# OBISHI: Objective Binaural Intelligibility Score for the Hearing Impaired

*Candy Olivia Mawalim, Benita Angela Titalim, Masashi Unoki, and Shogo Okada*

Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi, Ishikawa 923–1292 Japan
{candylim, s2110104, okada-s, unoki}@jaist.ac.jp

## Abstract

Speech intelligibility prediction for both normal hearing and hearing impairment is very important for hearing aid development. The Clarity Prediction Challenge 2022 (CPC1) was initiated to evaluate the speech intelligibility of speech signals produced by hearing aid systems. Modified binaural short-time objective intelligibility (MBSTOI) and hearing aid speech prediction index (HASPI) were introduced in the CPC1 to understand the basis of speech intelligibility prediction. This paper proposes a method to predict speech intelligibility scores, namely OBISHI. OBISHI is an intrusive (non-blind) objective measurement that receives binaural speech input and considers the hearing-impaired characteristics. In addition, a pre-trained automatic speech recognition (ASR) system was also utilized to infer the difficulty of utterances regardless of the hearing loss condition. We also integrated the hearing loss model by the Cambridge auditory group and the Gammatone Filterbank-based prediction model. The total evaluation was conducted by comparing the predicted intelligibility score of the baseline MBSTOI and HASPI with the actual correctness of listening tests. In general, the results showed that the proposed method, OBISHI, outperformed the baseline MBSTOI and HASPI (improved approximately 10% classification accuracy in terms of F1 score).

**Index Terms**: hearing impaired, speech intelligibility, binaural hearing, hearing aids, hearing loss model

## 1. Introduction

Hearing aids are a technology that contributes to assisting sensorineural hearing loss. The hearing loss phenomenon can be explained in several ways. First, the auditory threshold is lifted above 0 dB or above the auditory threshold in normal hearing (NH). Second, the contribution of hair cells in inner ear damage to the signal compression and auditory threshold is shifted to the higher range [1, 2]. These factors describe how the damage in the inner ear and the noise level affect speech perception, and hearing aids should compensate for the loss. Speech processing is needed in hearing aids to enhance speech quality and intelligibility, especially in noise and reverberation.

One of the important evaluations for hearing aids is the speech intelligibility metrics. Speech intelligibility often refers to how accurately speech is understood or the percentage of the number of words the listener correctly identifies [3, 4]. The hearing aid speech prediction index (HASPI) by Kates and Arehart [5, 6] is often considered in developing hearing aids as an objective speech intelligibility index. The HASPI model includes a comparison of the temporal amplitude envelope (TAE) and temporal fine structure (TFS) that makes the prediction accuracy in both NH and hearing-impaired (HI) processing improved [5]. Unfortunately, the HASPI model has several drawbacks; that is, evaluation is limited to the conditions provided in the training data, handles monaural listening, only considers the audiogram for the listener's hearing characteristics, and is invalid for tonal languages.

Another alternative to measuring objective speech intelligibility is the modified binaural short-time objective intelligibility (MBSTOI) [7]. This model was developed based on the STOI metric [8] and is an extended model of discrete binaural STOI (DBSTOI) [9]. The MBSTOI generates more accurate predictions than the DBSTOI because it overcomes the tendency of overestimation when the interferers are spatially distributed. However, this model utilized a hearing loss model [10] to approximate the HI auditory thresholds by adding internal noise and by attenuating the signals. Thus, the baseline model is sensitive to the level of the processed signal.

This study proposes an objective binaural intelligibility score for the hearing impaired (OBISHI) to improve the speech intelligibility performance of existing methods. For instance, unlike the HASPI model, the proposed method handles binaural listening. Additionally, the proposed method considers not only the listener's audiogram but also other HI characteristics, such as the digit-triplet test (DTT) results. The proposed prediction model integrates a pre-trained automatic speech recognition (ASR) system to predict the difficulty of the sentence regardless of the hearing loss conditions, an HI characteristics (HICs) predictor, and an intelligibility model built on a gammatone filterbank.

## 2. Hearing-Impaired Intelligibility Model

The Clarity Challenge[1] was formed as one part of contributing to the development of hearing aid technology to improve the signal processing in the hearing aids system and to predict the perceived speech in noise (SPIN). One of the main tasks of this challenge is to predict the speech intelligibility of HI listeners when they perceive noisy speech processed by a hearing aid system [4]. It provides audio signals from simulated hearing aids receiving SPIN with the corresponding reference signals & transcript, the HI listeners' characteristics, and the speech intelligibility score as the ground truth obtained from listening tests. The simplified baseline system consists of a hearing loss simulation and binaural speech intelligibility models. However, the configuration of the prediction model can be altered, for example, by combining the hearing loss and speech intelligibility model with a single model. Two HI intelligibility models are also introduced in CPC1: HASPI and the baseline MBSTOI models.

[1] http://claritychallenge.org/

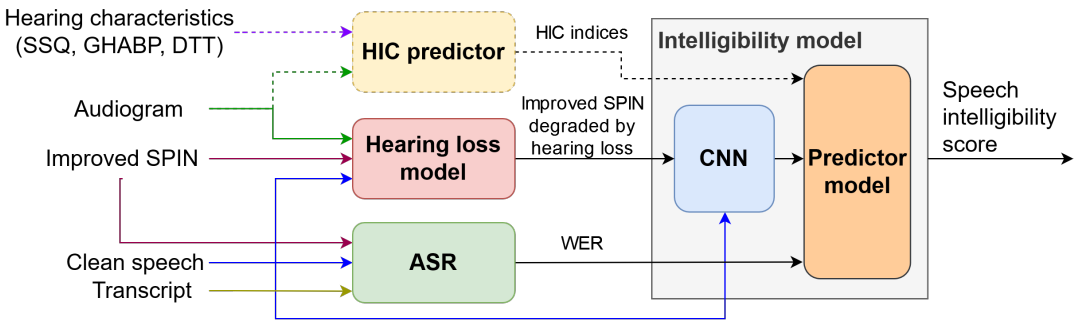


Figure 1: Block diagram of our proposed method. The dashed components were optionally included in the close-set prediction model. The HIC predictor is excluded in the open-set prediction model due to the unknown hearing characteristics of the unseen listener.

### 2.1. Hearing Aid Speech Prediction Index

HASPI is an index to predict speech intelligibility for NH and HI listeners. The model consists of analysis by a fourth-order gammatone filterbank (GTFB) with bandwidth changes for HI by a control filterbank. The control filterbank performs the processing of loudness recruitment and the signal intensity effect [11, 12] based on outer-hair cell (OHC). Besides, the damage to the inner hair cell (IHC), as a transducer that transmits mechanical force to an electrical signal to the brain, reduces transduction efficiency and results in additional attenuation [2]. The attenuation is represented by IHC firing-rate adaptation [13] in this model. The HASPI measurement generates cepstral coefficient sequences from the envelope output of the auditory periphery. The cepstral coefficient sequences then pass through modulation filters to give filtered sequences. The normalized cross-covariance is calculated and averaged across the basis function to produce an averaged vector of covariances. Finally, the speech intelligibility score is mapped using neural networks.

### 2.2. Modified Binaural Short-Time Objective Intelligibility

The baseline of the speech intelligibility prediction model for the Clarity Prediction Challenge 2022 [4] was the MBSTOI [14]. As the name suggests, MBSTOI is constructed based on the STOI metric that takes advantage of non-linear processing with no separate noise access requirement. The model also takes advantage of the first extended version, deterministic binaural STOI (DBSTOI), which covers the binaural processing and fluctuating interferers [7]. The intermediate correlation coefficient of the degraded and reference signals is calculated similarly to that in the STOI measurement. Then, the intelligibility score is defined as the average of the intermediate correlation across time and ⅓ octave frequency bands.

The baseline MBSTOI is constructed to predict the SPIN in broader types of noise, spatially different interferers, linear and non-linear processing, and reverberation. It corrects the broadband delay of the ear model due to hearing loss by running the Kronecker delta function. However, the baseline model did not fix the delay after the hearing aid processing. Moreover, a clean reference was not provided for the tests or evaluation set to correct the hearing aid processing delay in the challenge. Another problem with this model is that the MBSTOI is insensitive to the processed signal level. The measurement is based on the cross-correlation method, which may produce a highly intelligible result when the sound falls below auditory thresholds.

## 3. Proposed Method

The overall process in our method is shown in Fig. 1. The general inputs for both models are clean speech, improved SPIN (the output of the hearing aid system), and the audiogram of the HI listener. The HI characteristics (HIC) of the listener, including the results of speech, spatial, and qualities of hearing scale 12 (SSQ12) [15], the Glasgow hearing aid benefit profile (GHABP) [16], and the digit-triplet test (DTT) [17] were taken into account for inferring the HIC indices.

Our method has four main components: a HIC predictor, ASR, a hearing loss model, and an intelligibility model. The HIC predictor receives the SSQ, GHABP, DTT, and audiogram, resulting in the HIC indices representing each listener's characteristics. The imputation approach handled the missing data in the HIC characteristics, where the mean value is used to fill the missing data of the SSQ and GHABP of the available listeners. Meanwhile, we determined the DTT results using a prediction model by inputting the other listeners' characteristics and audiogram.

The ASR receives the clean speech and the output SPIN as inputs, and it outputs the word error rate (WER) of the predicted sentence of the output SPIN with the predicted sentence of the clean speech as the reference. We utilized a pre-trained ASR system [18] built using a factorized time delay neural network (TDNN-F) [19] that was trained on the LibriSpeech dataset [20]. The purpose of integrating an ASR in our model was to predict the sentence difficulty regardless of the HI condition (the recognition rate for the NH listener). The hearing loss model by the Cambridge auditory group [10] was utilized to estimate the improved SPIN degraded by hearing loss. The model, namely Moore, Stone, Baer, and Glasberg (MSGB) model, comprises simulations of acoustic transformation in the cochlea, spectral smearing and threshold elevation, and loudness recruitment.

Figure 2 shows a block diagram of how to generate the Convolutional Neural Network (CNN) input for our intelligibility model. We consider the speech inputs to be binaural signals. An infinite impulse response (IIR) time-domain GTFB[2] [21] with 32-channels was utilized to analyze the signals from both ears. Subsequently, we extracted the TAEs from the output of each channel in the GTFB analysis. These TAEs were then passed through a CNN (as shown in Fig. 3). The final predictor model consists of two layers of a fully-connected network with a recti-

[2] https://github.com/huynguyenqc/MOSA-Net-Cross-Domain/tree/VoiceMOSChallenge/gammatone_iir

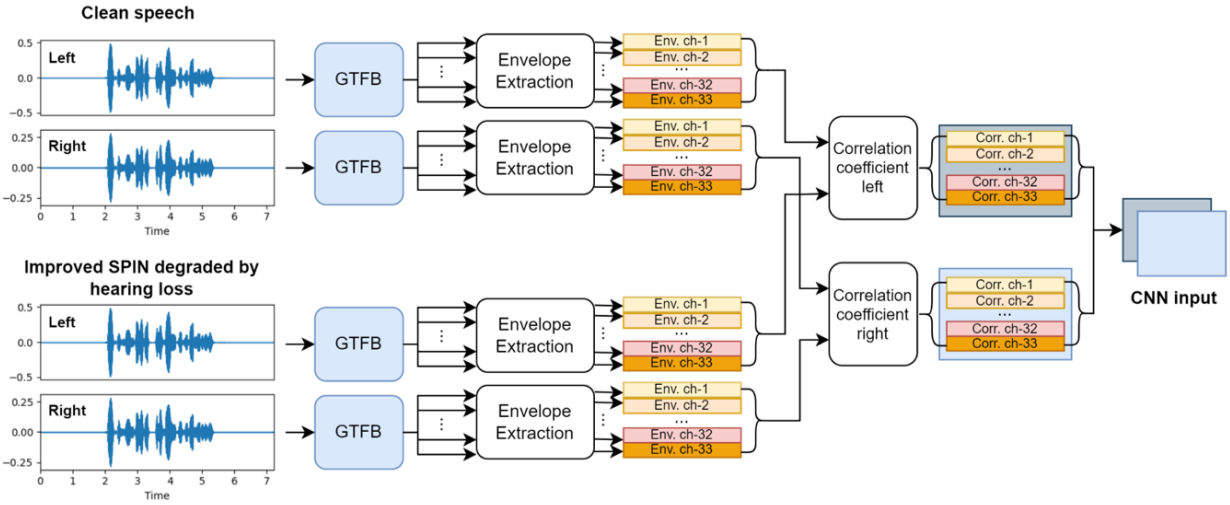


Figure 2: CNN input generation

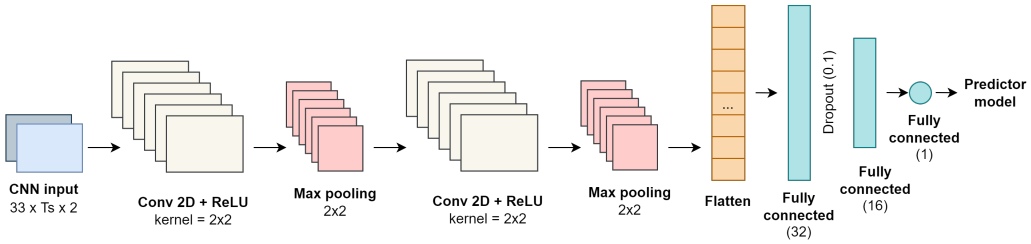


Figure 3: CNN architecture

fied linear unit (ReLU) activation function that receives the output of the CNN layer, the HIC indices, and the WER to predict the speech intelligibility score. We used the Adam optimizer algorithm and the mean-squared error (MSE) loss function in the training process.

## 4. Evaluation

### 4.1. Dataset

We utilized the dataset available in the Clarity Prediction Challenge 1 (CPC1)[3] [14]. Generally, it consists of a relatively large number of 44.1-kHz, 32-bit mono or stereo wav files and their corresponding metadata. The wav files are generated scenes, interferers, original target speech spoken by British English speakers, and improved SPINs (the output of SPINs after passing through hearing aid processors). The metadata provides detailed information related to the scenes, listeners, and transcripts. The dataset has six speakers, ten hearing aid processors in the first Clarity Enhancement Challenge [14], and 27 HI listeners. The hearing ability conditions of each listener are also available, including the pure-tone air-conduction audiogram for both ears, the DTT [17] results, and two self-assessment results (i.e., SSQ12 [15] and the GHABP questionnaire [16]). Unfortunately, some of the data of the DTT, SSQ12, and GHABP questionnaire results were missing for several listeners.

[3] https://claritychallenge.github.io/clarity_CPC1_doc/docs/cpc1_data

CPC1 has two tracks: track 1 (close-set) and track 2 (open-set). Each track has a different distribution of training/development and testing sets. The training/development and testing sets for both tracks do not overlap. We split the training/development data of track 1 (4,863 scenes) into 90% for training data and 10% for development data. The testing data of track 1 consists of 2,421 scenes. Track 2 consists of 3,580 training/development scenes and 632 test scenes. We split the training and development data by a leave-one-listener-and-one-system-out approach. This approach results in 2,933 scenes for training and 647 scenes for development data.

### 4.2. Evaluation Metrics

We used four metrics for evaluating our model, baseline MBSTOI, and HASPI: Pearson correlations ($\rho$), root-mean-square error (RMSE), F1 score (F1), and area under the curve (AUC) [22]. The speech intelligibility prediction model generates an intelligibility score ranging from 0 to 100, as defined in the CPC1 challenge [4]. Then, we converted the scale of baseline MBSTOI and HASPI from 0–1 to 0–100 by performing the RMSE minimization using a sigmoid function. We calculated the $\rho$ and RMSE of the predicted scores with the actual correctness of the subjective listening test. The F1 and AUC scores were obtained using binary classification (high and low). The score is classified as high when it is larger than 50 (middle point of 0–100); otherwise, it is classified as being a low score.

Table 1: Evaluation results of several speech intelligibility prediction models: MBSTOI (Baseline), HASPI left ear (HASPI (left)), HASPI right ear (HASPI (right)), and our proposed model with HIC predictor (OBISHI+HIC) and without HIC predictor (OBISHI).

| Dataset | Method | Track 1 (close-set) | | | | Track 2 (open-set) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $\rho$ | RMSE | F1 (%) | AUC (%) | $\rho$ | RMSE | F1 (%) | AUC (%) |
| Dev | **Baseline** | 0.63 | 33.65 ± 1.42 | 81.01 | 76.11 | 0.48 | 33.77 ± 0.92 | 84.57 | 67.18 |
| | **HASPI (left)** | 0.67 | 36.07 ± 1.34 | 73.13 | 71.91 | 0.43 | 43.58 ± 1.02 | 52.91 | 58.15 |
| | **HASPI (right)** | 0.67 | 35.57 ± 1.34 | 73.10 | 72.27 | 0.45 | 42.40 ± 1.01 | 57.16 | 58.67 |
| | **OBISHI** | 0.70 | 25.97 ± 1.21 | **88.55** | 85.23 | **0.60** | **22.81 ± 0.84** | **90.92** | **77.19** |
| | **OBISHI+HIC** | **0.77** | **23.97 ± 1.16** | 88.21 | **86.13** | | | | |
| Test | **Baseline** | 0.62 | 28.52 ± 0.58 | 81.83 | 75.74 | 0.53 | 36.52 ± 1.35 | 68.39 | 68.74 |
| | **HASPI (left)** | 0.60 | 37.72 ± 0.60 | 68.33 | 68.56 | 0.57 | 37.87 ± 1.20 | 67.88 | 68.58 |
| | **HASPI (right)** | 0.60 | 37.66 ± 0.60 | 68.33 | 68.56 | 0.55 | 38.61 ± 1.23 | 67.05 | 67.99 |
| | **OBISHI** | **0.68** | **27.86 ± 0.54** | 85.04 | 80.72 | **0.67** | **28.29 ± 1.06** | **82.90** | **78.69** |
| | **OBISHI+HIC** | 0.41 | 37.19 ± 0.72 | **85.16** | **87.11** | | | | |

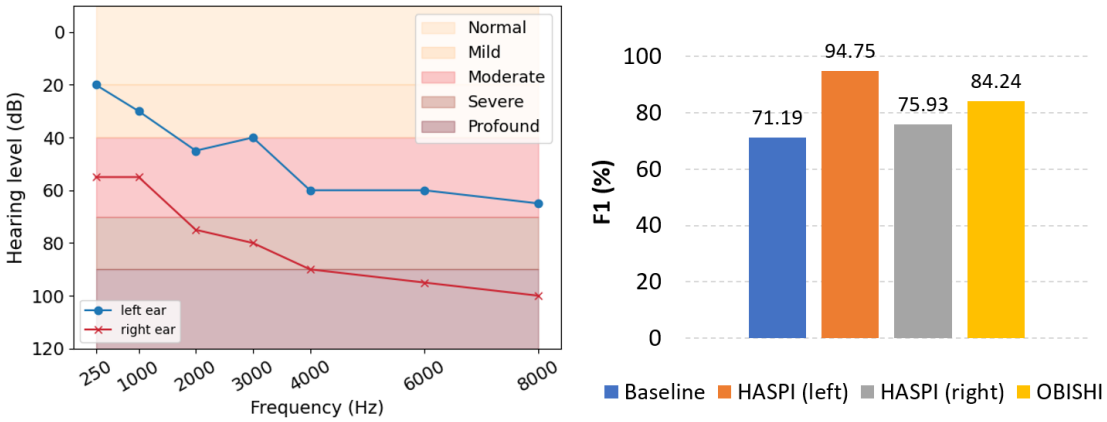


Figure 4: A case study on listener L0217. The left panel shows the audiogram of listener L0217. The right panel shows the speech intelligibility predictions results using comparative methods in terms of F1 score.

### 4.3. Results

Table 1 shows the total evaluation results of three models in both the development and testing phases. In general, our method improved the intelligibility prediction compared to the baseline and HASPI models.

**[Development phase]** In comparison with the baseline model, our prediction method significantly reduces the RMSE by approximately 10% and classification accuracy (F1 and AUC) by more than 5% for both tracks. An additional hearing characteristic predictor could also slightly improve the close-set scenario prediction. The performance of HASPI is generally not as high as that of the baseline and proposed models.

**[Testing phase]** Overall results during the testing phase also indicate that our proposed method has the best performance. However, although the classification accuracy could be improved, the additional HIC predictor in the proposed model (OBISHI+HIC) increased the RMSE and reduced the correlation. We predicted that this issue was caused by the rising number of missing hearing characteristics data occurring in the test set of track 1, which is larger than the development set. Although the imputation approach has been applied to the missing data, the model may fail to predict the relevant hearing characteristics beneficial for predicting speech intelligibility scores. Without the HIC predictor, our proposed method can improve the prediction accuracy, especially in track 2.

**[A case study]** We also plotted the classification prediction results and the audiogram of a specific HI listener 'L0217' in Fig. 4. We chose this listener because the hearing condition of the left ear is different than the right ear. The audiogram in Fig. 4 revealed listener L0217 has profound hearing loss in the right ear but a moderate hearing loss in the left ear at a higher frequency ($>$ 4 kHz). This condition is well represented by the HASPI model, where the left ear is better than the right ear. Meanwhile, the proposed model can balance the intelligibility prediction of both ears (F1 = 84.24%) better than the baseline model (F1 = 71.19%).

## 5. Conclusions

This paper proposed an objective binaural intelligibility score for the hearing impaired, OBISHI. The OBISHI belongs to an intrusive metric that considers the HI characteristics for predicting the speech intelligibility score. Additionally, we utilized an ASR system to infer the difficulty of the utterances in an NH condition. We integrated the MSBG hearing loss model with our constructed GTFB-based predictor model in the intelligibility model. The evaluation was conducted using a training test split method on two tracks (close-set and open-set). We also compared the predicted intelligibility score of the baseline MBSTOI and HASPI with the actual correctness from the listening test. The results showed that our method could significantly improve the prediction of the baseline MBSTOI and HASPI for both close-set and open-set tracks. In addition, our proposed method significantly improved the speech intelligibility prediction when the listener has different hearing impaired conditions of left and right ears compared to the baseline method.

## 6. Acknowledgements

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), and a JSPS KAKENHI grant (no. 22K21304). This work was also partially supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI (grant numbers 22H04860 and 22H00536) and JST AIP Trilateral AI Research, Japan (grant number JPMJCR20G6).

## 7. References

[1] C. J. Plack, V. Drga, and E. A. Lopez-Poveda, "Inferred basilar-membrane response functions for listeners with mild to moderate sensorineural hearing loss." *Journal of the Acoustical Society of America*, vol. 115 4, pp. 1684–95, 2004.

[2] B. C. J. Moore, D. A. Vickers, C. J. Plack, and A. J. Oxenham, "Inter-relationship between different psychoacoustic measures assumed to be related to the cochlear active mechanism." *Journal of the Acoustical Society of America*, vol. 106 5, pp. 2761–78, 1999.

[3] 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, 07 1995.

[4] 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 *INTERSPEECH 2022*. ISCA, 2022.

[5] J. Kates and K. Arehart, "The hearing-aid speech perception index (HASPI)," *Speech Communication*, vol. 65, 11 2014.

[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. de Haan, Z. Tan, and J. H. Jensen, "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," *Speech Commun.*, vol. 102, pp. 1–13, 2018.

[8] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[9] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "A method for predicting the intelligibility of noisy and non-linearly enhanced binaural speech," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4995–4999.

[10] 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," *Journal of the Acoustical Society of America*, vol. 102, pp. 603–15, 08 1997.

[11] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns." *Journal of the Acoustical Society of America*, vol. 74 3, pp. 750–3, 1983.

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

[13] D. M. Harris and P. Dallos, "Forward masking of auditory nerve fiber responses," *Journal of neurophysiology*, vol. 42, no. 4, pp. 1083–1107, 1979.

[14] S. Graetzer, J. Barker, T. J. Cox, M. Akeroyd, J. F. Culling, G. Naylor, E. Porter, and R. V. Muñoz, "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing," in *Interspeech 2021*. ISCA, 2021, pp. 686–690.

[15] K. Andersson, L. Andersen, J. Christensen, and T. Neher, "Assessing Real-Life Benefit From Hearing-Aid Noise Management: SSQ12 Questionnaire Versus Ecological Momentary Assessment With Acoustic Data-Logging," *American Journal of Audiology*, vol. 30, 12 2020.

[16] W. Whitmer, P. Howell, and M. Akeroyd, "Proposed norms for the Glasgow hearing-aid benefit profile (GHABP) questionnaire," *International journal of audiology*, vol. 53, 02 2014.

[17] E. V. den Borre, S. Denys, A. van Wieringen, and J. Wouters, "The digit triplet test: a scoping review," *International Journal of Audiology*, vol. 60, no. 12, pp. 946–963, 2021.

[18] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "The VoicePrivacy 2020 Challenge evaluation plan," 2020.

[19] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH 2015*, 2015, pp. 3214–3218.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE ICASSP*, 2015, pp. 5206–5210.

[21] M. Unoki and M. Akagi, "A method of signal extraction from noisy signal based on auditory scene analysis," *Speech Communication*, vol. 27, pp. 261–279, 4 1999.

[22] D. Freedman, R. Pisani, and R. Purves, "Statistics (international student edition)," *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*, 2007.