THE 2ND CLARITY PREDICTION CHALLENGE: A MACHINE LEARNING CHALLENGE FOR HEARING AID INTELLIGIBILITY PREDICTION

Jon Barker[1], *Michael A. Akeroyd*[2], *Will Bailey*[1], *Trevor J. Cox*[3], *John F. Culling*[4], *Jennifer Firth*[2], *Simone Graetzer*[3], *Graham Naylor*[2]

[1] Department of Computer Science, University of Sheffield, UK
[2] School of Medicine, University of Nottingham, UK
[3] Acoustics Research Centre, University of Salford, UK
[4] School of Psychology, Cardiff University, UK

ABSTRACT

This paper reports on the design and outcomes of the 2nd Clarity Prediction Challenge (CPC2) for predicting the intelligibility of hearing aid processed signals heard by individuals with a hearing impairment. The challenge was designed to promote new approaches for estimating the intelligibility of hearing aid signals that can be used in future hearing aid algorithm development. It extends an earlier round (CPC1, 2022) in a number of critical directions, including a larger dataset coming from new speech intelligibility listening experiments, a greater degree of variability in the test materials, and a design that requires prediction systems to generalise to unseen algorithms and listeners. This paper provides a full description of the new publicly available CPC2 dataset, the CPC2 challenge design, and the baseline systems. The challenge attracted 12 systems from 9 research teams. The systems are reviewed, their performance is analysed and conclusions are presented, with reference to the progress made since the earlier CPC1 challenge. In particular, it is seen how reference-free, non-intrusive systems based on pre-trained large acoustic models can perform well in this context.

Index Terms— speech-in-noise, speech intelligibility, hearing aids, hearing loss, machine learning

1. INTRODUCTION

According to the WHO [1], 430 million people have a disabling hearing loss at a cost of $750 billion. While hearing aids are the main treatment available, users find them problematic in some speech-in-noise scenarios. Machine learning approaches offer much promise in this situation; however, their development is held back by a lack of reliable evaluation metrics. Listening tests can be used but these are slow, expensive and cannot be built into the machine learning training workflow. More reliable instrumental, or 'objective', intelligibility metrics are required.

There had been much work on speech intelligibility prediction dating as far back as the 1950's in domains such as speech communication and coding. Feng and Chen provide a useful recent review of different non-intrusive methods, where the algorithm only has access to the speech-in-noise signal [2]. For intrusive approaches where the clean speech is also accessed by the algorithm, popular approaches include the Short-Time Objective Intelligibility measure (STOI) [3], the Speech Transmission Index (STI) [4] and the Speech Intelligibility Index (SII) [5]. However, research has mostly been driven by the needs of 'normal' hearing listeners.

HASPI (the Hearing-Aid Speech Perception Index) is a notable exception that considers the hearing abilities of listeners and the typical noise and distortion that might arise from hearing aids [6]. However, HASPI is an intrusive model that requires the clean reference speech signal, which is not always available in real situations. Further, HASPI is a single-ear model that does not consider binaural effects.

To foster new approaches to predicting speech intelligibility for people with a hearing loss, who are listening via hearing aids, the Clarity Prediction Challenge (CPC) series was launched in 2021. The first round, CPC1, led to many new solutions, many of which outperformed the HASPI baseline [7]. However, the static scenes and limited listening data available for CPC1 meant it was difficult to test the generalisibility of the entrants' approaches. Furthermore, many of the systems still required reference signals.

In this paper, we present the design and outcomes of the 2nd Clarity Prediction Challenge (CPC2), which ran from 2022 to 2023. It greatly extended the previous round in a number of significant ways. First, the challenge used speech-in-noise data that had come from a more complex and more realistic listening scenario involving multiple interferer sources and head motion. Second, the challenge used a wider variety of hearing aid signal processing algorithms. Third, a larger amount of listening data was provided for both training and evaluation. Fourth, the larger amount of data allowed a design that better tested generalisation with different systems and listeners in the training and tests sets. Finally, the challenge entrants had the benefit of the work from CPC1, which had shown the promise of non-intrusive systems that took advantage of statistical speech models adopted from the speech recognition community.

Although the challenge has now closed, all the software and data resources are now in the public domain and free for use for further research. Hence, this paper serves both as a review of the work that has been completed, and also as a reference for these public resources. The dataset of listening test responses is one of the largest of its kind in the public domain[1].

The remainder of this paper is structured as followed. Section 2 describes the CPC2 challenge materials including the hearing aid output signals and the listening test responses. Section 3 describes the construction of the machine learning challenge, including how

The Clarity project is funded by the UKRI (grants EP/S031448/1, EP/S031308/1, EP/S031324/1 and EP/S030298/1). We thank Amazon, the Hearing Industry Research Consortium and the Royal National Institute for the Deaf (RNID) for their support.

[1] The CPC2 data are available at https://claritychallenge.org/.

the training and evaluation data have been defined, the rules that were set to focus the research and the baseline system that was provided. Section 4 presents an overview of the systems that were submitted, with the performance of these systems and the challenge outcomes presented in Section 5. The paper concludes with thoughts on the progress that has been made and priorities for the next challenge round in Section 6.

2. MATERIALS

Entrants were given the task of predicting the intelligibility of signals, of varying SNRs (see below), that had been processed by a range of hearing aid algorithms and presented to listeners with a hearing loss. Below we describe the signals that were processed, the hearing aid systems, and the listeners and listening test procedures.

2.1. The Signals

The signals were a subset of 1500 scenes that had been generated for the 2nd Clarity Enhancement Challenge (CEC2) [8]. These were more complex than the scenes used for the previous prediction challenge (CPC1), involving more complex interferers and head movement. The target speech consisted of 7–10 word sentences [9]. This speech overlapped in time with 2–3 interfering noise sources. Unlike CPC1, the interferers could be music, speech or domestic appliance noises in any combination.

Each scene was constructed using a geometric room acoustic model [10]. All sources were randomly located within a room with the dimensions of a typical living room. The listener was at least 1 m from the room boundaries. Sources were at least 1 m from the listener and the walls. Sixth order Ambisonic room impulse responses were used. Head rotations by the listener were simulated in the Ambisonic domain. The listener turned their head towards the location of the target speaker around the time that the target utterances started. To proceed from Ambisonics to the hearing aid input signals, an HRTF (Head-Related Transfer Function) database was used [11]. This contained data for behind-the-ear hearing aids with 3 microphones on both the left and right aids. The level of the target speech was set to achieve a range of target SNRs from -12 dB to +6 dB, which had been chosen based on informal pilot listening tests using unprocessed signals and normal hearing listeners. Further details of the simulation and signal construction can be found in [8].

2.2. The Hearing Aid Systems

The hearing aid inputs were processed by the experimental hearing aid systems to produce the signals that were used in the challenge and listening tests. The hearing aid algorithms were those that had been submitted to CEC2. In CEC2, entrants had been provided with pairings of hearing aid input signals and the left and right audiograms of a specific listener with a hearing loss. They then produced the personalised stereo signals that were used in the listening tests.

CEC2 attracted 18 submissions from 7 teams, which varied widely in quality. Ten of these were used in the listening tests. The systems varied according to the strategies used for single channel enhancement, multichannel processing and signal amplification – see [8] and papers referred to therein. Importantly for the CPC2 prediction challenge, the signals covered a large range of intelligibility scores whether measured by listeners or predicted by objective intelligibility measures.

2.3. The Listening Tests

The human intelligibility scores that the CPC2 entrants were asked to predict come from an extensive set of listening tests involving 18 listeners with a hearing loss. The panel was recruited via the University of Nottingham and Nottingham Audiology Services. They were a subset of 27 listeners who had participated in the CEC1/CPC1 listening tests. Each was characterised by a bilateral audiogram measured at [250, 500, 1000, 2000, 3000, 4000, 6000, 8000] Hz. The main inclusion criteria were that listeners should not have losses in excess of 80 dB in two or more consecutive bands and should be users of acoustic hearing aids. Listeners with Ménière's disease, hyperacusis or severe tinnitus were excluded. Hearing loss when broadly characterised by averaging the dB loss between the frequencies 2 and 8 kHz, was found to be mild ($\leq$ 35 dB) for 1 listener, moderate (35–56 dB) for 5 listeners and severe (>56 dB) for 12 listeners. Most could be characterised as having sloping audiograms with milder loss at low frequencies, typical of age-related hearing loss.

Listeners were presented with a subset of CEC2 signals that had been generated for their audiograms. The listening tests were conducted by the listeners in their own homes [12, 7]. In brief, they used headphones, a tablet and were not wearing their hearing aids. Signals were presented in blocks with the same target talker, preceded by a number of enrollment sentences where they heard the target talker alone. They were asked to repeat the words spoken by the target talker. Responses were recorded and then transcribed using the Whisper Automatic Speech Recognition (ASR) engine [13]. Subsequently, these transcriptions were validated by human transcribers.

Each listener response was then transformed into a single numeric sentence intelligibility score by aligning the transcription with the original target sentence and calculating the percentage of words recognised correctly. In total, the listening tests provided 10,062 signal/response pairs.

3. CHALLENGE TASKS AND BASELINE

3.1. Task and Datasets

Entrants were provided with hearing aid output signals and the listener audiograms and tasked with predicting the sentence intelligibility. Both intrusive and non-intrusive systems were accepted. For intrusive systems, participants also had access to the anechoic, noise-free target speech, and also the target utterance text. Non intrusive systems were required to use only the hearing aid outputs.

Data were divided into training and evaluation data. For the training data, entrants were given the true listener correctness achieved in the listening tests. In addition, they had access to the complete transcription of the listener's response, and the full details of the signal construction (e.g., the component sources, room geometry and hearing aid algorithm). This information was not available for the evaluation data.

The training and evaluation datasets were constructed using 5,946 listener responses from the CEC2 listening tests. We also provided the 6,297 listener responses from the CEC1 listening test, which used the simpler scenes with a single interferer and a different set of hearing aid algorithms. The CEC1 data were made available as additional training data but did not appear in the evaluation set. The evaluation set was designed to use listeners and hearing aid systems not seen during training. Also, none of the target sentences appeared in both the training and evaluation sets. Due to only having 10 hearing aid algorithms, the evaluation used a three-fold strategy with three evaluation splits, i.e., test.1, test.2, test.3, and three corresponding, disjoint, training data splits, train.1, train.2 and train.3.

Participants were asked to train three separate versions of their models and to strictly only use the train.1 data for evaluation of the test.1 set, etc. Figure 1 shows the training and test set split indicating the number of responses, systems and listeners.

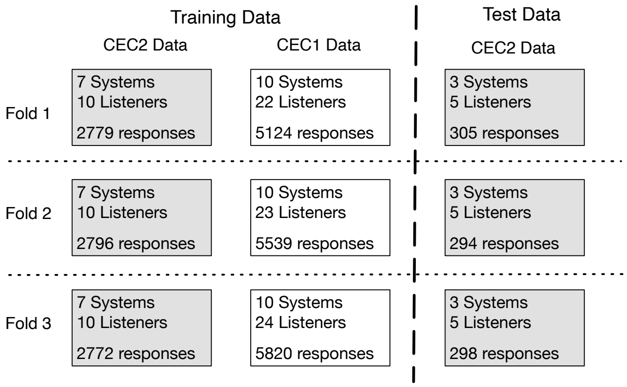

Fig. 1. The design of the CPC2 training and evaluation sets.

Entrants submitted predictions for remote evaluation by the organisers. Systems were evaluated by computing the Root Mean Square error (RMSE) and the Pearson correlation coefficient between the predicted intelligibilities and the true values. Responses for all three test sets were combined before computing a single RMSE value that was used to rank the entrants.

3.2. Baseline system

Participants were provided with an implementation of a baseline system based on HASPI. (In CPC1, the baseline used a hearing loss model combined with a binaural version of the STOI intelligibility metric, but the HASPI system was found to perform better.) The CPC2 baseline computes HASPI scores for the left and right ear separately and then takes the maximum of the two scores. This is to model the better ear effect of binaural listening. The HASPI scores, which range from 0.0 to 1.0, are mapped onto the correctness scores in the range 0.0 to 100.0 via a logistic function. The logistic function parameter values were obtained by a least-mean-square fitting using the training datasets. We call this better ear HASPI (beHASPI).

The RMSE score for the baseline was 28.7%. The overall performance is shown in the left plot of Figure 2. Averaging over groups of samples is done because there is a lot of unexplainable variability in the listener data that makes looking at scatter plots of individual samples less instructive. In the left box plot, samples are batched into 10-percentile intelligibility ranges and the mean of the true and predicted values for each batch is shown. While the HASPI scores are broadly correlated with the true intelligibilities, they do not span the full range and often severely over-predict the scores for the least intelligible samples (i.e., there is positive bias).

4. SUBMISSIONS

There were 12 systems in total originating from 9 separate teams. Systems have been categorised as intrusive or non-intrusive (see the columns 'Intr' and 'Non-Intr' in Table 1). Systems are summarised below with details available in the references provided in Table 1.

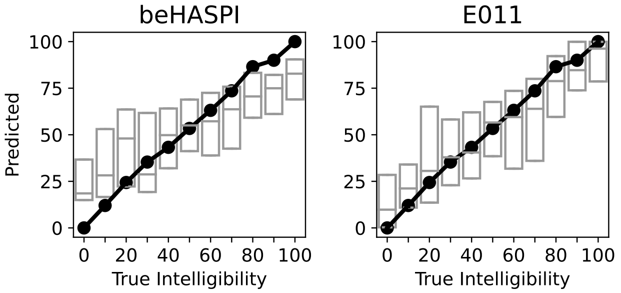

Fig. 2. Comparing the overall performance of the beHASPI baseline (left) and best system, E011 (right). Test set samples are grouped into 10-percentile intelligibility bands. For each percentile, the true mean intelligibility is shown by the solid line and black dots whereas the distribution of predicted values (mean and 25th to 75th percentile range) is shown by the box plot.

4.1. Modelling Hearing Loss

A variety of approaches have been used to model hearing loss that can be broadly categorised as either 'auditory' or 'statistical'. The most popular auditory approach has been to apply a 'hearing loss simulator' directly to the signal as a preprocessing stage. Systems E016 and E023 employed the Cambridge "MSBG" hearing loss simulator that was provided by the organisers, whereas E015 used the same approach but used the Irino hearing loss simulator [14]. Both these simulators use the audiogram but differ in how they emulate the effects of auditory bandwidth widening, temporal smearing, raised thresholds and loudness recruitment. System E020 operated by using a model of the impact of hearing loss on auditory nerve firing [15] and comparing the reference and noisy signals in the 'neurogram' domain.

The statistical approaches used the training data to understand the relationship between audiograms and reported intelligibilities. E009 directly used the listener audiograms as additional inputs to a neural-network regression model alongside an encoding of the acoustic features. E011, which used a deep-learning transformer model to process the audio, introduced the audiogram via a learnt linear transform to increase the dimensionality before concatenating it with the acoustic encoder output. E024 also used deep-learning, while taking an alternative approach that introduced a one-hot listener embedding layer in their model, which encoded listener identity but did not use specific audiogram thresholds. Teams E002, E022 and E025, surprisingly, did not employ the audiogram information at all. It had been shown in CPC1 that the audiograms were not highly correlated with intelligibility. Although counter-intuitive, this is consistent with the fact that the hearing aid signals had already been processed to compensate for listeners' raised auditory thresholds.

4.2. Binaural modelling

Approaches to allowing for binaural listening can be categorised according to the processing stage at which the left and right ear pathways were integrated. The baseline (beHASPI) makes predictions for the left and right ear independently and takes the highest score to simulate better ear listening. However, this does not explicitly model binaural unmasking, where listeners integrate information across the two ears. This can happen even when natural binaural cues have been disrupted [16]. System E009 addresses this by using a similar 'late-integration', taking the best of the left and right ear STOI measures but making decisions at the level of local time-frequency cells before integrating. Deep-learning systems E016 and E023 combined frame-level intelligibility estimates for the left and right ear with a final linear layer before averaging. E015 is similar but performed an earlier integration, merging the latent representation at a central network layer. E011 merged the left and right network branches using a cross-attention mechanism at an intermediate level of representation. System E024 was unique in performing an early integration, processing the left and right ear signals through an equalisation-cancellation (EC) model – a well-established model of spatial hearing used in the binaural form of STOI – and encoding the output signal [17].

4.3. System architectures

In contrast to the CPC1 challenge, most entrants chose to design non-intrusive systems (10 out of 12 systems). The dominant approach was to leverage pre-trained deep-learning models and use these to extract a representation of the hearing-aid outputs (with or without hearing-loss modelling – see Section 4.1). This also produced the best performance – see Section 5.

As an example of an intrusive approach, system E009 uses the wav2vec 2.0 [18] model and fine tunes it on the WSJCAM0 British English corpus [19] to estimate phone lattice representations of the reference signal and hearing aid outputs. Intelligibility is then based on correlation between the phone estimates. This is conceptually similar to other intrusive measures such as HASPI and STOI, except that it is performing the comparison using a higher-level of abstraction, i.e., phoneme probabilities, rather than a lower level such as frequency band envelope modulations.

Non-intrusive approaches exploited the implicit uncertainty of the internal representation of speech in pre-trained speech models. This assumes that this ambiguity of a speech signal is correlated with human-listener intelligibility. E025 uses this idea explicitly where the beam-search in the ASR inference stage is used to generate a sample of hypotheses and an entropy measure is then used to estimate the certainty of the most probable one.

Alternatively, a pre-trained speech model can be used more akin to a feature extractor by obtaining an 'embedding vector' from a layer within the network that can be passed through further layers trained to estimate the intelligibility score. This approach was taken by E016, E023, E002, E003, E024 and E011. Two well-known pre-trained models have been employed: Whisper [13] (E016, E023, E002, E011) and WavLM [20] (E003, E024, E011). E011 was unique in combining both and in finding some advantage in doing so. Although these systems all use pre-trained speech models they vary considerably in their architecture, e.g., models using a sequence of convolutional, recurrent and attention layers such as E016, E023 and E002, versus a transformer model with self-attention and cross-attention between the left and right ear pathways for E011.

5. RESULTS

Results for all systems are presented in Table 1 reporting the RMSE error score (with the ± indicating one standard error) and the Pearson correlation ('Corr') between the predicted and ground truth listening test scores. Systems are ordered by RMSE score with the best performing at the top. The table also includes the baseline system ('beHASPI') and the outcome that would be achieved by simply using the training data mean intelligibility score for all samples ('Prior').

Table 1. Evaluation of 12 submitted systems plus baselines showing RMS prediction error (RMSE) and ground-truth vs prediction Pearson correlation coefficient (Corr). Systems are categorised as intrusive ('Intr') or non-intrusive ('Non-Intr'). 'Prior' is a system that simply outputs the mean of the training data.

Team	System	Intr.	Non-Intr.	RMSE ↓	Corr ↑
T01	E011 [21]		X	**25.1 ± 0.8**	0.78
T02	E002 [22]		X	25.3 ± 0.8	0.77
T03	E009 [23]	X		25.4 ± 0.8	0.78
T04	E022 [24]	X		25.7 ± 0.9	0.77
T05	E023 [25]		X	26.4 ± 0.9	0.76
T05	E016 [25]		X	26.8 ± 0.9	0.75
T04	E025 [24]		X	27.9 ± 0.9	0.72
Base.	beHASPI	X		28.7 ± 1.0	0.70
T06	E003 [26]		X	31.1 ± 1.0	0.64
T06	E024 [26]		X	31.7 ± 1.0	0.62
T07	E015 [27]		X	35.0 ± 1.1	0.60
T08	E020 [28]		X	39.8 ± 1.3	0.33
Base.	Prior		X	40.0 ± 1.3	–
T09	E019		X	(withdrew)	–

Seven submissions were able to beat the beHASPI baseline which had an RMSE score of 28.7%. The very best system, E011, achieved a score of 25.1%. Figure 2 (right) shows that system's performance in more detail. Although the 3.6% overall improvement is small, Figure 2 shows how the improvement is focused on the most and least intelligible cases. The top scoring system was only fractionally better than the 2nd placed (25.1% vs 25.3%) but a paired one-tailed t-test showed the difference to be statistically significant ($t(896)=2.71$, $p=.003$).

Remarkably, the two top ranked systems (E011, E002) are both non-intrusive, and outperformed the best intrusive systems (E009, E022) despite not having access to the reference signal. This represents real progress since CPC1, where we noted that non-intrusive systems were able to come close to the intrusive system performance but did not surpass it. The primary difference is that although the non-intrusive systems in CPC1 were using similar approaches based on embeddings extracted from DNNs trained on speech signals, they were not using the recently-available large pre-trained models such as Whisper and WavLM that entrants used in CPC2. The two systems are highly complementary; a combined system produces a better performance than either individual system ($RMSE$=23.7, r=0.80).

6. CONCLUSIONS

The second round of the Clarity Prediction Challenges for speech intelligibility (CPC2) had data spanning broader scenarios than the first round (CPC1). This paper outlined the materials for the challenge, which are publicly available. The availability of large pre-trained speech models enabled non-intrusive approaches to be most successful. The development of speech intelligibility metrics that can be calculated without access to a clean reference signal is extremely useful and should facilitate new machine learning approaches for enhancing speech-in-noise for hearing aids. For the next prediction challenge, the plan is to broaden the scenarios even further to include listening outdoors. This will be based on data generated by Clarity's third enhancement challenge (CEC3), which will open early in 2024.

7. REFERENCES

[1] World Health Organization, "Deafness and hearing loss," 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

[2] Y. Feng and F. Chen, "Nonintrusive objective measurement of speech intelligibility: A review of methodology." *Biomedical Signal Processing and Control*, vol. 71, no. 103204, 2022.

[3] 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.

[4] T. Houtgast and H. Steeneken, "A physical method for measuring speech-transmission quality," *The Journal of the Acoustical Society of America*, vol. 67, pp. 318–326, 1980.

[5] ANSI, *American National Standard Methods for Calculation of the Speech Intelligibility Index.* Acoustical Society of America, 1997.

[6] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI)," *Speech Communication*, vol. 65, pp. 75–93, 2014.

[7] J. Barker, M. A. Akeroyd, T. J. Cox, J. F. Culling, J. Firth, S. Graetzer, H. Griffiths, L. Harris, R. V. Muñoz, G. Naylor, Z. Podwińska, and E. Porter, "The 1st Clarity Prediction Challenge: A machine learning challenge for hearing aid intelligibility prediction." in *Proceedings of Interspeech*, 2022, pp. 3508–3512.

[8] M. A. Akeroyd, W. Bailey, J. Barker, T. J. Cox, J. 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 *Proceedings of ICASSP*. IEEE, 2023.

[9] S. Graetzer, M. A. Akeroyd, J. Barker, T. J. Cox, J. F. Culling, G. Naylor, E. Porter, and R. V. Muñoz, "Dataset of British English speech recordings for psychoacoustics and speech processing research: The Clarity speech corpus," *Data in Brief*, vol. 41, no. 107951, Apr. 2022.

[10] D. Schröder and M. Vorländer, "RAVEN: A real-time framework for the auralization of interactive virtual environments," in *Forum Acusticum*, Denmark: Aalborg, 2021, pp. 1541–1546.

[11] F. Denk, S. M. Ernst, J. Heeren, S. D. Ewert, and B. Kollmeier, "The Oldenburg Hearing Device (OlHeaD) HRTF Database," University of Oldenburg, Tech. Rep., 2018.

[12] 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 heariing aid processing," in *Proceedings of Interspeech 2021*, Aug. 2021, pp. 686–690.

[13] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[14] T. Irino, "Hearing impairment simulator based on auditory excitation pattern playback: WHIS," *IEEE Access*, vol. 11, pp. 78 419–78 430, 2023.

[15] N. Mamun, W. A. Jassim, and M. S. A. Zilany, "Prediction of speech intelligibility using a neurogram orthogonal polynomial measure (NOPM)," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 760–773, 2015.

[16] V. Best, C. R. Mason, J. Kidd, Gerald, N. Iyer, and D. S. Brungart, "Better-ear glimpsing in hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 137, no. 2, pp. EL213–EL219, 02 2015. [Online]. Available: https://doi.org/10.1121/1.4907737

[17] R. Wan, N. I. Durlach, and H. S. Colburn, "Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers," *The Journal of the Acoustical Society of America*, vol. 128, no. 6, pp. 3678–3690, 12 2010. [Online]. Available: https://doi.org/10.1121/1.3502458

[18] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[19] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: a British English speech corpus for large vocabulary continuous speech recognition," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1995, pp. 81–84 vol.1.

[20] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1–14, 10 2022.

[21] S. Cuervo and R. Marxer, "Temporal-hierarchical features from noise-robust speech foundation models for non-intrusive intelligibility prediction," in *Proc. ISCA Clarity-2023*, Dublin, Ireland, Aug. 2023.

[22] R. Mogridge, G. Close, R. Sutherland, S. Goetze, and A. Ragni, "Pre-trained intermediate ASR features and Human memory simulation for non-intrusive speech intelligibility prediction in the Clarity Prediction Challenge 2," in *Proc. ISCA Clarity-2023*, Dublin, Ireland, Aug. 2023.

[23] 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*, Dublin, Ireland, Aug. 2023.

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

[25] R. E. Zezario, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Deep learning-based speech intelligibility prediction model by incorporating Whisper for hearing aids," in *Proc. ISCA Clarity-2023*, Dublin, Ireland, Aug. 2023.

[26] 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*, Dublin, Ireland, Aug. 2023.

[27] K. Yamamoto, "A non-intrusive binaural speech intelligibility prediction for Clarity-2023," in *Proc. ISCA Clarity-2023*, Dublin, Ireland, Aug. 2023.

[28] 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*, Dublin, Ireland, Aug. 2023.