DOI: 10.61782/fa.2025.0287

DEVELOPING A NOVEL MULTIMODAL SPEECH ENHANCEMENT INTELLIGIBILITY EVALUATION METRIC: ADDRESSING THE LIMITATIONS OF TRADITIONAL OBJECTIVE MEASURES

Adeel Hussain[1*] Mandar Gogate[1] Kia Dashtipour[1]
Nasir Saleem[1] Adele Goman[1] Arslan Tughrul[2]
Aziz Sheikh[3] Amir Hussain[1]
[1]Edinburgh Napier University, 10 Colinton Road, Edinburgh
[2]School of Engineering, The University of Edinburgh, Edinburgh
[3]Nuffield Dept of Primary Care Health Sciences, Radcliffe Observatory Quarter

ABSTRACT

The evaluation of speech intelligibility is crucial for optimising speech-based systems. Existing objective metrics primarily focus on acoustic analysis, often neglecting the audiovisual (AV) nature of speech. To address this limitation, this study proposes AVIntell, a deep learning-based model that integrates subjective intelligibility AV data for intelligibility prediction. In addition, we introduce NAPE-AV, a novel dataset specifically designed for the assessment of AV intelligibility. The model uses the complementary strengths of convolutional neural network (CNN) and long short term memory (LSTM) to predict speech intelligibility by comparing processed audio with reference speech. Experimental results demonstrate a strong correlation with human perceptual scores, surpassing the state-of-the-art speech intelligibility metrics including STOI and MOSA Net+ across all evaluation metrics. These findings confirm the advantages of integrating AV intelligibility data collected for a more accurate and robust assessment of speech intelligibility.

Keywords: *intelligibility assessment, audiovisual, self-supervised learning, neural network*

1. INTRODUCTION

Speech intelligibility assessment metrics are critical for optimising the performance of wide range of speech-based communication systems. Intelligibility is typically defined as the proportion of correctly identified words within a given sentence [1]. These metrics can be broadly categorised into subjective and objective metrics. Subjective evaluation involves human listeners to assess how well speech can be understood, typically through structured listening tests. These tests are widely recognised as the gold standard for assessing speech intelligibility [2–4]. While these tests yield highly accurate measurements, they necessitate the involvement of trained personnel and are often time-consuming [3, 4]. To overcome the challenges, studies have developed complementary assessment approaches combining both subjective and objective methods. Subjective evaluation involves human listeners who assess how well speech can be understood, typically through structured listening tests. Evaluation metrics have been developed to estimate intelligibility automatically, enabling efficient and scalable assessment without the need for human listeners.

Objective intelligibility metrics can be further categorised into intrusive [5] and non-intrusive [6]. Intrusive metrics require a clean reference signal, whereas non-intrusive metrics do not require a reference signal. Although non-intrusive methods are more practical for real-world scenarios, their generalisation abilities are limited compared to intrusive approaches [7]. However, it is im-

*[*Corresponding author*: adeel.hussain@napier.ac.uk.]*

Copyright: *©2025 Hussain et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.*

portant to note that both intrusive and non-intrusive metrics typically assess only the audio modality, whereas real-world communication is inherently audiovisual (AV), which may impact perceived speech quality and intelligibility.

In the literature, extensive work has been done to develop speech evaluation metrics that aim to correlate with human perceptual judgments [1, 8]. However, these metrics face two primary limitations. Firstly, they have not been extensively validated across diverse datasets beyond their training sets, raising concerns about their generalisability. Secondly, they rely exclusively on audio-only (AO) data, despite real-world speech perception being inherently multimodal, incorporating both auditory and visual cues. This limitation reduces the ecological validity of these metrics, as they fail to capture the role of visual articulatory cues in speech comprehension, which is particularly beneficial for both normal-hearing and hearing-impaired individuals [9]. Integrating AV human intelligibility data into model training significantly enhances the robustness and real-world applicability of these metrics.

This study introduces an end-to-end neural network architecture designed to predict speech intelligibility scores by jointly analysing processed and reference speech signals. The model is built with CNN and LSTM to exploit the spatio-temporal nature of the input data. The proposed model is evaluated against state-of-the-art objective metrics, including STOI, eSTOI, and MOSA-Net+. MOSA-Net+ is a non-intrusive model whereas the proposed model follows an intrusive approach. Comparing both intrusive and non-intrusive methods ensures a comprehensive assessment of intelligibility scores across different evaluation paradigms.

Furthermore, many existing intelligibility metrics have not been extensively validated on datasets beyond those originally used for their development. The predominant reliance on AO datasets fails to capture real-world conditions where visual cues, such as the facial expressions of the speaker, significantly improve intelligibility [9]. When comparing AV subjective data to current objective metrics, a notable discrepancy emerges, highlighting the limitations of traditional approaches. This study aims to bridge this gap by developing a metric that integrates both AO and AV modalities, thereby improving ecological validity and the accuracy of intelligibility assessments.

The key contributions of this study are as follows:

- We develop a novel dataset containing both AO and AV subjective speech intelligibility evaluations. The dataset incorporates clean, noisy, and enhanced speech samples, with enhancement performed using two state-of-the-art systems: an AO denoiser [10] and an AV speech enhancement model [11]. This collection represents the first intelligibility dataset featuring these specific conditions while utilising a balanced British English speech corpus.
- We propose a novel hybrid deep learning framework for intelligibility prediction. This innovative approach is trained and evaluated using first of its kind AV subjective intelligibility data, combining the strengths of multiple neural architectures to achieve robust performance.
- We perform extensive evaluation of our proposed approach with state-of-the-art objective evaluation metrics including STOI, eSTOI and MOSA Net+.

2. RELATED WORK

The assessment of intelligibility metrics has a long history, dating back to the 1940s with the development of the articulation index [12]. Subsequent advancements in experimental methodologies led to significant refinements of articulation index, culminating in the speech intelligibility index (SII) [13]. Both articulation index and SII quantify speech intelligibility by analysing the contribution of different frequency bands to overall understanding. These contributions are modelled as a function of the signal-to-noise ratio (SNR) within each band, where higher SNR values indicate better speech perception. SII, in particular, incorporates weighting factors to account for the varying importance of different frequency bands in human speech perception, offering a more refined and standardised approach to intelligibility assessment. An intelligibility score is then calculated by taking a weighted average across frequency bands, which has been shown to correlate well with subjective intelligibility for stimuli degraded by additive noise. The speech transmission index (STI) is another well-known intelligibility metric widely used for assessing speech intelligibility in various acoustic environments [14]. The STI methodology is similar to the SII, as both are based SNRs across multiple frequency bands. However, in the STI framework, the SNR for each frequency band is specifically related to the reduction of amplitude modulations caused by the transmission system. The SII and STI are widely established metrics used by researchers and audiologists, yet they present significant limitations. First, both metrics, being based on long-term statistics, fail to accurately account for degradations caused by non-linear, time-varying noise sources such as competing talkers and wind [15]. Second, neither metric adequately addresses distortions introduced by speech enhancement algorithms [16]. Third, these models are trained exclusively on audio, limiting their ability to incorporate multimodal contextual cues that could enhance speech quality and intelligibility assessment.

To address these limitations of the SII and STI, researchers have developed range of intelligibility metrics targeting specific types of signal degradation. These metrics consider various forms of distortion, including additive noise, reverberation, auditory thresholds, bandwidth reduction, interframe transitions (IFTs), environmental noise, and linear distortions. While each proposed metric demonstrates particular strengths in addressing specific aspects of speech degradation, a comprehensive solution remains elusive. The well-known intelligibility metrics that overcome some of these limitations include the coherence SII [17], the Extended SII (ESII) [15], the Quasi-Stationary STI (QSTI) [18], the Normalised Covariance Measure (NCM) [19], the Temporal Fine-Structure Spectrum-based Index (TFSS) [20], the Hearing-Aid Speech Perception Index (HASPI) [21], the Christiansen-Pedersen-Dau metric (CPD) [22], the Short-Time Objective Intelligibility (STOI) [23] and Extended STOI (eSTOI) [24].

In addition, the following metrics are described in detail, as they were used for comparative evaluation of the proposed model.

STOI is considered a benchmark objective intelligibility metric due to its strong correlation with the results of the subjective listening test [23]. STOI calculates temporal envelopes from both clean and modified speech, producing values between 0 and 1, where 1 represents perfect intelligibility. The metric uses a time-frequency (T-F) dependent intermediate intelligibility measure that decomposes signals into T-F regions, followed by energy clipping and normalisation. Intelligibility predictions are derived from cross-correlations between processed and clean signals across different T-F cells.

eSTOI [24] was developed to address the limitations in STOI, specifically its poor performance with modulated noise sources, such as amplitude-modulated Gaussian noise. While STOI computes correlation between clean and modified envelopes over short time segments, eSTOI operates in spectral domain, enabling more effective identification of clean speech glimpses. Moreover, the clipping procedure was eliminated to enhance mathematical tractability. The implementation of the eSTOI metric used in this study was gained from the developers official repository.

MOSA-Net+ [25] is a deep neural network (DNN) non-intrusive multi-objective speech assessment framework that incorporates whisper, a large-scale weakly supervised model, to generate robust acoustic features for speech quality and intelligibility prediction. The architecture employs cross-domain features from three distinct sources: (1) traditional spectral features extracted via Short-Time Fourier Transform (STFT), (2) waveforms processed using adaptable filters from a convolutional network, and (3) latent representations generated by whisper.

Overall, STOI and eSTOI remain widely used benchmarks for objective intelligibility assessment, demonstrating strong correlations with subjective listening tests. While these intrusive methods offer reliable performance, they rely on access to a clean reference signal, limiting their applicability in real-world scenarios. In contrast, MOSA-Net+ represents a more recent advancement in non-intrusive intelligibility assessment, leveraging deep learning and large-scale pre-trained models to improve generalisation across diverse conditions. However, despite these developments, existing intelligibility metrics remain constrained by their reliance on AO data, overlooking the multimodal nature of speech perception. This limitation highlights the need for novel approaches that integrate AV information to enhance the ecological validity of intelligibility assessments. The proposed model in this study aims to address these gaps by incorporating AV-based speech intelligibility data, offering a more comprehensive and robust evaluation framework.

3. PROPOSED FRAMEWORK

3.1 Framework DNN architecture

Figure 1 presents a novel intelligibility framework. The proposed framework was evaluated using a trained from scratch model. The proposed intelligibility framework builds upon the architecture of InQSS [26], a non-intrusive quality assessment method. This framework was further refined and trained using a novel NAPE-AV Speech Intelligibility (NAPE-AV) dataset, which was specifically collected to facilitate human speech intelligibility assessment. The result is an intrusive metric that leverages both audio and visual modalities to evaluate speech intelligibility.

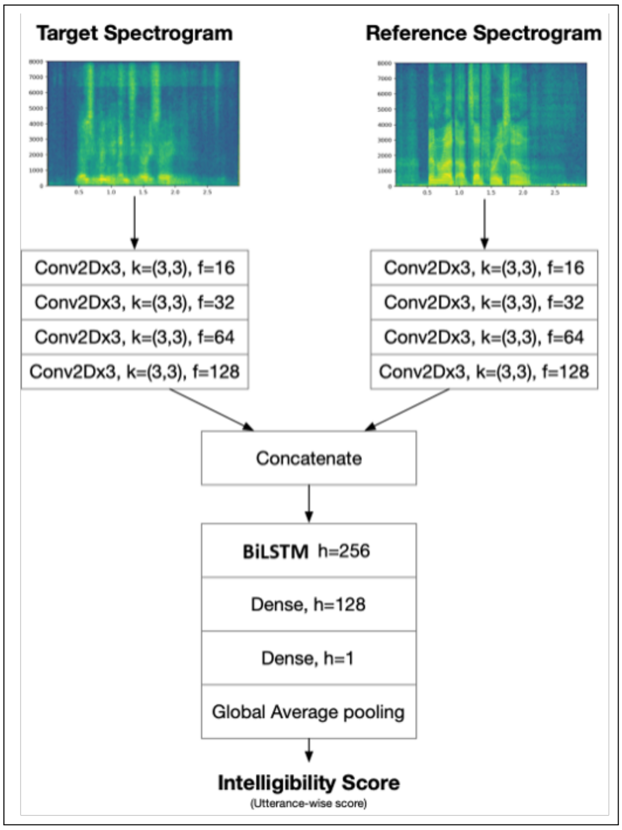

Figure 1. Model structure of the DNN framework.

The input to the framework consists of two spectrograms: the target (S_T) and the reference (S_R), each represented as matrices of dimensions (H×W), where H denotes the height (representing time steps) and W represents the width (representing frequency bins). These spectrograms are then passed through a series of convolutional layers for feature extraction. Specifically, each spectrogram undergoes four 2D convolution operations, with kernel size K = (3×3) and increasing filter counts f=16, 32, 64, 128. The output of each convolutional layer generates feature maps, where the number of filters increases at each layer. After the feature extraction step, the features from the target and reference spectrograms are concatenated into a single feature vector. The concatenated features are then processed by a Bidirectional Long Short-Term Memory (BiLSTM) layer with 128 hidden units, which captures temporal dependencies and contextual information.

Finally, the processed features pass through two fully connected dense layers. The first dense layer has 128 units, while the second has a single unit, producing the final output, which is an intelligibility score. Before the final classification, a global average pooling operation is applied to reduce the dimensionality of the feature maps, allowing the model to output a scalar value representing the utterance-wise intelligibility score.

4. EXPERIMENTS

In this section, we evaluate the proposed approach using IEEE sentences dataset.

4.1 Dataset

The dataset utilised in this study is based on the British IEEE sentences [27], comprising 72 lists of 10 phonetically balanced and homogeneously structured utterances. The utterances were recorded by a male speaker for both testing and training sets. The video data captured using an iPad Pro (12.9-inch, 5th generation) at 30 fps in 4K resolution, while audio was recorded via the omnidirectional lavalier microphone (Zoom F2) attached to the speaker's collar. In order to create noisy conditions, a male multi-talker babble was generated using IEEE sentence lists recorded by four different male speakers, which were then randomly mixed. All recordings, including the clean utterances and the multi-talker babble, were conducted in the auralisation suite at Edinburgh Napier University [28]. Each sentence in the dataset is evaluated based on five keywords per trial.

The SNR for the noisy utterances was randomly selected between the range of 20 dB to -20 dB. These noisy utterances were processed using two different SE models for the AO state of the art Facebook denoiser [10] and for the AV [11] was used. Finally, the clean, noisy and enhanced utterances were combined to form the listening test utterance pool. In this study, the proposed assessment model was evaluated using the NAPE-AV dataset. The dataset comprised a total of 4,920 utterances, with 984 samples designated for testing and the remaining 3,936 samples used for training and validation.

4.2 Listening test for development of the Dataset

The speech in noise tests were performed on 36 participants (18 females, 18 males), with ages ranging from 21 to 93 years. The participants were divided into two groups; group 1 consisted of 18 participants within normal limits of hearing (9 females, 9 males), group 2 consisted of a further 18 participants with hearing losses ranging from mild to severe (9 females, 9 males). The test for individuals with normal hearing was performed in noisy and enhanced conditions. To gain the intelligibility score the participants were instructed to identify five keywords in each sentence, with the intelligibility score being calculated as the number of correctly identified keywords, out of five per utterance across random SNRs. For the hearing loss group the noisy condition was performed in unaided and aided conditions as well as aided in enhanced condition. Both groups were provided a practice list in both AO and AV conditions to familiarise themselves with the test set up. These utterances where excluded from the data sample used in the model. For the testing 30 sentences were used for each condition. This totalled 4920 samples for our dataset.

4.3 Model Training

The model was implemented using PyTorch and trained on a system comprising an Intel i9 processor, 64 GB RAM, and dual NVIDIA RTX 2080 Ti GPUs (12 GB VRAM each). The model is trained over 20 epochs and optimised by ADAM. The Mean Squared Error (MSE) loss function is used during model training to measure the average squared difference between the predicted values and the actual target values.

4.4 Data pre-processing

The AVintell model employs a systematic preprocessing pipeline to standardise input audio data. All audio files are resampled to 16 kHz and normalised to a fixed duration of six seconds, ensuring temporal consistency across samples. Raw intelligibility scores, originally rated on a 1–5 scale, are normalised to a continuous 0–1 range while preserving intermediate values (e.g., 0.2, 0.4). This design choice aligns with AVintell's formulation as a regression model rather than a classification system, enabling it to predict intelligibility along a continuous spectrum. To extract time-frequency representations, the preprocessing pipeline applies a Short-Time Fourier Transform (STFT) with a 512-point FFT, a 256-sample hop length, and a Hamming window. The resulting magnitude spectrograms are normalised before being reshaped to match the model's input format.

5. RESULTS & DISCUSSION

To evaluate the proposed AVIntell model, we adopted three evaluation metrics: MSE, Linear Correlation Coefficient (LCC), and Spearman's Rank Correlation Coefficient (SRCC). Lower MSE scores indicate that the predicted scores are closer to the ground-truth assessment scores, whereas higher LCC and SRCC values (ranging from -1 to 1) indicate stronger correlations between predicted and ground-truth assessment scores.

Table 1 shows a comparative analysis of AV Intell against conventional intrusive intelligibility metrics (STOI and eSTOI), a state-of-the-art non-intrusive deep neural network model (MOSA-Net+), and human intelligibility scores. We evaluated all models using LCC and SRCC, with subjective intelligibility scores serving as the reference standard. As shown in Table I, the proposed AV Intell model demonstrates the highest alignment with human perception, achieving the best correlation scores (LCC = 0.426, SRCC = 0.438). This indicates that it consistently predicts intelligibility in a manner that reflects human evaluation more accurately than competing models.

In contrast, traditional objective metrics such as STOI and eSTOI exhibit weaker correlations with human scores. STOI achieves LCC = 0.298 and SRCC = 0.305, while eSTOI shows a marginal improvement in SRCC (0.321) but slightly lower LCC (0.296), suggesting that both fail to capture the nuances of human intelligibility perception effectively. Although MOSA-Net+ attains the highest predicted intelligibility score (74.9%), its correlation values (LCC = 0.307, SRCC = 0.307) remain comparable to those of STOI, indicating limited improvement in perceptual alignment. Furthermore, its higher MSE (0.155) compared to AV Intell suggests less precise prediction capability.

The strong correlation scores achieved by AV Intell, coupled with its closely matched predicted score (58.041%) to the human intelligibility score (58.05%), highlight the effectiveness of incorporating visual cues. These results underscore the limitations of acoustic-only approaches and the need for multimodal strategies to improve intelligibility prediction. The improvements in LCC and SRCC demonstrate the proposed model's superior ability to model human perception, addressing key shortcomings of existing intelligibility metrics.

Figure 2 illustrates the relationship between intelligibility scores and SNRs. The observed trend shows that the intelligibility improves as SNR increases, confirming that higher SNR conditions lead to clear speech perception. At low SNR levels (-20 to -10 dB), intelligibility remains relatively low ($\sim$0.3 to 0.5) primarily due to significant presence of background noise. Since different listeners have

Table 1. Comparison of the intelligibility metrics. (Avg Intell indicates average intelligibility)

Model	Avg Intell	LCC	SRCC	MSE
Subjective	0.580	-	-	-
STOI	0.469	0.298	0.305	0.147
*e*STOI	0.337	0.296	0.321	0.191
MOSA-Net+	0.749	0.307	0.307	0.155
AV Intell	**0.580**	**0.426**	**0.438**	**0.107**

diverse hearing profiles, their perception of intelligibility across SNR levels differs. Individuals with hearing impairments face greater challenges at low SNRs, as they face challenges to distinguish speech from noise. Age-related declines in auditory processing also affect some listeners. As SNR increases to moderate levels (-10 to 0 dB), intelligibility steadily improves, and at higher SNRs (5 to 15 dB), it peaks (∼0.7–0.8), with minimal differences between actual and predicted intelligibility values. The predicted intelligibility scores closely match the actual scores, with a minimum difference of 0.02 and maximum of 0.08 indicating that the model effectively estimates intelligibility across various SNRs.

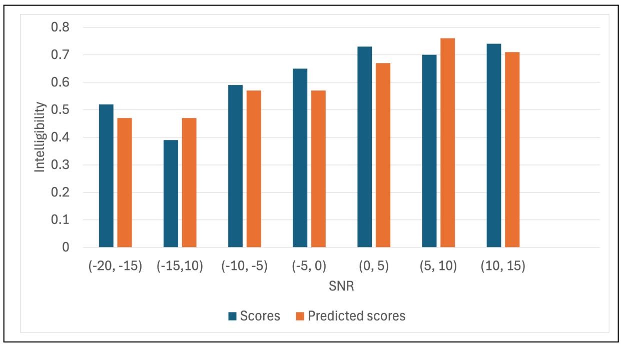

Figure 2. A bar graph showing the model predicted and human scores at the various SNR ranges

Figure 3 shows an analysis of MSE, LCC, and SRCC across various SNR levels, demonstrating the impacts of noise conditions on intelligibility prediction accuracy.

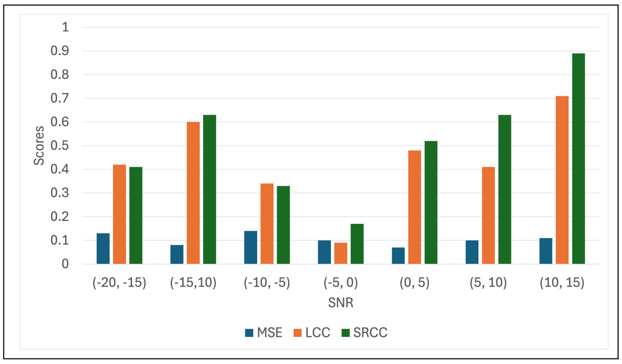

Figure 3. A bar graph showing the comparison of model evaluation metrics at the various SNR ranges

The highest SRCC value (0.89) is observed in the (10, 15) dB range, indicating a strong rank correlation under high-SNR conditions. In contrast, the lowest SRCC value (0.17) occurs in the (-5, 0) dB range, similar to the LCC trend, suggesting a weak rank-based correlation in this region. SRCC values remain relatively stable in the (-20, -15) dB (0.41) and (5, 10) dB (0.63) ranges, reflecting a moderate correlation under mid-range SNR conditions. Similarly, the highest LCC value (0.71) is found in the (10, 15) dB range, demonstrating strong linear correlation when noise is minimal. Conversely, the lowest LCC (0.09) occurs in the (-5, 0) dB range, implying that predictions in this SNR range are less aligned with the ground truth. Moderate LCC values in the (-20, -15) dB (0.42) and (5, 10) dB (0.41) ranges further indicate a partial correlation between predicted and actual values under these conditions.

Since for human listeners with hearing loss and different hearing profiles, the low correlation between predicted and actual intelligibility scores at lower SNRs is attributed to the compounded effect of both noise and auditory processing discrepancies. Individuals with hearing impairments often have a reduced ability to distinguish speech from background noise, especially in noisy environments where the SNR is low. This makes it challenging for them to perceive key speech cues, further deteriorating speech intelligibility [29]. As a result, even if predictive models capture some speech features, they fail to accurately reflect the true intelligibility as experienced by listeners with hearing loss, leading to lower correlation values in these conditions.

Another factor contributing to the overall lower LCC and SRCC results across all SNR ranges, compared to other models, is the diversity of conditions present in our dataset. Typically, models are trained under a single condition; however, our dataset incorporates multiple conditions, including noisy, aided noisy, enhanced, and aided enhanced environments. Additionally, our data encompasses a wide range of hearing abilities, from individuals with normal hearing to those with severe hearing loss. This increased variability in both noise conditions and hearing profiles introduces more complexity, which may explain the lower correlation scores. Nevertheless, despite these lower scores, the model more accurately reflects real-world listening scenarios, making it highly valuable for a broad range of testing protocols in future research and applications.

Across the SNR range, the MSE values remain consistently low, indicating strong model performance. The lowest MSE (0.07) is observed in the (0,5) dB range, suggesting that the model achieves optimal performance at this level. One possible reason for this could be the higher representation of data within these SNR ranges, allowing the model to better learn the underlying trends. Conversely, the highest MSE values (0.14 and 0.13) occur in the (-20, -15) dB and (-10, -5) dB ranges, where noise conditions are more severe, leading to reduced model performance. However, despite the increased noise, the MSE values remain relatively low, indicating that the model maintains a reasonable level of accuracy even in challenging conditions.

6. CONCLUSIONS

This study introduces AVIntell, a novel deep learning-based model for speech intelligibility prediction that integrates AV speech intelligibility data. Comparative experimental results demonstrates strong alignment of AVIntell with human perceptual scores, outperforming traditional audio-only intelligibility metrics such as STOI and eSTOI and deep learning based metrics including MOSA-Net+ across all evaluation metrics. The inclusion of AV speech intelligibility data in AVIntell substantially enhances intelligibility predictions, particularly in noisy environments where traditional acoustic-based methods often struggle. In conclusion, AVIntell introduces a novel intelligibility prediction model and highlights the potential for multimodal integration to significantly enhance the robustness and accuracy of speech intelligibility prediction systems. Ongoing work include development of a non-intrusive objective intelligibility measure based on AVIntell. In future, we intend to incorporate visual cues alongside audio inputs to develop a fully multimodal speech intelligibility metric.

7. ACKNOWLEDGMENTS

This research is supported by the UK EPSRC COG-MHEAR programme (Grant No. EP/M026981/1).

8. REFERENCES

[1] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An evaluation of intrusive instrumental intelligibility metrics," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2153–2166, 2018.

[2] H.-T. Chiang, K.-H. Hung, S.-W. Fu, H.-C. Kuo, M.-H. Tsai, and Y. Tsao, "Study on the correlation between objective evaluations and subjective speech quality and intelligibility," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–7, IEEE, 2023.

[3] 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," *arXiv preprint arXiv:2204.03305*, 2022.

[4] P. C. Loizou, "Speech quality assessment," in *Multimedia analysis, processing and communications*, pp. 623–654, Springer, 2011.

[5] M. B. Pedersen, A. H. Andersen, S. H. Jensen, and J. Jensen, "A neural network for monaural intrusive speech intelligibility prediction," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 336–340, IEEE, 2020.

[6] 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 *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 482–486, IEEE, 2020.

[7] Y. Feng and F. Chen, "Nonintrusive objective measurement of speech intelligibility: A review of methodology," *Biomedical Signal Processing and Control*, vol. 71, p. 103204, 2022.

[8] K. Shen, D. Yan, J. Hu, and Z. Ye, "Non-intrusive speech quality assessment: A survey," *Neurocomputing*, vol. 580, p. 127471, 2024.

[9] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *The journal of the acoustical society of america*, vol. 26, no. 2, pp. 212–215, 1954.

[10] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," *arXiv preprint arXiv:2006.12847*, 2020.

[11] M. Gogate, K. Dashtipour, and A. Hussain, "Towards robust real-time audio-visual speech enhancement," *arXiv preprint arXiv:2112.09060*, 2021.

[12] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *The journal of the Acoustical society of America*, vol. 19, no. 1, pp. 90–119, 1947.

[13] A. N. S. Institute, *American National Standard: Methods for Calculation of the Speech Intelligibility Index*. Acoustical Society of America, 1997.

[14] H. J. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *The Journal of the Acoustical Society of America*, vol. 67, no. 1, pp. 318–326, 1980.

[15] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2181–2192, 2005.

[16] C. Ludvigsen, C. Elberling, and G. Keidser, "Evaluation of a noise reduction method–comparison between observed scores and scores predicted from sti.," *Scandinavian audiology. Supplementum*, vol. 38, pp. 50–55, 1993.

[17] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," *The journal of the acoustical society of America*, vol. 117, no. 4, pp. 2224–2237, 2005.

[18] B. Schwerin and K. Paliwal, "An improved speech transmission index for intelligibility prediction," *Speech Communication*, vol. 65, pp. 9–19, 2014.

[19] R. L. Goldsworthy and J. E. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *The Journal of the Acoustical Society of America*, vol. 116, no. 6, pp. 3679–3689, 2004.

[20] F. Chen, L. L. Wong, and Y. Hu, "A hilbert-fine-structure-derived physical metric for predicting the intelligibility of noise-distorted and noise-suppressed speech," *Speech Communication*, vol. 55, no. 10, pp. 1011–1020, 2013.

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

[22] C. Christiansen, M. S. Pedersen, and T. Dau, "Prediction of speech intelligibility based on an auditory preprocessing model," *Speech Communication*, vol. 52, no. 7-8, pp. 678–692, 2010.

[23] 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.

[24] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[25] R. E. Zezario, Y.-W. Chen, S.-W. Fu, Y. Tsao, H.-M. Wang, and C.-S. Fuh, "A study on incorporating whisper for robust speech assessment," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, 2024.

[26] Y.-W. Chen and Y. Tsao, "Inqss: a speech intelligibility and quality assessment model using a multi-task learning network," *arXiv preprint arXiv:2111.02585*, 2021.

[27] E. H. Rothauser, "IEEE recommended practice for speech quality measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, 1969.

[28] E. Prokofieva, C. Luciani, and I. McGregor, "Design and Development of Auralization Room at Edinburgh Napier University," in *Audio Engineering Society Convention 136*, Audio Engineering Society, 2014.

[29] E. W. Healy and S. E. Yoho, "Difficulty understanding speech in noise by the hearing impaired: underlying causes and technological solutions," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 89–92, IEEE, 2016.