Auditory Model Optimization with Wavegram-CNN and Acoustic Parameter Models for Nonintrusive Speech Intelligibility Prediction in Hearing Aids

Candy Olivia Mawalim, Benita Angela Titalim, Shogo Okada, and Masashi Unoki
Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
Ishikawa, Japan
{candylim, s2110104, okada-s, unoki}@jaist.ac.jp

***Abstract*—Nonintrusive speech intelligibility (SI) prediction is essential for evaluating many speech technology applications, including hearing aid development. In this study, several factors related to hearing perception are investigated to predict SI. In the proposed method, we integrated a physiological auditory model from two ears (binaural EarModel), wavegram-CNN model and acoustic parameter model. The refined EarModel does not require clean speech as input (blind method). In EarModel, the perception caused by hearing loss is simulated based on audiograms. Meanwhile, the wavegram-CNN and acoustic parameter models represent the factors related to the speech spectrum and acoustics, respectively. The proposed method is evaluated based on the scenario from the 1st Clarity Prediction Challenge (CPC1). The results show that the proposed method outperforms the intrusive baseline MBSTOI and HASPI methods in terms of the Pearson coefficient (ρ), RMSE, and R^2 score in both closed-set and open-set tracks. Based on the results from listener-wise evaluation results, the average ρ could be improved by more than 0.3 using the proposed method.**

***Index Terms*—hearing aids, clarity challenge, speech intelligibility, nonintrusive method, auditory model**

I. Introduction

Speech intelligibility (SI) refers to how well someone understands speech or the percentage of words identified correctly [1], [2]. SI prediction is crucial in effective real-world communication to identify perceivable speech that is inseparable from background noise and reverberation. Speech may be less intelligible when it reaches human ears [3]. This phenomenon is worse for individuals with hearing loss [4], [5]. Consequently, hearing aids are beneficial to enhance SI for hearing-impaired listeners.

In the past few decades, several speech enhancement techniques have been proposed for hearing aids. However, until the development of objective SI prediction methods around 1950s, SI evaluation was limited and relied on subjective listening experiments [6]–[8]. For instance, the Hearing-Aid Speech Perception Index (HASPI) has been considered for evaluating SI for hearing aids [9]. A remarkable aspect of this model is that it includes an auditory model to simulate hearing loss in hearing-impaired listeners [10]. The main limitation of HASPI is primarily due to monaural signal processing and target speech that cannot be emphasized in a noisy environment.

On the other hand, a baseline model using modified binaural short-time objective intelligibility (MBSTOI) [11] and the Cambridge hearing loss model (MBSG model) [12] were proposed in the Clarity Challenge[1] [2] as an alternative to HASPI [13]. Since the Clarity Challenge focused on binaural processing, the aim of the baseline model for the prediction challenge is to find an algorithm that accurately predicts the speech intelligibility of binaural signals. Despite this advantage, the baseline model cannot be used to fix the delay after hearing aid processing, and using a correlation function in the measurement leads to signal-level insensitivity.

We propose a nonintrusive method to improve the training efficiency and the process for SI prediction and facilitate binaural processing based on the remaining issues with high consideration of auditory perception. Unlike our prior work [14] that utilizes directly the auditory model from HASPI [9], the proposed method enhances auditory model to simulate hearing loss perception without clean speech as input. Additionally, features related to the speech spectrum (wavLM [15] with the wavegram-CNN model) and acoustic parameters (eGeMAPS [16]) are also utilized. We hypothesize that our proposed method can provide higher SI prediction accuracy than the baseline MBSTOI because (1) we include several speech parameters, (2) we solve the delay that occurs after hearing aid processing, and (3) our auditory model is derived from HASPI, which has been used as a standard in hearing aid development.

II. Related Work

Speech intelligibility (SI) prediction with the consideration of hearing impaired condition is one of the modules in the Clarity Challenge. The Clarity Challenge[1] was initiated to allow the research community to contribute to solving problems in hearing aid processing by providing a general scenario, dataset, baseline system, and fundamental knowledge through open-source software, tutorials, etc. [13]. The task in Clarity Prediction Challenge 1 (CPC1) is to predict the SI in speech-in-noise (SPIN) perceived by listeners with a hearing aid system [2]. SI is defined as the percentage of words that are accurately identified in a given sentence (7-to-10-words

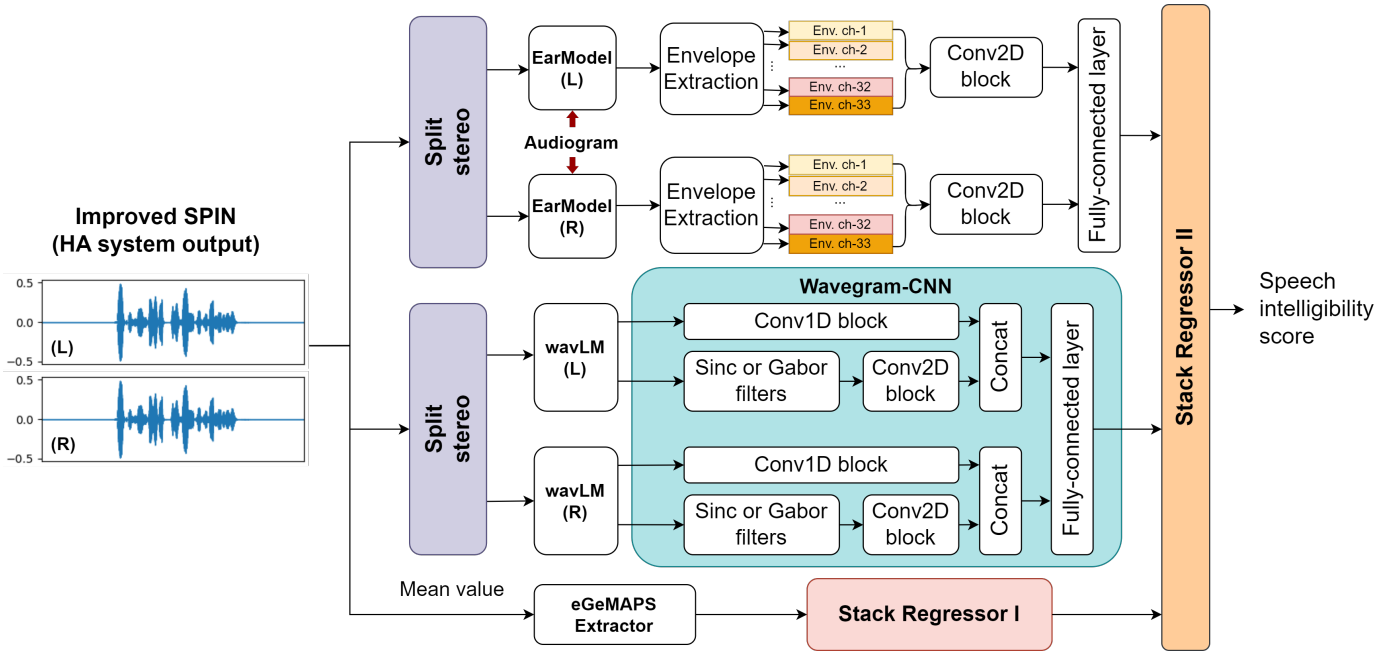

Fig. 1. Block diagram of the proposed method

long) [2]. For reference, a baseline system based on MBSTOI [11] was integrated with the MBSG model [12].

As a comparison, another SI index, HASPI, was proposed by Kates and Arehart [9]. This model is usually preferable in hearing aid development because it includes an auditory model [10] that can simulate both normal hearing and hearing loss perception. Hearing loss perception is simulated based on outer hair cell (OHC) [17], [18] and inner hair cell (IHC) loss [19]. In the auditory model, the temporal envelope is extracted in several channels, passed through the modulation analysis, and mapped using a neural network to calculate or predict the intelligibility index. The HASPI was trained on IEEE sentences and scored based on the proportion of entire sentences correct [9].

In prior work [14], we addressed the delay problem in a baseline system based on MBSTOI and supported binaural processing using the auditory model in HASPI (EarModel) [9]. In our prior work, it appears that the evaluation results have a higher correlation and smaller error compared to the baseline system and the model in HASPI. However, since the input of the original EarModel includes clean speech, this method is regarded as a nonblind (intrusive) method.

The requirement of clean speech as input is often unrealistic in a real-world environment. Moreover, in the machine learning model for the WavLM input in our prior work, less consideration was given to hearing perception. The main objective of this study is to propose a nonintrusive SI prediction method by considering several relevant factors in hearing perception, including the listener's hearing condition, the spectrum of speech, and the acoustic parameters.

[1]https://claritychallenge.org/

III. Proposed Method

Figure 1 shows the overall block diagram of our proposed method, which consists of three models: the auditory model, wavegram-CNN model, and acoustic parameter model. The outputs of all models are combined with a stack regressor to obtain the speech intelligibility (SI) score.

A. Auditory Model

An auditory model proposed in HASPI [9] is utilized in the proposed method to represent normal and impaired hearing perception. However, a modification is applied to the auditory model by excluding signal processing from the input clean signal. Thus, the SI prediction method with the modified auditory model, denoted as EarModel, can be regarded as a nonintrusive method.

The top part of Fig. 1 shows the process of the auditory model. There are two EarModels, which process the left and right improved SPIN signals. The first processing step in the EarModel involves resampling to 24 kHz and bandpass filtering, which resembles the middle ear filter. The filter limits the bandwidth of the improved SPIN signal to 350-5000 Hz. Next, the signal is passed through an auditory filter that was developed based on the relation between the degree of OHC loss and signal intensity at each audiometric frequency. The OHC loss is modeled in the gammatone filterbank with an increasing filter bandwidth within the 32 frequency channels. Based on the work reported by [19], the filter bandwidth change because of OHC loss BW^{HI} with respect to normal hearing BW^{NH} below 50 dB SPL can be approximated by

Eq. 1.

$$BW^{\text{HI}} = \left(1 + \left(\frac{\text{attnOHC}}{50}\right) + 2 \times \left(\frac{\text{attnOHC}}{50}\right)^{6}\right) BW^{\text{NH}} \tag{1}$$

where attnOHC is the OHC damage in dB that causes hearing loss, and the value 50 indicates the maximum attenuation in dB SPL. For signal intensities between 50 and 100 dB SPL, the control filterbank is used to determine the bandwidth changes based on linear interpolation.

In addition, the dynamic range compression is modeled based on the OHC function. Compression ratios of 1.25:1 to 3.5:1 in the frequency range of 80-8000 Hz are reduced due to OHC damage. The compression gain in the auditory model is expressed as follows.

$$G = -\text{attnOHC} - \left(1 - \frac{1}{CR}\right)\left(\theta_{\text{low}} - \hat{E}_c\right) \tag{2}$$

where

$$\hat{E}_c = \max\left(\theta_{\text{low}}, \left(\min\left(\hat{E}_c, \theta_{\text{high}}\right)\right)\right). \tag{3}$$

where $\hat{E}_c$, θ_{low}, θ_{high}, and CR are the control signal envelope in dB, a lower threshold equal to $(\text{attnOHC} + 30)$, the highest threshold equal to 100 dB, and the compression ratio, respectively [10], [20].

Due to the IHC damage, the additional attenuation contributes to auditory filterbank changes [19]. Moreover, after the signal decomposition in the filterbank, the IHC loss is determined by using the short-term IHC firing adaptation function based on the equivalent RC circuit model [21], [22]. Then, the differential equations are transformed using 1st-order backward differences in the digital domain. After the loss parameters are applied to the input signal, our proposed method utilized the short-term temporal envelopes (20 ms window with a stride of 10 ms) of the left and right signals from the 32 channels. These temporal envelopes were fed into a Conv2D block (three two-dimensional (2D) CNN layers with a fully connected layer) to predict SI.

The Conv2D block was chosen over a simple CNN encoder in previous work [14] because the Conv2D block allows for capturing spatial information. The temporal envelope extracted from EarModel contains temporal dynamics but lacks spatial information. Thus, Conv2D, which operates on a 2D grid, is used to leverage the convolutional layers and pick up spatial patterns or structures within the data. Also, it can perform hierarchical learning, which can capture low-level and high-level features from the envelope extracted, enabling the model to understand fine-grained details that contribute to speech intelligibility. Lastly, due to the computational complexity of the proposed method, the convolutional layer in the Conv2D with a fully-connected layer reduces the number of parameters, making the model more computationally efficient.

B. Wavegram-CNN Model

In prior research from CPC1 participants, it was shown that features derived from automatic speech recognition (ASR) significantly improved the prediction accuracy of SI [23]–[25]. In this study, we utilized a better downstream ASR model, namely, the pretrained self-supervised learning (SSL) model, specifically the wavLM model [15]. Instead of using the one-dimensional (1D) CNN structure from our prior work [14], we developed a wavegram-CNN model to learn the wavLM feature.

The wavegram architecture learns the time-frequency representation using a 2D CNN block with input obtained from Sinc filters [26] or Gabor filters [27]. Although wavegrams can learn new kinds of features over handcrafted spectrograms, important 1D temporal features can be extracted from time-domain 1D CNN blocks [28]. We utilized the SincConv and GaborConv1d functions implemented in SpeechBrain[4] [29] for processing the wavLM features with a sampling frequency of 16 kHz. By combining the advantages of the wavegram architecture with the extraction of temporal features through SincConv and GaborConv1d functions, our approach aimed to enhance the model to capture both the time-frequency representation and relevant temporal characteristics in predicting speech intelligibility.

C. Acoustic Parameter Model

We constructed an acoustic parameter model using eGeMAPS [16] as the input feature. The eGeMAPS is often used as a minimalist feature set in various speech processing tasks. While the eGeMAPS feature set may not be the primary choice for speech intelligibility prediction, in the context of hearing loss perception and hearing aid development, we include eGeMAPS as an additional feature set along with the auditory and wavegram-CNN model. In this case, the acoustic parameter model became a part of a comprehensive analysis modeling approach to understand the broader aspects of speech processing and perception in individuals with hearing loss.

The openSMILE [30] toolkit was utilized to extract the eGeMAPS features, including features related to frequency, energy, and spectral parameters. The distortion due to the hearing-impaired condition causes changes in pitch perception, frequency discrimination, and amplitude modulation. These acoustic parameters were reported to be highly associated with SI [3]. For instance, perceiving speech spoken by a female speaker is difficult for those with hearing loss since it has a higher fundamental frequency, lower spectral energy below 4 kHz, and higher spectral energy above 4 kHz [31]. A stack regressor consisting of a linear regressor, a support vector machine regressor, and a random forest regressor was used to learn the eGeMAPS features (as implemented in [14]). By using the combination of these three regressors, we can benefit from their individual strengths: providing interpretability, handling non-linear relationships, and capturing complex interactions, respectively. Together, they are used to improve the accuracy and robustness of speech intelligibility prediction from eGeMAPS features.

TABLE I

EVALUATION RESULTS OF SI PREDICTION MODELS: THE BASELINE MBSTOI (MBSTOI [11] + MBSG MODEL [12]), HASPI [9] (LEFT AND RIGHT), AND OUR PROPOSED METHODS. WE ALSO PROVIDE THE ABLATION TEST RESULTS (BY EXCLUDING ADDITIONAL FEATURES, I.E., EGEMAPS AND WAVLM). THE DIRECTION OF THE ARROW INDICATES THE BETTER CRITERIA OF EACH EVALUATION METRIC. THE ONE-WAY ANOVA CONDUCTED ON THE DATA REVEALED A SIGNIFICANT DIFFERENCE BETWEEN THE METHODS ($p < 0.01$).

Method	**Binaural**	**Non-intrusive**	**Track 1 (closed-set)**			**Track 2 (open-set)**		
			$\rho \uparrow$	**RMSE** $\downarrow$	$R^2 \uparrow$	$\rho \uparrow$	**RMSE** $\downarrow$	$R^2 \uparrow$
Baseline	Yes	No	0.62	28.52 ± 0.58	0.39	0.53	36.52 ± 1.35	−0.02
HASPI (left)	No	No	0.60	37.72 ± 0.60	−0.08	0.57	37.87 ± 1.20	−0.10
HASPI (right)	No	No	0.60	37.66 ± 0.60	−0.07	0.55	38.61 ± 1.23	−0.14
Proposed (Sinc)	Yes	Yes	0.72	25.46 ± 0.52	0.51	0.56	30.06 ± 1.19	0.31
Proposed (Gabor)	Yes	Yes	**0.74**	**24.90 ± 0.51**	**0.53**	0.64	28.04 ± 1.11	0.40
Ablation (Proposed (Gabor) + Excluded Feature)								
eGeMAPS			0.61	28.95 ± 0.59	0.37	0.43	33.54 ± 1.33	0.14
wavLM			0.74	25.09 ± 0.51	0.52	**0.64**	**27.85 ± 1.11**	**0.41**
eGeMAPS + wavLM			0.63	30.49 ± 0.61	0.30	0.50	31.76 ± 1.26	0.23

IV. EXPERIMENTS

A. Dataset

We conducted our experiment on the CPC1 dataset[2]. A large amount of speech data processed by hearing-aid (HA) systems were recorded in various scenes, and the corresponding metadata were provided in this dataset. Each recorded speech data point was a mixture of clean speech (target) and an interference signal in an anechoic cuboid room[3]. Subsequently, the improved speech in noise (SPIN) is defined as the recorded speech data enhanced by a machine-learning-based HA system. Six British English speakers and ten HA systems were involved in generating the improved SPIN. Furthermore, the SI label was obtained from the listening tests of 27 hearing-impaired listeners. A pure-tone air conduction audiogram was provided for each listener. CPC1 consists of (1) a *closed-set* track, which includes all unseen scenes from seen listeners and HA systems and; (2) a *open-set* track, which includes all unseen scenes from unseen listeners and HA systems.

We followed the data distribution from the challenge to train and evaluate our proposed method. To derive the optimal hyperparameters, we perform a grid search on the batch size and hidden units, ranging from 16 to 64. During the model training, the Adam optimizer was used with a learning rate of 0.001. The number of channels for the wavegram-CNN model is set to 32.

B. Evaluation

Three metrics were considered for evaluation: the Pearson correlation coefficient (ρ), root-mean-square error (RMSE), and coefficient of determination (R^2). These metrics were generally utilized for performance analysis of a regression task (SI prediction, which ranges from 0 to 100, is the CPC1 task). Additionally, we compared our proposed method with the baseline MBSTOI [2] and HASPI [9]. The one-way analysis of variance (ANOVA) was also performed to compare the differences among the predicted results obtained from each method.

[2]https://claritychallenge.org/clarity_CPC1_doc/docs/cpc1_data

[3]https://claritychallenge.org/clarity_CPC1_doc/docs/cpc1_scenario

[4]https://speechbrain.github.io/

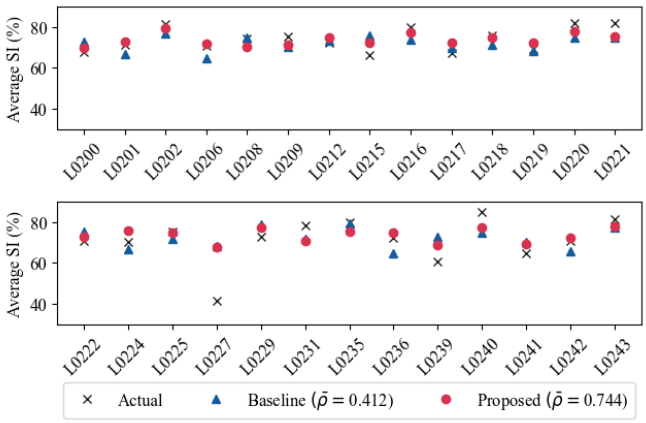

Fig. 2. Listening test results using a closed-set track. Actual is the label of SI (correctness in the listening test). Baseline and proposed represent the baseline MBSTOI and the proposed (Gabor) method, respectively, in Table I. $\bar{\rho}$ indicates the average ρ of the predicted SI and the actual correctness for each listener.

C. Results

The results are summarized in Table I. In general, SI prediction could be significantly improved using our proposed methods in comparison to HASPI, which was also developed using the EarModel [10]. The proposed methods are also non-intrusive and consider input from both ears (binaural). The best results for the closed-set track were achieved by combining all features and utilizing Gabor filters in the wavegram-CNN model. Meanwhile, the best results for the open-set track were achieved by the same model but excluding the wavLM input. We predict that this is due to more factors from unknown scenes with unknown systems and listeners embedded in the wavLM feature in the evaluation dataset in the open-set track.

Since the aim of our proposed method is to investigate the contribution of the auditory model based on hearing loss conditions, we also plot the results from the listening tests

in Fig. 2. The results show that our proposed method could be used to predict SI better than the baseline MBSTOI for almost all listeners (the $\bar{\rho}$ value improved by more than 0.3). Subsequently, the results show that the SI of listener L0227 is hard to predict, possibly due to the moderate hearing loss in both ears for almost all frequency ranges (based on the audiogram).

While this study provides valuable insights into the prediction of SI, it is important to acknowledge certain limitations that may impact the generalizability and robustness of the findings. Firstly, the dataset might not fully capture the diversity of listener characteristics and the complexity of real-world speech degradation scenarios. For instance, the average SI in Fig. 2 mostly lies in the interval of [60,80] because the existing data for each listener is imbalanced, especially for the score between 0 and 100. As a future direction, a more diverse and extensive dataset encompassing a wide range of degradation scenarios would enhance the robustness of the proposed method.

V. Conclusion

In this paper, a nonintrusive speech intelligibility prediction method was proposed by considering several relevant factors in hearing perception, including the listener's hearing condition, the spectrum of speech, and the acoustic parameters. Three main models, i.e., the auditory model, wavegram-CNN model, and acoustic parameter model, were exploited to retrieve the relevant information. Our experiment was conducted based on the protocol in CPC1. The results showed that the proposed method outperformed the intrusive baseline MBSTOI and HASPI methods. In addition, our proposed method, which is a nonintrusive (blind) method, does not require clean speech as input. Our future direction will focus on refining the binaural perceptual model and further analyzing hearing-impaired conditions in noisy environments.

Acknowledgment

This work was supported by the SCOPE Program of Ministry of Internal Affairs and Communications (No. 201605002), a Grant-in-Aid for Scientific Research (B) (No. 21H03463), the Japan Society for the Promotion of Science (JSPS) KAKENHI grant (No. 22K21304, No. 22H04860, and No. 22H00536), and JST AIP Trilateral AI Research, Japan (No. JPMJCR20G6).

References

[1] M. Munro and T. Derwing, "Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech," *Language and speech*, vol. 38 (3), pp. 289–306, 1995.

[2] J. Barker et al., "The 1st Clarity Prediction Challenge: A machine learning challenge for hearing aid intelligibility prediction," in *Proc. of Interspeech*. 2022, pp. 3508–3512, ISCA.

[3] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, pp. 90–119, 1945.

[4] M. Stone and B. Moore, "Effect of the speed of a single-channel dynamic range compressor on intelligibility in a competing speech task," *J. Acoust. Soc. Am.*, vol. 114, pp. 1023–34, 2003.

[5] A. Heinrich, H. Henshaw, and M. Ferguson, "The relationship of speech intelligibility with hearing sensitivity, cognition, and perceived hearing difficulties varies for different speech perception tests," *Frontiers in Psychology*, vol. 6, 2015.

[6] T. H. Falk et al., "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *IEEE Sig. Proc. Mag.*, vol. 32 (2), pp. 114–124, 2015.

[7] K. Miles et al., "Measuring speech intelligibility and hearing-aid benefit using everyday conversational sentences in real-world environments," *Frontiers in Neuroscience*, vol. 16, pp. 789565, 2022.

[8] Yong Feng and Fei Chen, "Nonintrusive objective measurement of speech intelligibility: A review of methodology," *Biomedical Signal Processing and Control*, vol. 71, pp. 103204, 2022.

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

[10] J. Kates, "An auditory model for intelligibility and quality predictions," *J. Acoust. Soc. Am.*, vol. 133, pp. 3560, 2013.

[11] A. H. Andersen et al., "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," *Speech Comm.*, vol. 102, pp. 1–13, 2018.

[12] Y. Nejime and B. Moore, "Simulation of the effect of threshold elevation and loudness recruitment combined with reduced frequency selectivity on the intelligibility of speech in noise," *J. Acoust. Soc. Am.*, vol. 102, pp. 603–15, 1997.

[13] S. Graetzer et al., "Clarity-2021 Challenges: Machine Learning Challenges for Advancing Hearing Aid Processing," in *Proc. of Interspeech*. 2021, pp. 686–690, ISCA.

[14] B. A. Titalim et al., "Speech intelligibility prediction for hearing aids using an auditory model and acoustic parameters," in *Proc. of APSIPA*, 2022, pp. 1076–1084.

[15] S. Chen et al., "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *ArXiv*, vol. abs/2110.13900, 2022.

[16] F. Eyben et al., "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Trans. on Affective Computing*, vol. 7, 2015.

[17] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns.," *J. Acoust. Soc. Am.*, vol. 74 3, 1983.

[18] J. Kiessling, "Current approach to hearing aid evaluation," *Canadian Journal of Speech-Language Pathology and Audiology*, vol. 17, no. 4, pp. 39–49, 1993.

[19] B. C. J. Moore et al., "Inter-relationship between different psychoacoustic measures assumed to be related to the cochlear active mechanism," *J. Acoust. Soc. Am.*, vol. 106 5, pp. 2761–78, 1999.

[20] Z. Tu, N. Ma, and J. Barker, "DHASP: Differentiable Hearing Aid Speech Processing," in *Proc. of ICASSP*, 2021, pp. 296–300.

[21] L. A. Westerman and R. L. Smith, "Conservation of adapting components in auditory-nerve responses," *J. Acoust. Soc. Am.*, vol. 81, pp. 680–691, 1987.

[22] J. Kates, "A time-domain digital cochlear model," *IEEE Trans. on Sig. Proc.*, vol. 39, pp. 2573–2592, 1991.

[23] Z. Tu, N. Ma, and J. Barker, "Unsupervised uncertainty measures of automatic speech recognition for non-intrusive speech intelligibility prediction," in *Proc. of Interspeech*. 2022, ISCA.

[24] C. O. Mawalim et al., "OBISHI: Objective Binaural Intelligibility Score for the Hearing Impaired," in *Proc. of SST*, 2022, pp. 111–115.

[25] R. E. Zezario et al., "MBI-Net: A Non-Intrusive Multi-Branched Speech Intelligibility Prediction Model for Hearing Aids," *arXiv*, 2022.

[26] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *SLT*. 2018, pp. 1021–1028, IEEE.

[27] N. Zeghidour et al., "LEAF: A Learnable Frontend for Audio Classification," in *ICLR*. 2021, OpenReview.net.

[28] Q. Kong et al., "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," *IEEE ACM TASLP*, vol. 28, pp. 2880–2894, 2020.

[29] M. Ravanelli et al., "SpeechBrain: A General-Purpose Speech Toolkit," *CoRR*, vol. abs/2106.04624, 2021.

[30] F. Eyben et al., "Recent Developments in OpenSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proc. of ACM MM*, 2013, p. 835–838.

[31] B. Larsby et al., "The influence of female versus male speakers' voice on speech recognition thresholds in noise: Effects of low- and high-frequency hearing impairment," *Speech, Language and Hearing*, vol. 18, pp. 83–90, 2015.