Non-Intrusive Intelligibility Prediction for Hearing Aids: Recent Advances, Trends, and Challenges

Ryandhimas E. Zezario
Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
E-mail: ryandhimas@citi.sinica.edu.tw

***Abstract*—This paper provides an overview of recent progress in non-intrusive speech intelligibility prediction for hearing aids (HA). We summarize developments in robust acoustic feature extraction, hearing loss modeling, and the use of emerging architectures for long-sequence processing. Listener-specific adaptation strategies and domain generalization approaches that aim to improve robustness in unseen acoustic environments are also discussed. Remaining challenges—such as the need for large-scale, diverse datasets and reliable cross-profile generalization—are acknowledged. Our goal is to offer a perspective on current trends, ongoing challenges, and possible future directions toward practical and reliable HA-oriented intelligibility prediction systems.**

I. Introduction

Speech intelligibility has long been a key metric for evaluating hearing aid (HA) performance, aiming to estimate how well audio signals can be understood by listeners. While human-based evaluations are considered the gold standard due to their reliability, they require a sufficient number of listeners to obtain unbiased evaluation scores. Conventional speech intelligibility methods often rely on signal processing and psychoacoustic models to approximate human auditory perception. Prominent examples include the Speech Intelligibility Index (SII) [1], Extended SII (ESII) [2], Speech Transmission Index (STI) [3], Short-Time Objective Intelligibility (STOI) [4], Modified Binaural STOI (MBSTOI) [5], and the Hearing Aid Speech Perception Index (HASPI) [6]. Although these methods have demonstrated notable performance, they depend on the availability of clean reference signals for more accurate estimation, limiting their applicability in real-world scenarios where such references are often unavailable.

With the advancement of deep learning, there has been growing interest in applying neural network models for non-intrusive speech intelligibility assessment [7]–[11]. Depending on the type of ground-truth labels employed, deep learning-based non-intrusive intelligibility prediction methods can be categorized into objective-based and subjective-based approaches. In objective-based methods, models are trained to predict scores from objective intelligibility metrics such as STOI or HASPI. In contrast, subjective-based methods use human listening test results as target labels. Early efforts in this field primarily focus on predicting intelligibility for normal-hearing listeners. For example, STOI-Net [8] uses a convolutional neural network (CNN) and a bidirectional long short-term memory (BLSTM) architecture to estimate STOI scores from speech features in a non-intrusive manner. In the subjective-based domain, one of the earliest models [7] employs CNN to predict intelligibility scores collected from human listeners. More recently, multi-target models such as MTI-Net [10] combine CNN-BLSTM architectures with diverse acoustic features—including spectral features, raw waveforms, and self-supervised representations—to jointly predict multiple intelligibility metrics, including human ratings, STOI, and word error rate (WER), using a multi-task learning framework.

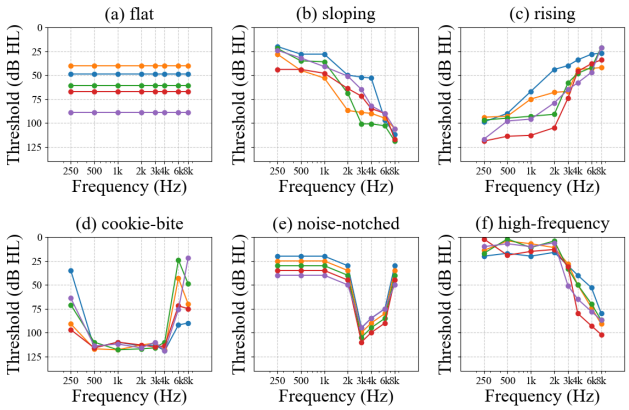

Fig. 1. Examples of Hearing Loss Audiogram Profiles.

Building on the promising results achieved in normal-hearing intelligibility prediction, recent studies [12]–[21] have begun to explore applications in HA scenarios, where accurate intelligibility assessment becomes even more critical. Unlike normal-hearing conditions, intelligibility for hearing-impaired (HI) listeners is influenced by a wide range of factors, including the severity of hearing loss, the characteristics of the HA processing pipeline (e.g., noise reduction, beamforming, dynamic range compression), and environmental conditions. This added complexity makes intelligibility assessment more challenging but also more clinically relevant. As such, there is a growing need for models that can generalize to diverse listening profiles and HA systems. In this survey, we provide a comprehensive overview of recent advances, highlight emerging trends in model design and evaluation, and discuss key challenges in developing reliable non-intrusive intelligibility prediction models for HA.

979-8-3315-7206-8/25/$31.00

TABLE I
AUDIOGRAM THRESHOLDS (IN DECIBELS HEARING LEVEL(DB HL).)

Frequency (Hz)	250	500	1000	2000	4000	8000
Right Ear (dB HL)	10	15	20	25	30	35
Left Ear (dB HL)	15	20	25	30	35	40

II. RECENT ADVANCES AND TRENDS

A. Listener Audiograms and Acoustic Features

One of the most important aspects in deploying non-intrusive speech intelligibility prediction for HA is the availability and integration of audiogram information. Audiograms provide essential insights into an individual's hearing thresholds across different frequencies, as shown in Table I, enabling models to personalize intelligibility predictions based on the user's unique hearing profile, as further illustrated in Fig. 1. This information helps ensure that the system accounts for frequency-specific hearing deficits, which is essential for achieving more accurate speech intelligibility estimation.

The use of listener audiograms can generally be categorized into two distinct strategies. The first is direct input embedding, in which the listener's audiogram is concatenated with the corresponding acoustic features as model input [12], [18], [19], [30], [31]. The second strategy involves incorporating the audiogram into hearing loss simulation models, such as the Moore, Stone, Baer, and Glasberg (MSBG) model [32]–[35] or other ear models [36]. Several models adopt this simulation-based approach to simulate the perceptual effects of hearing loss and generate the corresponding audio representations [15]–[17], [21], [23], [24], [27], [29].

Furthermore, in addition to audiogram information, acoustic features play an important role in accurate intelligibility prediction, given that most datasets [13], [14] consist of challenge scenarios that include unseen speakers, environments, and listener hearing profiles. While handcrafted features such as Mel-Frequency Cepstral Coefficients (MFCC), spectral features, and extended Geneva minimalistic acoustic parameter set (eGeMAPS) have been shown to be effective features [12], [21]–[23], models that incorporate large speech pre-trained models have shown higher prediction performance. These models could either use automatic speech recognition, self-supervised learning (SSL) representations (e.g., HuBERT [37], WavLM [38]), or weakly supervised models such as Whisper [39]. These models are capable of extracting high-level acoustic and linguistic information that generalizes well across conditions, due to the advantage of being trained on large-scale datasets, thereby improving robustness and intelligibility prediction accuracy. In some scenarios, WavLM tends to achieve higher prediction than HuBERT [16], while in more general conditions, Whisper tends to provide richer acoustic features and achieve higher prediction performance [19], [20], [26], [29]. As a result, the combination of personalized hearing profiles (via audiograms) and suitable acoustic features (via SSL or Whisper) has emerged as a key trend in the development of higher-performing and generalizable models.

B. Model Architecture

Along with richer acoustic features, the choice of a suitable model architecture has also been shown to be a crucial aspect in the development of non-intrusive intelligibility prediction systems. Given the sequential and temporal nature of audio data, most existing methods [12], [16], [18], [20], [25], [26], [29] adopt a bidirectional long short-term memory (BLSTM) framework as the backbone of the model. The BLSTM architecture is particularly well-suited for modeling long-range dependencies in audio signals, enabling the system to capture both past and future context, which is essential for accurately predicting speech intelligibility.

To further enhance the modeling capacity of BLSTM-based frameworks, several works have proposed the integration of attention mechanisms [12], [18], [25]–[27]. These mechanisms allow the model to focus on more important regions of the input sequence, thereby improving its ability to respond to important acoustic or linguistic cues. Another line of research introduces CNN into the BLSTM architecture [16], [20], [29], as front-end feature extractors. The use of CNN helps capture local spectro-temporal patterns that strengthen the sequence modeling performed by BLSTM.

Beyond the BLSTM-based approaches, several alternative architectures have been explored to improve performance or reduce model complexity. For instance, time-delay neural networks (TDNN) [22] have been investigated for their ability to model temporal context with fewer parameters and lower computational cost. CNN-based architectures [17], [21], [23] have also been proposed, leveraging deep hierarchical feature representations. More recently, transformer-based models [15], [19], [30] have gained attention due to their ability to capture global dependencies and their success in other speech-related tasks. Furthermore, the emerging MAMBA architecture [28], designed to handle long sequences, presents an alternative module for non-intrusive speech intelligibility prediction in HA.

C. Objective Functions

To optimally train the model, it is crucial to define an appropriate objective function that effectively guides the adjustment of model parameters. A commonly adopted approach is direct estimation, which minimizes the mean square error (MSE) between the model's predicted scores and the corresponding human-labeled intelligibility scores [17], [21], as defined in the following equation:

$$\mathcal{L}_{\text{utterance}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{1}$$

where y_i and $\hat{y}_i$ denote the reference and predicted intelligibility scores for the i-th utterance.

In addition to utterance-level loss, several studies have explored the benefits of incorporating frame-level objectives as auxiliary losses. This approach aims to capture more fine-grained temporal patterns during training, which may improve

TABLE II
COMPARISON OF NON-INTRUSIVE INTELLIGIBILITY PREDICTION METHODS WITH AUDIOGRAM AND ACOUSTIC FEATURE INTEGRATION.

Model / Paper (Year)	Acoustic Input	Audiogram / HL Use	Architecture	Loss / Objective
Chiang et al. (2021) [12]	Spectral features + audiogram (concatenated)	Direct input embedding	BLSTM + Attention	Utterance-level MSE + Frame-level MSE
Close et al. (2022) [17]	Spectral features	MSBG simulation	CNN	Utterance-level MSE
Tu et al. (2022) [15]	ASR-based intermediate features	MSBG simulation	Transformer	CTC loss
Zezario et al. (2022) [16]	WavLM or HuBERT	MSBG simulation	CNN + BLSTM + Attention	Utterance-level MSE + Frame-level MSE
Robach et al. (2022) [22]	MFCC	None	TDNN	Lattice-free maximum mutual information (LF-MMI) + Cross Entropy
Titalim et al. (2022) [23]	WavLM+ MFCC + eGeMAPS	EarModel simulation	CNN + WaveNet + Meta regressor + Stack regressor	Utterance-level MSE
Chiang et al. (2023) [18]	WavLM + Audiogram (concatenated)	Direct input embedding	BLSTM + Attention	Utterance-level MSE + Frame-level MSE
Mawalim et al. (2023) [21]	WavLM+ MFCC + eGeMAPS	EarModel simulation	CNN	Utterance-level MSE
Mawalim et al. (2023) [24]	WavLM+ MFCC + eGeMAPS	EarModel simulation	CNN + Stack regressor	Utterance-level MSE
Close et al. (2023) [25]	SSL	None	BLSTM + Attention	Utterance-level MSE
Mogridge et al. (2024) [26]	Whisper + Whisper exemplar	None	BLSTM + Attention	Utterance-level MSE
Cuervo et al. (2024) [19]	Whisper + Audiogram (concatenated)	Direct input embedding	Transformer	Huber loss
Zezario et al. (2024) [20]	Spectral + Waveform + Whisper	MSBG simulation	CNN + BLSTM + Attention	Utterance-level MSE + Frame-level MSE + Cross Entropy
Zhou et al. (2025) [27]	WavLM	MSBG simulation	LSTM + LightGBM	Utterance-level MSE
Yamamoto et al. (2025) [28]	Whisper + audiogram	Direct input embedding	MAMBA	Huber loss
Zezario et al. (2025) [29]	Spectral + Waveform + Whisper	MSBG simulation	(Early Attention) CNN + BLSTM + Attention	Utterance-level MSE + Frame-level MSE + Cross Entropy
Zhou et al. (2025) [30]	Whisper + Score	None	Transformer	Huber loss

generalization and stabilize learning [12], [16], [18], [20], [29].

$$\mathcal{L}_{\text{frame}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i}\left(y_{i,t}-\hat{y}_{i,t}\right)^2 \tag{2}$$

where T_i is the number of frames in the i-th utterance. In general, prior works [12], [16], [18], [20], [29] combine $\mathcal{L}_{\text{utterance}}$ and $\mathcal{L}_{\text{frame}}$ to form the overall training loss.

Some studies [20], [29] further adopt a multi-task learning strategy, combining losses from different metrics such as HASPI, along with a cross-entropy loss to classify different HA systems, as follows:

$$\mathcal{L}_{\text{multi-task}} = \alpha\cdot\mathcal{L}_{\text{intell}} + (1-\alpha)\cdot\mathcal{L}_{\text{HASPI}} + \mathcal{L}_{\text{CE}} \tag{3}$$

where $\alpha \in [0,1]$ controls the task weighting. Each main loss ($\mathcal{L}_{\text{intell}}$ and $\mathcal{L}_{\text{HASPI}}$) is defined as a weighted combination of utterance-level and frame-level losses:

$$\mathcal{L}_{\text{intell}} = \beta\cdot\mathcal{L}_{\text{intell,utt}} + (1-\beta)\cdot\mathcal{L}_{\text{intell,frame}} \tag{4}$$

$$\mathcal{L}_{\text{HASPI}} = \beta\cdot\mathcal{L}_{\text{HASPI,utt}} + (1-\beta)\cdot\mathcal{L}_{\text{HASPI,frame}} \tag{5}$$

The individual loss terms are defined as:

$$\mathcal{L}_{\text{intell,utt}} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i^{\text{intell}}-\hat{y}_i^{\text{intell}}\right)^2 \tag{6}$$

$$\mathcal{L}_{\text{intell,frame}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i}\left(y_{i,t}^{\text{intell}}-\hat{y}_{i,t}^{\text{intell}}\right)^2 \tag{7}$$

$$\mathcal{L}_{\text{HASPI,utt}} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i^{\text{HASPI}}-\hat{y}_i^{\text{HASPI}}\right)^2 \tag{8}$$

$$\mathcal{L}_{\text{HASPI,frame}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i}\left(y_{i,t}^{\text{HASPI}}-\hat{y}_{i,t}^{\text{HASPI}}\right)^2 \tag{9}$$

Moreover, alternative loss formulations such as the Huber[40] loss have also been investigated for their robustness to outliers and their balanced treatment between L1 and L2 behaviors [19], [28], [30]:

$$\mathcal{L}_{\text{Huber}} = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}\frac{1}{2}\left(y_i-\hat{y}_i\right)^2 & \text{if } |\hat{y}_i-y_i|\le\delta\\ \delta\left(|y_i-\hat{y}_i|-\frac{1}{2}\delta\right) & \text{otherwise}\end{cases} \tag{10}$$

where δ is a threshold parameter.

In a slightly different scenario, where ASR-based systems are used to extract confidence measures, the Connectionist Temporal Classification (CTC) loss is often selected as the training objective.

D. Dataset

Due to the data-driven nature of non-intrusive speech intelligibility prediction methods, the availability of suitable datasets is an important factor in system development. The Clarity Prediction Challenge (CPC) series—CPC1 [13] and CPC2 [14]—plays an important role in advancing this field, with most existing deep learning-based approaches for HA assessment relying on these datasets for training and evaluation.

CPC1 (2022) [13] provides data from ten HA systems used in the 2021 Clarity Enhancement Challenge [41]. Listening tests involve 25 HA users, each tasked with repeating what they hear from presented speech samples. Scores range from 0 to 100, with higher scores indicating better intelligibility. Bilateral pure-tone audiograms are estimated for each listener using hearing thresholds at [250, 500, 1000, 2000, 3000, 4000, 6000, 8000] Hz. The dataset is organized into two tracks: Track 1 with 4,863 training utterances and Track 2 with 3,580 training utterances, alongside corresponding test sets of 2,421 and 632 utterances, respectively, with no overlap between training and test material.

Building upon CPC1, CPC2 (2023) [14] expands the dataset to include recordings from six talkers processed by ten HA systems originating from the 2022 Clarity Enhancement Challenge [42]. Intelligibility ratings from 25 listeners are collected for three tracks: Track 1 (2,779 utterances), Track 2 (2,796 utterances), and Track 3 (2,772 utterances). The test sets—containing 305, 294, and 298 utterances for Tracks 1–3, respectively—feature unseen listeners and HA systems. Evaluation relies on root mean square error (RMSE), linear correlation coefficient (LCC), and Spearman's rank correlation coefficient (SRCC) [43], where lower RMSE and higher LCC/SRCC indicate better performance.

III. Key Challenges

Based on recent advancements, several key challenges remain in the development of non-intrusive intelligibility prediction models for HA users. One of the most important challenges is the effective integration of a listener's hearing loss audiogram. While many models incorporate audiogram data by directly concatenating it with acoustic features, this naïve embedding approach may not fully capture the complex perceptual information of hearing impairment. On the other hand, integrated simulation-based hearing loss models such as MSBG and EarModel may introduce additional complexity, can often be non-differentiable, and may limit the efficiency of end-to-end learning. Furthermore, in scenarios where the systems are tested on new listener data, the systems may struggle to generalize well if they were previously trained on specific listener audiograms. As a result, recent developments mainly focus on improving generalization under more adverse conditions. With the progression from CPC 1 [13] and CPC2 [14] to the current CPC3 dataset, the emphasis has shifted toward using general severity levels as substitutes for listener-specific audiograms. This highlights a clear trend toward developing more generalizable systems capable of operating effectively under diverse and challenging conditions.

Another fundamental problem is ensuring robustness and generalizability across diverse acoustic environments. Many existing systems perform well on seen datasets but degrade significantly in real-world scenarios involving unseen speakers, noise types, room acoustics, and hearing-aid settings. Additionally, the selection of acoustic features also plays a pivotal role in model performance. Earlier approaches primarily relied on handcrafted features such as Mel-Frequency Cepstral Coefficients (MFCC) or other spectral features, which are computationally efficient but limited in modeling long-range temporal patterns and higher-order contextual dependencies. More recent works have incorporated self-supervised learning (SSL) models such as HuBERT, WavLM, and Whisper, which provide rich and hierarchical representations learned from large-scale speech corpora. Although these features improve prediction accuracy, they also make the model more computationally intensive. Therefore, developing appropriate strategies to reduce computational cost while maintaining generalization remains an important step for better system deployment.

Lastly, there is a need to design objective functions that better align with human perception. While most models optimize for mean squared error (MSE) between predicted and ground-truth intelligibility scores, this metric may not adequately capture perceptual information. Recent efforts have explored auxiliary frame-level losses or robust alternatives like the Huber loss, but perceptually inspired or listener-specific loss functions remain limited.

In summary, future progress in non-intrusive intelligibility prediction will depend on improving hearing loss modeling in ways that are both perceptually meaningful and computationally feasible, enhancing robustness to unseen conditions, leveraging rich yet efficient acoustic representations, and aligning training objectives more closely with human-perceived intelligibility.

IV. Conclusion and Discussions

In this paper, we reviewed recent developments in deep learning-based non-intrusive intelligibility prediction for HA. A variety of input features have been explored, ranging from traditional acoustic features like MFCC to self-supervised embeddings such as Whisper and WavLM. At the same time, hearing loss modeling has been further explored from simple audiogram concatenation to more perceptually accurate simulations using models like MSBG. Architecturally, researchers have transitioned from CNNs and BLSTMs to more advanced structures, including attention mechanisms, transformers, and memory-based designs. Despite these improvements, most models are still trained with loss functions such as MSE or Huber loss, which may not fully capture the perceptual nature of intelligibility as rated by hearing-impaired listeners.

Looking forward, future development of non-intrusive intelligibility prediction models for HA may benefit from incorporating hearing loss modeling approaches that balance perceptual relevance with computational efficiency. This may involve the design of learnable and differentiable auditory

models, or hybrid strategies that combine severity-based representations with listener-specific adaptation to improve generalization under mismatched audiogram conditions. Enhancing robustness to unseen acoustic scenarios will also be important, potentially achieved through domain generalization techniques, data augmentation, and model architectures that can adapt to varying noise and reverberation conditions. Furthermore, balancing prediction accuracy with computational efficiency emphasizes the value of lightweight feature representations that preserve the contextual richness of SSL features while supporting real-time deployment. Finally, exploring perceptually guided and listener-aware loss functions, possibly within multi-task learning frameworks, may further align model optimization with human-perceived intelligibility and enhance practical applicability in HA scenarios.

REFERENCES

[1] ANSI Std. S3.5 1997, "Methods for calculation of the speech intelligibility index," in *Acoustical Society of America*, 1997.

[2] T. Houtgast and H. J. M. Steeneken, "Evaluation of speech transmission channels by using artificial signals," *Acustica*, vol. 25, no. 6, pp. 355–367, 1971.

[3] 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.

[4] 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.

[5] 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.

[6] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI) version 2," *Speech Communication*, vol. 131, pp. 35–46, 2021.

[7] A. H. Andersen, J. M. D. Haan, Z.-H. Tan, and J. Jensen, "Nonintrusive speech intelligibility prediction using convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1925–1939, 2018.

[8] R. E. Zezario, S.-W. Fu, C.-S. Fuh, Y. Tsao, and H.-M. Wang, "STOI-Net: A deep learning-based non-intrusive speech intelligibility assessment model," in *Proc. APSIPA ASC*, 2020, pp. 482–486.

[9] R. E. 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.

[10] R. E. 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.

[11] Y.-W. Chen and Y. Tsao, "InQSS: a speech intelligibility and quality assessment model using a multi-task learning network," in *Proc. INTERSPEECH*, 2022, pp. 3088–3092.

[12] H.-T. Chiang, Y.-C. Wu, C. Yu, *et al.*, "HASA-Net: A non-intrusive hearing-aid speech assessment network," in *Proc. ASRU*, 2021, pp. 907–913.

[13] J. Barker, M. Akeroyd, T. J. Cox, *et al.*, "The 1st Clarity prediction challenge: A machine learning challenge for hearing aid intelligibility prediction," in *Proc. INTERSPEECH*, 2022, pp. 3508–3512.

[14] J. P. Barker, M. A. Akeroyd, W. Bailey, *et al.*, "The 2nd Clarity prediction challenge: A machine learning challenge for hearing aid intelligibility prediction," in *Proc. ICASSP*, 2024, pp. 11 551–11 555.

[15] 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.

[16] 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.

[17] G. Close and S. Hollands and S. Goetze and T. Hain, "Non-intrusive speech intelligibility metric prediction for hearing impaired individuals," in *Proc. INTERSPEECH*, 2022, pp. 3483–3487.

[18] 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, pp. 3574–3587, 2024.

[19] S. Cuervo and R. Marxer, "Speech foundation models on intelligibility prediction for hearing-impaired listeners," in *Proc. ICASSP*, 2024, pp. 1421–1425.

[20] 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.

[21] 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. 109 663, 2023.

[22] J. Roßbach and R. Huber and S. Röttges and C. F. Hauth and T. Biberger and T. Brand and B. T. Meyer and J. Rennies, "Speech intelligibility prediction for hearing-impaired listeners with the LEAP model," in *Proc. INTERSPEECH*, 2022, pp. 3498–3502.

[23] B. A. Titalim, C. O. Mawalim, S. Okada, and M. Unoki, "Speech intelligibility prediction for hearing aids using an auditory model and acoustic parameters," in *Proc. APSIPA ASC*, 2022, pp. 1076–1084.

[24] C. O. Mawalim, B. A. Titalim, S. Okada, and M. Unoki, "Auditory model optimization with wavegram-cnn and acoustic parameter models for nonintrusive speech intelligibility prediction in hearing aids," in *Proc. EUSIPCO*, 2023, pp. 211–215.

[25] G. Close, T. Hain, and S. Goetze, "Non intrusive intelligibility predictor for hearing impaired individuals using self supervised speech representations," in *Proc. ASRU 2023 SPARKS workshop*, 2023.

[26] R. Mogridge, G. Close, R. Sutherland, *et al.*, "Non-intrusive speech intelligibility prediction for hearing-impaired users using intermediate ASR features and human memory models," in *Proc. ICASSP*, 2024, pp. 306–310.

[27] X. Zhou, C. Olivia Mawalim, and M. Unoki, "Speech intelligibility prediction using binaural processing for hearing loss," *IEEE Access*, vol. 13, pp. 25 817–25 836, 2025.

[28] K. Yamamoto and K. Miyazaki, "Non-intrusive binaural speech intelligibility prediction using mamba for hearing-impaired listeners," in *Proc. INTERSPEECH*, 2025, pp. 5463–5467.

[29] R. E. Zezario and S. M. Siniscalchi and F. Chen and H.-M. Wang and Y. Tsao, "Feature importance across domains for improving non-intrusive speech intelligibility prediction in hearing aids," in *Proc. INTERSPEECH*, 2025, pp. 5473–5477.

[30] H. Zhou and C. Mo and B. Cao and L. Li and S. X. Wang, "No audiogram: Leveraging existing scores for personalized speech intelligibility prediction," in *Proc. INTERSPEECH*, 2025, pp. 5468–5472.

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

[32] B. C. J. Moore, B. R. Glasberg, and T. Baer, "A model for the prediction of thresholds, loudness, and partial loudness," *Journal of the Audio Engineering Society*, vol. 45, pp. 224–240, 1997.

[33] T. Baer and B. C. J. Moore, "Effects of spectral smearing on the intelligibility of sentences in noise," *The Journal of the Acoustical Society of America*, vol. 94, pp. 1229–1241, 1993.

[34] T. Baer and B. C. J. Moore, "Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech," *The Journal of the Acoustical Society of America*, vol. 95, no. 4, pp. 2050–2062, 1993.

[35] B. C. J. Moore and B. R. Glasberg, "Simulation of the effects of loudness recruitment and threshold elevation on the intelligibility of speech in quiet and in a background of speech," *The Journal of the Acoustical Society of America*, vol. 94, no. 4, pp. 2050–2062, 1993.

[36] J. Kates, "An auditory model for intelligibility and quality predictions," *The Journal of the Acoustical Society of America*, vol. 133, no. 5 Supplement, pp. 3560–3560, 2013.

[37] 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.

[38] S. Chen, C. Wang, Z. Chen, *et al.*, "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.

[39] 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.

[40] P. J. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.

[41] S. Graetzer, J. Barker, M. A. T. J. Cox, *et al.*, "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing," in *Proc. INTERSPEECH*, 2021, pp. 686–690.

[42] M. A. Akeroyd, W. Bailey, J. Barker, *et al.*, "The 2nd Clarity enhancement challenge for hearing aid speech intelligibility enhancement: Overview and outcomes," in *Proc. ICASSP*, 2023, pp. 1–5.

[43] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.