learning models have shown great potential in assessing speech quality and intelligibility in a non-intrusive manner, and most current deep learning-based speech assessment systems integrate large pre-trained models (such as self-supervised learning (SSL) models [11, 12, 14] or Whisper [13, 15, 16]) and utilize ensemble learning [14, 15, 16].

With two Clarity Prediction Challenges [17, 18], there is growing interest in deploying non-intrusive speech intelligibility prediction models for hearing aids (HA) [15, 16, 19, 20, 21, 22]. In the first Clarity Prediction Challenge [18], the best-performing non-intrusive system combines hidden layer representations of automatic speech recognition (ASR) [21] and hidden layers of SSL models [23] to achieve robust performance. In the more recent challenge [17], the use of Whisper-generated acoustic features proved to be more effective in improving robustness, as the three best non-intrusive models all employed Whisper to generate acoustic features [15, 16, 19]. Moreover, we note that the robustness of acoustic features plays a crucial role in achieving higher prediction performance. Therefore, further research on the most effective strategies for extracting acoustic features is necessary.

Inspired by the concept of estimating feature importance in computer vision [24], which has been shown to be effective in helping models focus on important regions from the early stages of the training process, we aim to adopt this concept and propose feature importance across domains (FiDo). To demonstrate proof-of-concept of our approach, we aim to incorporate the FiDo approach into MBI-Net+ [19]. We choose MBI-Net+ as our base model because of its simple module implementation and its excellent performance in the latest Clarity Prediction Challenge. In our implementation, FiDo adopts a different feature extraction strategy compared to the original cross-domain feature extraction of MBI-Net+. First, we introduce feature-based concatenation when processing cross-domain acoustic features. This ensures that we maintain the original number of frames, thus avoiding an excessive number of frames when concatenating in the temporal dimension, as implemented in the original MBI-Net+. Second, we introduce multi-head self-attention (MHSA) [25] with cross-domain feature concatenation to effectively capture the importance of acoustic features. Since FiDo utilizes three different types of acoustic features: spectral (PS), filterbank (FB), and Whisper (WS) features, we first concatenate the MHSA outputs of PS and FB along the feature dimension. These features are then processed through a convolutional neural network and combined with the Whisper features processed using MHSA and an adapter layer. The final output of the FiDo feature extraction

The work of Fei Chen was supported by the National Natural Science Foundation of China (Grant No. 62371217).

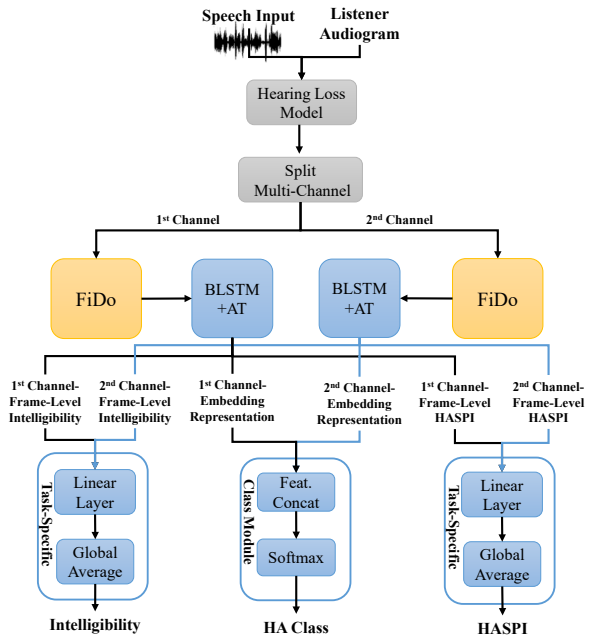


Figure 1: *Architecture of the MBI-Net+ model with FiDo.*

method is the combination of these two feature outputs. For the assessment module, we follow the original MBI-Net+ architecture, which includes BLSTM with attention, task-specific modules for estimating intelligibility, HASPI, and HA classes. Experimental results show that integrating FiDo into MBI-Net+ results in an relative RMSE reduction of 7.62% (from 26.10% to 24.11%). Furthermore, MBI-Net+ with FiDo achieves an relative RMSE reduction of 3.98% over the best-performing system in the 2023 Clarity Prediction Challenge. These results confirm FiDo's effectiveness in improving neural speech intelligibility assessment in HA.

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

2. FiDo

The motivation behind FiDo is to estimate attention weights immediately after feature extraction, with the expectation that it can improve prediction accuracy by making the model focus on salient acoustic features from the beginning. Unlike existing speech assessment methods that apply attention mechanisms at deeper layers [16, 19, 15], FiDo introduces early attention modulation to refine feature representations before higher-level processing. The overall integration of FiDo and MBI-Net+ is shown in Fig. 1.

The FiDo acoustic feature extraction approach is applied immediately after splitting the simulated audio (compensated for specific audiogram characteristics) into two separate channels (y^l, y^r). As illustrated in Fig. 2, two distinct FiDo modules are used to process each audio channel input. In the detailed process, given the simulated audio input of the left channel y^l (the steps for the right channel are the same), the corresponding features (PS, FB, and WS) are extracted as follows:

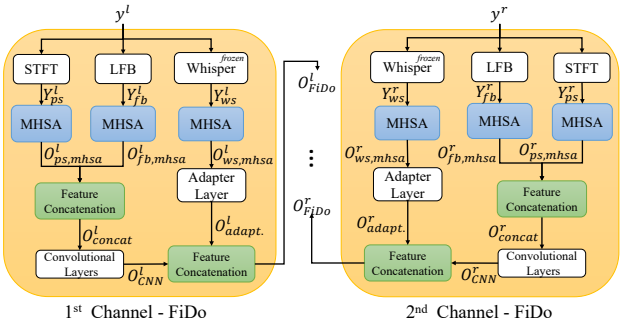


Figure 2: *A detailed procedure of FiDo for both branches.*

$$
\begin{aligned}
Y^l_{ps} &= \text{STFT}(y^l), \\
Y^l_{fb} &= \text{LFB}(y^l), \\
Y^l_{ws} &= \text{Whisper}(y^l).
\end{aligned} \tag{1}
$$

To ensure that the three types of features extracted from an audio input have the same number of time frames, we perform zero padding to 7 seconds on each input waveform and adjust Whisper's chunk size to 7-seconds. In addition, for Y^l_{ps} and Y^l_{fb}, we apply the same configuration to the STFT operation and the SincNet [26] layer of learnable filterbank (LFB) to be consistent with Whisper's frame rate.

After obtaining the three corresponding features—$Y^l_{ps} \in \mathbb{R}^{T\times(d_{ps})}$, $Y^l_{fb} \in \mathbb{R}^{T\times(d_{fb})}$, and $Y^l_{ws} \in \mathbb{R}^{T\times(d_{ws})}$—where T represents the number of time frames, and d_{ps}, d_{fb}, and d_{ws} denote the feature dimensions, the process of estimating feature importance is formulated as follows:

$$
\begin{aligned}
Q^l_{ps} = Y^l_{ps} W^l_{ps,Q}, \quad K^l_{ps} = Y^l_{ps} W^l_{ps,K}, \quad V^l_{ps} = Y^l_{ps} W^l_{ps,V}, \\
A^l_{ps} = \text{softmax}\left(\frac{Q^l_{ps} {K^l_{ps}}^\top}{\sqrt{d_k}}\right), \\
O^l_{ps,h} = A^l_{ps} V^l_{ps}, \\
O^l_{ps,mhsa} = \text{Concat}(O^l_1, O^l_2, \ldots, O^l_H) W^l_O.
\end{aligned} \tag{2}
$$

Y^l_{ps} represents the input feature; $W^l_{ps,Q}$, $W^l_{ps,K}$, and $W^l_{ps,V}$ are learnable weight matrices that project the input Y^l_{ps} into the query (Q^l_{ps}), key (K^l_{ps}), and value (V^l_{ps}) matrices, respectively. The attention matrix A^l_{ps} is calculated using the scaled dot-product of the queries and keys, scaled by the square root of the key dimension d_k. The attention output for the h-th attention head, $O^l_{ps,h}$, is calculated as the product of the attention weights A^l_{ps} and the value matrix V^l_{ps}. The outputs $O^l_1, O^l_2, \ldots, O^l_H$ correspond to the results of H attention heads, computed in parallel. Finally, W^l_O is the final projection matrix that combines the outputs of all attention heads. Following the operation described in Eq. (2), by inputting Y^l_{fb} and Y^l_{ws} with their corresponding attention weights, we can obtain $O^l_{fb,mhsa}$ and $O^l_{ws,mhsa}$, respectively.

In the subsequent steps of FiDo, the concatenation of $O^l_{ps,mhsa}$ and $O^l_{fb,mhsa}$ is processed through feature-based concatenation, followed by projecting the concatenated features into the CNN module for higher-level projection steps. The detailed process is as follows.

$$
\begin{aligned}
O^l_{concat} = \text{Concat}(O^l_{ps,mhsa}, O^l_{fb,mhsa}) \in \mathbb{R}^{T\times(d_{ps}+d_{fb})}, \\
O^l_{CNN} = \text{CNN}(O^l_{concat}).
\end{aligned} \tag{3}
$$

In the final stage when FiDo obtains $O^l_{\text{FiDo}} \in \mathbb{R}^{T\times(d_{\text{cnn}}+d_{\text{adapt.}})}$, O^l_{CNN} is concatenated with the adapter-processed $O^l_{ws,mhsa}$. The detailed process is as follows:

$$O^l_{\text{FiDo}} = \text{Concat}(O^l_{\text{CNN}}, \text{Adapter}(O^l_{\text{ws,mhsa}})). \quad (4)$$

O^r_{FiDo} for the right channel is obtained through the same steps. O^l_{FiDo} and O^r_{FiDo} are then used as input to the assessment modules, where we adopt the original implementation of MBI-Net+ [19]. Specifically, BLSTM+AT is applied to obtain frame-level intelligibility and HASPI scores and the embedding representation input to the class module (HA class). The frame-level intelligibility and HASPI scores are processed by their corresponding task-specific modules to predict the final intelligibility and HASPI scores, respectively. The loss function for training MBI-Net+ with FiDo is defined as follows:

$$O = \gamma_1\mathcal{L}_{Int} + \gamma_2\mathcal{L}_{HASPI} + \gamma_3\mathcal{L}_{CE}, \quad (5)$$

where $\mathcal{L}_{CE}$ represents the cross-entropy loss between the estimated enhancement type of the class module and the reference enhancement type (the front-end processor in the HA system). The weights between the losses are represented by γ_1, γ_2, and γ_3. $\mathcal{L}_{Int}$ and $\mathcal{L}_{HASPI}$ represent the MSE losses between the corresponding ground-truth scores and the predicted scores.

3. Experiments

3.1. Experimental Setup

The Clarity Prediction Challenge (CPC) 2023 dataset [27] includes recordings from six talkers, ten enhancement methods (corresponding to ten hearing aid (HA) systems) from the 2022 Clarity Enhancement Challenge [17], and intelligibility ratings provided by 25 listeners. The dataset is categorized into three tracks: Track 1 contains 2,779 utterances, Track 2 has 2,796 utterances, and Track 3 consists of 2,772 utterances. For each track, 90% of the utterances were allocated for training, while the remaining 10% were used for development. The test sets for the three tracks comprise 305, 294, and 298 utterances, respectively, featuring unseen listeners and HA systems. The models were trained and evaluated separately on all three tracks, and the overall performance across them is referred to as "All." The evaluation relied on three key metrics: root mean square error (RMSE), linear correlation coefficient (LCC), and Spearman's rank correlation coefficient (SRCC) [28]. A lower RMSE signifies better alignment with ground-truth scores, while higher LCC and SRCC values indicate stronger correlations between predictions and actual scores. All models presented in this study were trained and assessed using the CPC 2023 dataset.

3.2. Comparison of different feature concatenation strategies

In the first experiment, we aim to compare the performance of MBI-Net+ using different feature concatenation strategies. This experiment intends to confirm the most appropriate concatenation method when employing cross-domain features.

Following the original implementation of MBI-Net+ [19], we adopt a structure consisting of a cross-domain feature extraction module, a four-layer convolutional block (with 16, 32, 64, and 128 channels, respectively), a one-layer Bidirectional Long Short-Term Memory (BLSTM) (with 128 nodes), and a fully connected layer (with 128 neurons). The model also incorporates an attention mechanism and a single neuron for frame-level scoring. Finally, two task-specific modules are leveraged

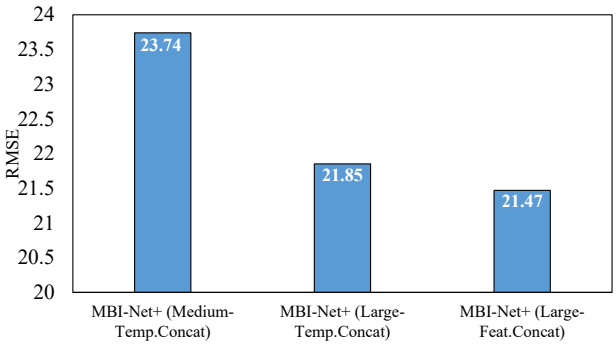


Figure 3: *Performance comparison of different feature concatenation strategies.*

for estimating subjective intelligibility and HASPI, along with a classification module trained to classify ten enhancement system types. The model is trained using the Adam [29] optimizer with a learning rate of 1e-4. We set the tunable weights to γ_1 = 1, γ_2 = 0.4, and γ_3 = 0.2, assigning higher weights to more important metrics and ensuring that the losses remain on a similar scale. Specifically, we deployed three versions of MBI-Net+: (1) MBI-Net+ with the Whisper Medium model using temporal concatenation, (2) MBI-Net+ with the Whisper Large model using temporal concatenation, and (3) MBI-Net+ with the Whisper Large model using feature-based concatenation. For temporal concatenation, the frames of PS, FB, and WS are combined, resulting in a frame size equal to the sum of their frame numbers.

In this experiment, we specifically focus on Track 3, which has a smaller training set, to evaluate the feature concatenation strategy. As shown in Fig. 3, we first notice that by simply replacing Whisper Medium (the original implementation of MBI-Net+ [19]) with Whisper Large, we can achieve a relative RMSE reduction of 7.96% (from 23.74 to 21.85). This confirms that feature extraction plays an important role in improving prediction performance. Furthermore, replacing temporal concatenation with feature-based concatenation further improves the RMSE score. Therefore, this confirms that feature-based concatenation provides an effective strategy while also avoiding the lengthy number of frames present in temporal concatenation.

3.3. Comparing MBI-Net+ with FiDo and MBI-Net+

In the second experiment, we investigate the impact of FiDo on the performance of MBI-Net+ across different tracks. FiDo employs eight attention heads for each input feature: PS, FB, and WS. After applying attention, PS and FB features are concatenated. This concatenated output is then processed by a four-layer convolutional block with 16, 32, 64, and 128 channels, respectively. For the MHSA output of WS features, an adapter layer is introduced to perform task-specific adaptation and reduce the dimensionality to align with the output of the convolutional block. Finally, the outputs of the convolutional block and the adapter layer are concatenated and fed into the BLSTM attention layer and the corresponding assessment module, following the original MBI-Net+ configuration [19].

As shown in Table 1, for Track 1, FiDo achieves an LCC of 0.754 and an SRCC of 0.728, reducing the RMSE from 28.370 to 26.710. In Track 2, the improvements are more pronounced, with LCC increasing from 0.754 to 0.791 and SRCC improv-

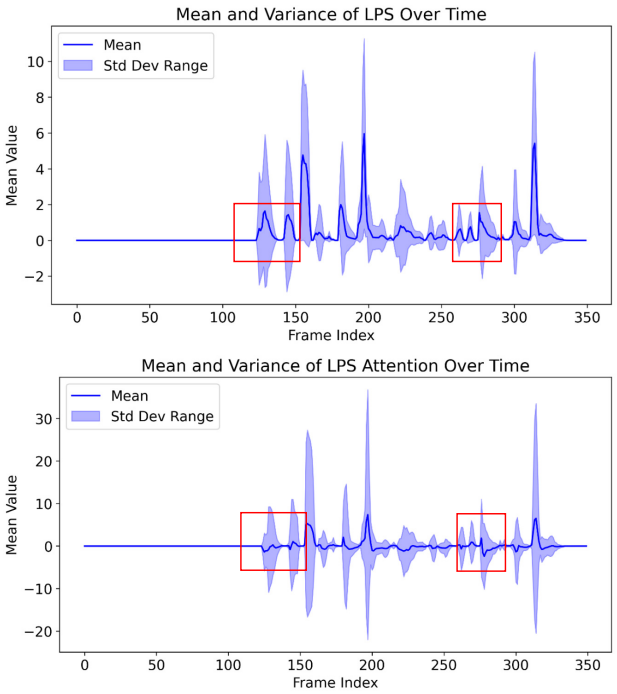


Figure 4: *Mean and standard deviation analysis between PS features before and after performing the attention mechanism.*

ing from 0.737 to 0.787, while RMSE decreases from 25.920 to 23.850. In Track 3, applying FiDo boosts the LCC from 0.813 to 0.850, improves the SRCC from 0.814 to 0.861, and reduces the RMSE from 23.740 to 21.420. Overall, across all tracks, FiDo improves LCC from 0.764 to 0.796, SRCC from 0.767 to 0.793, and reduces RMSE from 26.100 to 24.110. These results confirm that FiDo's early attention mechanism and feature-based concatenation strategy help improve the model's prediction performance.

In the next experiment, we visualize the effect of attention by comparing the mean and standard deviation of PS features before and after applying the attention mechanism. As shown in Fig. 4, the attention mechanism changes the shape of the feature, which further confirms that the attention mechanism indeed emphasizes specific regions of the audio representation.

3.4. Comparison with Other Prediction Models for HA

In the last experiment, we further compare MBI-Net+ with FiDo against other state-of-the-art systems on the Clarity Prediction Challenge 2023 dataset. The results are shown in Table 2. The first observation is that models employing large-scale pre-trained models (E011 [15], E002 [16], E025 [27], and MBI-Net+ [23]) generally outperform models that rely on conventional signal processing techniques (E003, E024, E015, and E020). Interestingly, when compared to intrusive-based systems, three models, including MBI-Net+ with FiDo, which employs Whisper as the feature extraction module, achieve lower RMSE values than the intrusive models. Furthermore, MBI-Net+ with FiDo achieves a relative RMSE reduction of 3.98% (from 25.1 to 24.1) compared to the best-performing model E011 [15] in the 2023 Clarity Prediction Challenge, which adopts a more complex and deeper model architecture. This further confirms that by extracting acoustic features more effectively and focusing on important regions at an early stage, the model can achieve better prediction performance.

Table 1: *LCC, SRCC, and RMSE results of MBI-Net+ and MBI-Net+ with FiDo on the complete test set.*

Model	LCC	SRCC	RMSE
Track 1			
MBI-Net+ [19]	0.721	0.714	28.370
MBI-Net+ with FiDo	**0.754**	**0.728**	**26.710**
Track 2			
MBI-Net+ [19]	0.754	0.737	25.920
MBI-Net+ with FiDo	**0.791**	**0.787**	**23.850**
Track 3			
MBI-Net+ [19]	0.813	0.814	23.740
MBI-Net+ with FiDo	**0.850**	**0.861**	**21.420**
Track All			
MBI-Net+ [19]	0.764	0.767	26.100
MBI-Net+ with FiDo	**0.796**	**0.793**	**24.110**

Table 2: *RMSE and LCC results of different systems (Intrusive and Non-Intrusive) from the First and Second Clarity Challenges on the test set.*

System	Non-Intrusive	RMSE	LCC
MBI-Net+ with FiDo	✓	**24.1**	**0.80**
E011 [15]	✓	25.1	0.78
E002 [16]	✓	25.3	0.77
E009 [30]	✗	25.4	0.78
E022 [31]	✗	25.7	0.77
MBI-Net+ [19]	✓	26.1	0.76
E025 [31]	✓	27.9	0.72
Baseline [27]	✗	28.7	0.70
E003 [32]	✓	31.1	0.64
E024 [32]	✓	31.7	0.62
E015 [33]	✓	35.0	0.60
E020 [34]	✓	39.8	0.33
Prior [3]	✗	40	-

4. Conclusion

In this paper, we have proposed Feature Importance across Domains (FiDo), which integrates multi-head self-attention (MHSA) and feature-based concatenation to capture the importance of different acoustic features across multiple domains. Experimental results confirm that incorporating FiDo into MBI-Net+ improves overall prediction performance, achieving a substantial RMSE reduction of 7.62% over the original MBI-Net+ model. Furthermore, MBI-Net+ with FiDo achieves a 3.98% RMSE reduction over the top system in the 2023 Clarity Prediction Challenge. These results validate the effectiveness of FiDo in enhancing speech intelligibility assessment in hearing aids and highlight the importance of capturing more effective acoustic features. In future work, we will focus on further refining FiDo's feature extraction process by exploring advanced attention mechanisms and feature fusion strategies.

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

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

[7] 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.

[8] 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.

[9] 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.

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

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

[12] 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.

[13] R. E. Zezario, Y.-W. Chen, S.-W. Fu, Y. Tsao, H.-M. Wang, and C.-S. Fuh, "A study on incorporating Whisper for robust speech assessment," in *Proc. ICME*, 2024.

[14] 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.

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

[16] 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.

[17] 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.

[18] J. Barker, M. Akeroyd, T. J. Cox, J. F. Culling, J. Firth, S. Graetzer, H. Griffiths, L. Harris, G. Naylor, Z. Podwinska, E. Porter, and R. V. Munoz, "The 1st clarity prediction challenge: A machine learning challenge for hearing aid intelligibility prediction," in *Proc. INTERSPEECH*, 2022.

[19] R. E. Zezario, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Non-intrusive speech intelligibility prediction for hearing aids using whisper and metadata," in *Proc. INTERSPEECH*, 2024, pp. 3844–3848.

[20] 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," *J. Acoust. Soc. Am.*, vol. 195, p. 3574–3587, 2024.

[21] 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.

[22] 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.

[23] R. E. 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.

[24] K. H. Lee, C. Park, J. Oh, and N. Kwak, " LFI-CAM: Learning Feature Importance for Better Visual Explanation ," in *Proc. IEEE/CVF ICCV*, 2021, pp. 1335–1343.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[26] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," in *Proc. SLT*, no. 1021-1028, 2018.

[27] 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.

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

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

[30] M. Huckvale and G. Hilkhuysen, "Combining acoustic, phonetic, linguistic and audiometric data in an intrusive intelligibility metric for hearing-impaired listeners," in *Proc. ISCA Clarity 2023*, 2023.

[31] Z. Tu, N. Ma, and J. Barker, "Intelligibility prediction with a pretrained noise-robust automatic speech recognition model," in *Proc. ISCA Clarity 2023*, 2023.

[32] C. O. Mawalim, X. Zhou, S. Okada, and M. Unoki, "A nonintrusive speech intelligibility prediction using binaural cues and time-series model with one-hot listener embedding," in *Proc. ISCA Clarity 2023*, 2023.

[33] K. Yamamoto, "A non-intrusive binaural speech intelligibility prediction for clarity-2023," in *Proc. ISCA Clarity 2023*, 2023.

[34] N. Mamun, S. Ahmed, and J. H. L. Hansen, "Prediction of behavioral speech intelligibility using a computational model of the auditory system," in *Proc. ISCA Clarity 2023*, 2023.