2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) | 979-8-3503-0067-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/APSIPAASC58517.2023.10317260

Incorporating the Digit Triplet Test in A Lightweight Speech Intelligibility Prediction for Hearing Aids

Xiajie Zhou, Candy Olivia Mawalim, Benita Angela Titalim, Masashi Unoki
* Japan Advanced Institute of Science and Technology, Japan
1-1 Asahidai, Nomi, Ishikawa 923–1292 Japan
E-mail: {s2210112, candylim, titalim, unoki}@jaist.ac.jp

***Abstract*—Recent studies in speech processing often utilize sophisticated methods for solving a task to obtain high-accuracy results. Although high performance could be achieved, the methods are too complex and require high-performance computational power that might not be available for a wide range of researchers. In this study, we propose a method to incorporate the low dimensional and the recent state-of-the-art acoustic features for speech processing to predict the speech intelligibility in noise for hearing aids. The proposed method was developed based on the stack regressor on various traditional machine learning regressors. Unlike other existing works, we utilized the results of the digit triplet test, which is usually used to measure the hearing ability in the existence of noise, to improve the prediction. The evaluation of our proposed method was carried out by using the first Clarity Prediction Challenge dataset. This dataset is utilized for speech intelligibility prediction that consists of speech signals output of hearing aids that were arranged in various simulated scenes with interferers. Our experimental results show that the proposed method could improve speech intelligibility prediction. The results also show that the digit triplet test results are beneficial for speech intelligibility prediction in noise.**

I. Introduction

Hearing loss is a widespread and significant problem that affects people's lives. With age and other factors, an individual's hearing function gradually deteriorates, decreasing speech comprehension and communication skills [1], [2]. Hearing loss directly impacts the clarity of an individual's speech. Hearing loss at specific frequencies makes it difficult for individuals to discriminate certain speech sounds, reducing speech intelligibility and comprehension [3]. This effect is further exacerbated by background noise and complex listening environments [4], [5]. As individuals' hearing function gradually deteriorates, they often suffer from physical burdens, further exacerbated by emotional depression due to the inability to communicate effectively with others.

A significant goal of hearing aids is to improve the speech intelligibility of individuals with hearing loss to better participate in social activities, communicate with others, and receive and understand verbal information effectively. Evaluation of hearing abilities takes place through various methods such as audiograms, speech tests, auditory questionnaires, auditory neurophysiological measures, and cognitive tests [6]–[8]. Existing methods for speech intelligibility prediction in hearing aids often rely only on audiograms to represent the status of hearing loss, which has certain limitations.

The traditional audiogram assessments only provide information about hearing sensitivity and frequency characteristics. In contrast, speech comprehension involves more complex auditory and cognitive processes, and audiograms do not fully reflect changes in these aspects [9]. Moreover, audiograms do not provide information about an individual's performance in a given speech environment, which is essential when assessing the speech intelligibility of hearing aids. The effects of environmental noise, speech intelligibility, and individual auditory processing abilities prevent conventional audiograms from providing detailed information on these aspects. To obtain more comprehensive and accurate hearing assessment results in assessing the speech intelligibility of hearing aids, we need to combine other assessment methods, such as the digit triplet test (DTT) as a method of assessing hearing ability in noise.

The DTT has essential applications in assessing speech perception in complex listening environments, such as background noise or competing sounds [10], [11]. The test uses digit triplets to measure an individual's ability to perceive and discriminate speech sounds in the presence of hearing loss, especially in the high-frequency range [12]. By measuring the speech reception threshold (SRT) obtained from the DTT, we can reveal an individual's ability to comprehend speech under adverse conditions and correlate it with the overall speech intelligibility in the real world. Lower SRT values indicate that individuals are more capable of successfully recognizing and discriminating speech in the DTT, which implies a higher level of discourse intelligibility. In addition, the DTT provides practical guidance for hearing aid assessment, selection, and individualized rehabilitation planning by measuring an individual's ability to perceive and discriminate speech at lower levels. It can also predict their overall speech intelligibility in the real world.

The purpose of this paper is to propose a speech intelligibility prediction method that considers the results of the DTT for hearing aids in complex listening environments. Our approach builds on a stacked regressor that combines the functionality of traditional machine learning regressors, including linear regressor, support vector regressor, and random forest regressor. Since the DTT can assess the hearing ability in the presence of noise, we hypothesize that incorporating the results of the DTT into our prediction model has the potential to improve the prediction further.

The paper next consists of five sections. Section 2 includes details of the related work, i.e., hearing condition assessment, baseline of the modified Binaural Short-time Objective Intelligibility (MBSTOI), and speech intelligibility prediction.

TABLE I
SUMMARY OF SPEECH INTELLIGIBILITY ASSESSMENT METHODS

Assessment Methods	Phoneme Tests	Word Recognition Tests	Sentence Intelligibility Tests	Digit Triplet Test
Purpose	Assess discrimination of phonemes, syllables	Evaluate recognition of individual words	Measure understanding of sentences	Assess speech perception in connected speech
Response	Discrimination or identification	Identification or repetition	Repetition or understanding	Identification or repetition
Stimuli	Individual speech sounds or minimal pairs	Isolated words	Sentences	Triplet of connected speech sounds
Types of Tests	Phoneme Discrimination Tests, Phoneme Recognition Tests	Consonant-Vowel Tests	Speech Intelligibility Index, Hearing in Noise Test	Digit Triplet Test
Measurement Outcome	Discriminative thresholds, discrimination accuracy	Word recognition score	Threshold level for speech intelligibility	Accuracy or threshold level for speech intelligibility

Section 3 presents the proposed method, i.e., feature extraction and regression analysis. Section 4 describes the experiments, such as the dataset, evaluation metrics, and results. Section 5 shows the conclusion of our study.

II. Related Work

A. Hearing Condition Assessment

An audiogram measures the sensitivity of an individual's hearing at different frequencies. The threshold of the different sound intensities they can hear is measured by performing a hearing test and drawing an audiogram. The shape and results of the audiogram can provide detailed information about an individual's hearing status. A physician or audiologist can use the audiogram to help determine if there is a hearing loss and its degree and type. Recording the hearing thresholds of individuals at different frequencies plots the audiogram. These hearing thresholds represent the minimum sound intensity that an individual can hear. The pure tone average (PTA) provides the audiogram with a measure of the degree of hearing loss. It summarizes an individual's hearing sensitivity over a specific frequency range. A common way to calculate the PTA is to select frequency points. These frequency points are the most critical frequency ranges for speech and communication. A higher PTA value indicates a higher hearing threshold, i.e., an individual has a lower auditory sensitivity to sound, suggesting the presence of hearing loss.

Existing methods for assessing speech intelligibility with hearing aids include a series of tests to assess an individual's ability to perceive and understand language [11], [13]–[16]. These assessments aim to quantify the degree of speech intelligibility and provide valuable information for diagnosing hearing impairment, assessing treatment effectiveness, and guiding intervention strategies. All of the methods assess an individual's ability to perceive and comprehend language. The acceptable discrimination required by phoneme tests allows for detailed speech analysis but may only partially reflect the challenges faced in complex listening environments. The word recognition tests provide insight into recognition at the lexical level but may not reflect an understanding of communication at the sentence level. The sentence intelligibility tests present higher cognitive and linguistic demands, potentially affecting an individual's ability to understand and accurately respond to the entire sentence, which places a higher cognitive burden on individuals with hearing loss.

The DTT has demonstrated reasonable validity and reliability in assessing speech intelligibility [11]. The DTT presents a series of digraph triplets consisting of three speech sounds [10]. During the test, individuals recognized different digit triplets at different signal-to-noise ratios. By setting different signal-to-noise ratios in the test environment, the DTT provides insight into a person's ability to perceive speech in complex listening environments and helps to assess their overall communication ability more accurately. Compared to isolated phonemes, words, or sentences, the DTT uses a triplet of digital speech with specific phonetic features, and the test results more closely mimic real-life speech intelligibility.

The advantages of the DTT in complex listening environments and for individuals with hearing loss make it a promising option for assessing speech intelligibility in challenging listening conditions. The DTT has higher sensitivity and specificity than other testing methods, allowing for a more accurate assessment of an individual's ability to perceive and discriminate speech sounds. Moreover, the DTT results correlate strongly with audiogram PTA measurements, which serve as an index to assess the degree of hearing loss and reflect an individual's hearing threshold at different frequencies. The correlation with the PTA suggests that the DTT can be used as one indicator to complement the hearing status assessment, providing more comprehensive information for a hearing assessment. The DTT is the ability to capture the individual's ability to perceive the nuances of speech features, thus providing a more comprehensive assessment of the individual's speech intelligibility.

B. Baseline Modified Binaural Short-time Objective Intelligibility (MBSTOI)

The MBSTOI is an improved binaural short-time objective intelligibility index for speech intelligibility prediction in hearing aids. The short-time objective intelligibility (STOI) measures have successfully predicted the intelligibility of noisy speech processed by time-frequency (TF) weighting methods [17]. However, a limitation of STOI is its reliance on monophonic speech signals, which ignores valuable dual-channel

cues available to listeners.

The binaural short-time objective intelligibility (BSTOI) is based on the principles of STOI but introduces some modifications [18]. In BSTOI, the equalisation cancellation (EC) model uses information from binaural cues to simulate how the human auditory system works [19]. During the EC phase, the linear combination of the left and right ears is described by Eq.(1). $x_{k,m}^{(l)}$ is the signal of the TF unit corresponding to the left ear in the m-th time period and k-th frequency bin. The factor λ implements the equalization step in the EC stage.

$$\hat{x}_{k,m} = \lambda \cdot \hat{x}_{k,m}^{(l)} - \lambda^{-1} \cdot \hat{x}_{k,m}^{(r)} \tag{1}$$

The MBSTOI performs independent EC stages at each time range and each frequency band to estimate EC parameters. The MBSTOI improves binaural processing and optimized equalization stages over the BSTOI to more accurately predict speech intelligibility. Such improvements allow MBSTOI to better capture and utilize binaural information in the speech signal and improve the performance of speech intelligibility assessment.

C. Speech Intelligibility Prediction

In the field of hearing aids, predicting the speech intelligibility of listeners is an important research task. Such prediction can be achieved by objective measurements, guiding audiologists in selecting appropriate signal processing algorithms when fitting hearing aids, and providing valuable tools in developing machine learning-based hearing aid algorithms and other speech enhancement methods [20]. The objective measurement of speech intelligibility refers to analyzing various features and attributes of the speech signal. These features can include spectral characteristics of speech, time domain characteristics, and acoustic parameters. By modeling the relationship between these characteristics and listener comprehension, it is possible to predict the speech intelligibility of listeners with hearing loss under listening conditions.

The widespread use of machine learning methods in speech intelligibility prediction is due to their ability to learn patterns and correlations from large-scale speech data and apply them to new data samples for prediction. These methods use large datasets and features to train models to establish associations between features and speech intelligibility. Standard machine learning algorithms include support vector machines, random forests, and deep learning models. Through objective speech intelligibility measures and machine learning methods, audiologists and researchers can better understand listeners' speech intelligibility and select appropriate signal processing algorithms to improve listeners' listening experience.

III. Proposed Method

As the proposed model, the noisy speech after hearing aid processing is taken as input speech for feature extraction. The result of extracted speech features also combines with the bilateral audiogram and the DTT results. The aims to incorporate both audiogram and the DTT is to account for the individual listener's hearing profile and tailor the prediction accordingly. Besides, we use this holistic approach to enhance the accuracy and relevance of speech intelligibility predictions by taking into account the listener's individual auditory capability. After the feature extraction stage, the proposed model concludes with regression analyses to predict speech intelligibility.

A. Feature Extraction

In the feature extraction, we corporate various feature extraction process, including Geneva minimalistic acoustic parameter set (GeMAPS) and its extended version (eGeMAPS), wav2vec2, hidden unit bidirectional encoder representation from transformer (HuBERT), and waveform language model (wavLM). The details of each feature extraction are described in the following section.

1) GeMAPS and eGeMAPS: The open-source toolkit openSMILE incorporates to perform extraction of the GeMAPS and eGeMAPS parameter sets. Our proposed method includes these sets to extract features related to signal-to-noise ratio, spectral envelope, pitch, and temporal dynamics. The valuable information about speech characteristics might capture the relevant factors affecting speech intelligibility in the context of hearing loss perception under complex listening environments. The GeMAPS is a minimalistic parameter set that incorporating essential prosodic, excitation, vocal tract, and spectral features [21], [22]. Meanwhile, the eGeMAPS, as an extended version, includes additional cepstral features to enhance recognition accuracy beyond using only prosodic and spectral parameters.

In detail, the set of features extracted by the GeMAPS includes frequency-related, amplitude-related, and spectral features. Frequency-related features include pitch (measured in semitones), jitter (deviations in F0 period lengths), and formant frequencies. Amplitude-related features include a shimmer (difference in peak amplitudes), loudness (perceived intensity), and harmonics-to-noise ratio (HNR). Finally, Spectral features include:

- Alpha ratio: energy ratio between specific frequency ranges
- Hammarberg Index: ratio of energy peaks
- Spectral slope: regression slopes in frequency bands
- Formant relative energy
- Harmonic difference of the first harmonic to the second harmonic (H1-H2) and the third formant range (H1-A3)

These parameters undergo smoothing using a moving average filter and various functions to generate a set of parameters. These functionals include arithmetic mean, coefficient of variation, percentiles, range, and slope-related measures. Additionally, temporal features, such as the rate of loudness peaks and statistics of voiced and unvoiced regions, provide a comprehensive analysis of speech characteristics. In total, 62 parameters were extracted by GeMAPS.

As aforementioned, the extension of GeMAPS, the eGeMAPS, includes cepstral and dynamic parameters to model the affective states in speech analysis. The eGeMAPS introduces additional parameters, including Mel-frequency cepstral coefficients (MFCCs), spectral flux difference, and formant bandwidth. Functionals like arithmetic mean and coefficient

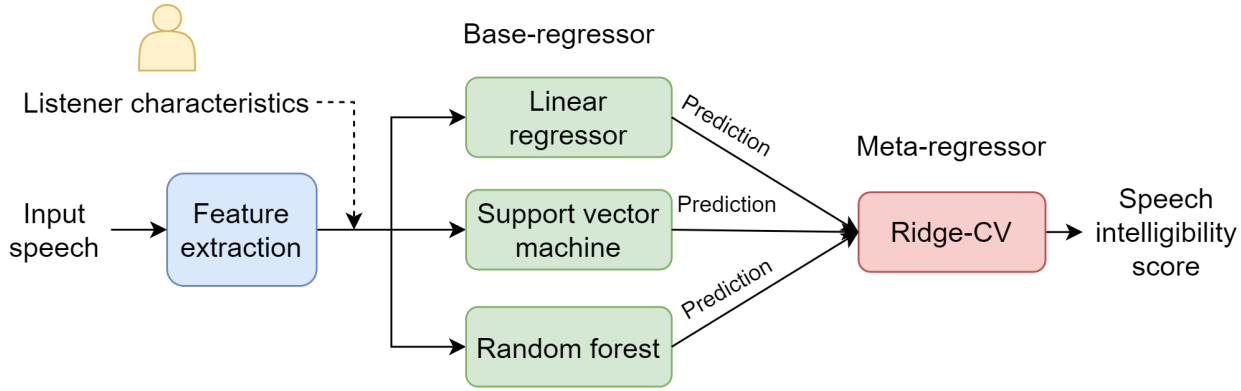

Fig. 1. Block diagram of our proposed method

of variation are applied to these parameters. The eGeMAPS set, combined with the GeMAPS, generates 88 parameters.

2) wav2vec2: The wav2vec2 is a framework of self-supervised learning in representing speech audio that is fine-tuned on transcribed speech [23]. We chose self-supervised learning because predicting speech intelligibility for hearing loss conditions using a speech recognition model only applies in real applications if it can deal with unlabeled data or explicit annotation. Besides, wav2vec2 extracts high-level acoustic features that capture important aspects of the speech signal useful for downstream tasks of predicting speech intelligibility.

The speech features extracted by wav2vec2 include the phonetic and linguistic properties by learning to model the temporal relationship between different phonemes or speech units. These representations encode information about the distinct speech and their temporal patterns, which are crucial for speech intelligibility. The wav2vec2 also captures the prosodic cues such as pitch variation, rhythm, and intonation, which play a role in conveying meaning and syntactic structure. Lastly, wav2vec2 extracts spectral information related to the frequency content, indicating phonetic distinctions and acoustic cues relevant to intelligibility.

3) HuBERT: Another self-supervised learning framework we consider is HuBERT [24]. HuBERT consists of three main components: the feature encoder, transformer encoder, and classification head, which appeared as the extension of wav2vec2. The feature encoder extracts low-level acoustic features from the input signal and converts them into time-frequency representation. Then, the transformer encoder captures the contextual information through unidirectional and bidirectional transformers. Finally, the classification head maps the acoustic features to the predicted speech intelligibility.

Although HuBERT shares the same overall architecture as the pre-trained model, HuBERT introduces enhancements and modifications to improve speech representation learning. It also utilizes larger-context modeling by increasing the context window size during pre-training to capture more long-range dependencies and contextual information in the speech signal. Unlike the unidirectional transformer layers used in wav2vec2, HuBERT incorporates both unidirectional and bidirectional transformer layers. This combination enables the model to effectively capture the local and global context in the speech signal, improve the ability to understand the intricacies of speech, and extract relevant features for speech intelligibility prediction.

4) wavLM: The last self-supervised learning model to learn the representation for speech-related tasks that we consider is the wavLM [25]. It is introduced as the extension of the HuBERT framework that enables the pre-trained model to work on speech recognition and related tasks. We include the wavLM because of the ability to extract fine-grained details within an audio signal. Besides, the wavLM trained using a language modeling objective, resulting in rich representations of speech and its linguistic properties, which can be beneficial for tasks related to speech intelligibility.

In terms of feature extraction to predict speech intelligibility, the wavLM model does not explicitly extract handcrafted features. Instead, it directly operates on the raw waveform data, allowing it to capture complex patterns and dependencies present in the speech signal. The wavLM model learns to encode various aspects of speech intelligibility within its internal representations and capture acoustic features such as pitch, spectral characteristics, phonetic content, and temporal dynamics, among others. In addition, by training on a large corpus of speech data, the WavLM model learns to implicitly extract relevant features for the task at hand, making it a powerful tool for speech intelligibility prediction.

B. Regression analysis

The extracted features were fed to the base-regressor and the meta-regressor to obtain the final speech intelligibility score. While speech is indeed a complex and multidimensional signal, we use the base-regressor, which consists of linear regressor, support vector machines, and random forest, to make predictions. Although linear regression models have been successfully used in various speech-related tasks, in the case of complex listening environments and listener individuals that affect speech intelligibility, the pattern captured by single linear model cannot be adequate. Therefore, we utilize support vector machines and random forests, alongside linear regressor, to capture broader range of relationships and better handle non-linearities in the data. Finally, each of the prediction results

was fed to the Ridge-CV as the meta-regressor and ensemble model to generate the final speech intelligibility score.

IV. Experiment

A. Dataset

The dataset utilized in the experiment is associated with 6 speakers, 10 hearing aid systems from the entrants of the first Clarity Enhancement Challenge, and 27 listeners [26]. The noisy speech in the dataset was generated from the various living room scenarios of sound propagation through the room and interaction to the human head. The target sentences consist of 7-10 words in length with the subset of 1500 utterances [27]. On the other hand, the interferer was recorded from sounds of daily electronic appliances, such as washing machines, vacuum cleaners, and kettles, to demonstrate the non-impulsive noise in the real world condition. Next, to create the consistence scenarios, the position of target speaker, listeners, and interferer is adjusted. For each scene, input signals are convoluted with appropriate geometric room acoustic model and head related transfer function (HTRF) database, which include measurements for hearing aid microphones to create suitable inputs to hearing aid system.

The information about the listeners is also given in the dataset and characterized by the bilateral pure-tone audiograms. The audiogram is specified with average hearing loss in dB in frequency between 2 to 8 kHz. Based on the measurement obtained, there is 1 listener with mild hearing loss (15–35 dB), 9 listeners with moderate hearing loss (35–56 dB), and 17 listeners with severe hearing loss (>56 dB) [28]. The other hearing ability measurement is also provided, that is, the DTT, Glasgow hearing-aid benefit profile questionnaire (GHABP), and Speech, Spatial, & Qualities of Hearing questionnaire (SSQ12). In this paper, more concentration will be devoted to the analysis of audiogram and the DTT results. Since the results of the DTT are not available for several listeners, we remove those data in investigating the speech intelligibility.

There are two tracks provided in the dataset. Track 1 is a closed-set track with the known listeners and hearing aid processors in the training set (4812 responses) and test set (2421 responses). Track 2 is an open set track with one hearing aid processor and five listeners unseen in the training set. The remaining 22 listeners and 9 hearing aid processors are available in both training and test sets, generated (3545 responses). The purpose of including track 2 is to demonstrate how well the method generalizes the prediction to the unseen data. Since the DTT results are unavailable for all listeners, we remove all the data with missing values. The total number of listeners is nineteen people with the characteristics detailed in Table III. Note that due to this removal process, the total number of data for this experiment is less than those in our prior works and the existing works in the Clarity Prediction Challenge [29].

B. Evaluation Metrics

Five universal metrics used for regression task are utilized to assess the results, including Pearson correlation coefficient (ρ_p), Spearman correlation coefficient (ρ_s), root mean square error (RMSE), standard error (SE), and the coefficient of determination (R^2).

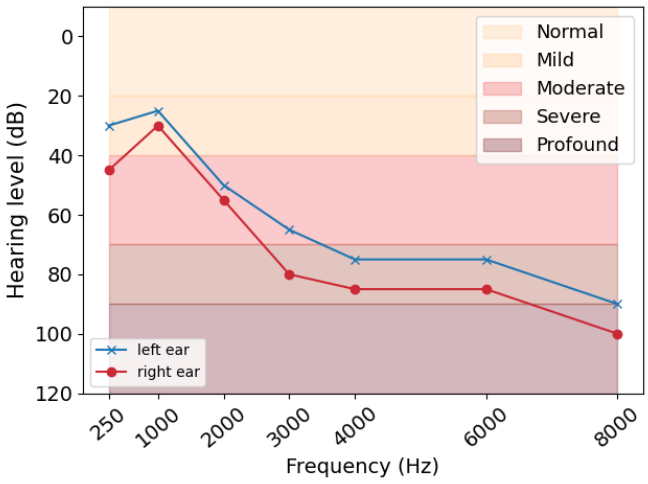

Fig. 2. The audiogram of listener L239. The hearing ability is categorized into severe hearing loss since the average hearing loss level in the audiogram with frequency in {250, 500, 1000, 2000, 3000, 4000, 6000, 8000} Hz is $>$ 56 dB. Additionally, the SRT obtained from the DTT can be categorized as high, which indicates the minimum hearing level to recognize 50% of the digit tests.

The ρ_p assesses the linear correlation between two variables. In the context of speech intelligibility prediction, ρ_p is used to determine the relationship between the actual and predicted speech intelligibility score. The ρ_s is also corporated to measure the association between the actual and predicted speech intelligibility score, even if they have a non-linear relationship. To assess the prediction accuracy, RMSE is corporate. In addition, SE helps to determine the precision of the prediction and reliability of the experiment results. Moreover, R^2 represents the goodness of fit between the predicted and actual intelligibility scores.

C. Results

In this section, we present the results of the experiment using the proposed lightweight machine learning method with several speech features for predicting speech intelligibility. Table II shows the experiment results of the proposed method compared to the baseline MBSTOI [29] in both Track 1 and Track 2. The baseline MBSTOI was developed based on the MBSTOI metric [17] with the hearing loss model developed by Nejime et al. [30]. The speech intelligibility model was built by fitting the output MBSTOI score of the training data with a logistic mapping model with a sigmoid function.

Since we aim at reducing the computational complexity, we experiment on several low dimensional acoustic features (i.e., eGeMAPS) and the embedding features in the state-of-the-art features for automatic speech recognition (ASR) (i.e., wav2vec2, HuBERT, and wavLM). The listener characteristics, such as audiogram and the DTT, were also simplified into categorical data. For audiograms, we use three categories: mild, moderate, and severe. Meanwhile, based on the SRT for the DTT results, we use two categories: low and high. These

TABLE II
EXPERIMENT RESULTS OF THE PROPOSED METHOD WITH SEVERAL ACOUSTIC FEATURES COMPARE TO THE BASELINE MBSTOI.

Method	**Feature dim.**	**Track 1 (closed-set)**					**Track 2 (open-set)**				
		$\rho_p \uparrow$	$\rho_s \uparrow$	**RMSE** $\downarrow$	**SE** $\downarrow$	$R^2 \uparrow$	$\rho_p \uparrow$	$\rho_s \uparrow$	**RMSE** $\downarrow$	**SE** $\downarrow$	$R^2 \uparrow$
Baseline MBSTOI [29]	-	0.61	0.53	28.92	0.71	0.37	0.54	0.53	35.51	1.53	0.05
Proposed method (mean value)											
GeMAPS	62	0.71	0.58	25.77	0.63	0.50	0.58	0.53	30.85	1.36	0.29
eGeMAPS	88	0.72	0.60	25.26	0.62	0.52	0.60	0.54	30.09	1.34	0.32
wav2vec2	1024	0.66	0.55	27.48	0.67	0.43	0.51	0.41	31.56	1.45	0.25
HuBERT	1024	0.70	0.59	26.08	0.64	0.49	0.55	0.47	30.76	1.40	0.29
wavLM	1024	**0.75**	**0.63**	**24.30**	**0.59**	**0.56**	**0.64**	**0.58**	**28.25**	**1.29**	**0.40**
Proposed method (concatenate)											
GeMAPS	62 x #sec.	0.71	0.59	25.68	0.63	0.50	0.60	0.53	29.59	1.34	0.34
eGeMAPS	88 x #sec.	0.72	0.60	25.49	0.62	0.51	0.60	0.53	29.67	1.34	0.34
wav2vec2	1024 x #sec.	0.63	0.52	28.39	0.69	0.39	0.50	0.41	31.72	1.45	0.25
HuBERT	1024 x #sec.	0.68	0.57	26.85	0.66	0.46	0.50	0.39	32.20	1.46	0.22
wavLM	1024 x #sec.	**0.74**	**0.62**	**24.54**	**0.60**	**0.55**	**0.59**	**0.54**	**29.51**	**1.35**	**0.35**
Proposed method (concatenate + feature selection (512))											
wav2vec2	512	0.64	0.53	28.11	0.69	0.41	0.37	0.23	34.47	1.57	0.11
HuBERT	512	0.69	0.57	26.55	0.65	0.47	0.55	0.46	30.73	1.39	0.29
wavLM	512	**0.74**	**0.62**	**24.52**	**0.60**	**0.55**	**0.60**	**0.55**	**29.40**	**1.35**	**0.35**

TABLE III
LISTENER CHARACTERISTICS FROM THE DIGIT TRIPLET TEST RESULTS AND AUDIOGRAM

Listener	L200	L201	L202	L209	L215	L216	L218	L219	L220	L222
SRT	-6.7	-8.8	-7.6	-6.1	-8.5	-6.3	-6.4	-6.1	-17.0	-8.2
DTT category	High	Low	High	High	High	High	High	High	Low	High
Hearing loss level	Severe	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate	Severe	Moderate	Moderate
Listener	L225	L227	L229	L231	L235	L239	L241	L242	L243	
SRT	-8.3	-6.5	-11.4	-11.0	-10.0	-6.1	-7.1	-6.9	-12.1	
DTT category	High	High	Low	Low	Low	High	High	High	Low	
Hearing loss level	Moderate	Severe	Moderate	Moderate	Mild	Severe	Moderate	Severe	Moderate	

categories were based on the calculation of the average SRT of all listeners that represents a collective level of speech intelligibility. If an individual's SRT value is significantly higher than the mean value, the individual needs a higher signal-to-noise ratio to discriminate speech accurately. The individual has poor speech intelligibility under specific hearing conditions and may have a hearing loss or hearing impairment. Conversely, an individual with a significantly lower than average SRT value indicates that the individual can discriminate speech accurately at a relatively low signal-to-noise ratio, which implies that the individual has better speech intelligibility and auditory discrimination and may have better hearing ability. Table III shows the categorization of the listener characteristics.

Table II indicates the results from the baseline MBSTOI and the proposed method with three ways to reduce the dimension of the features. For wav2vec2, HuBERT, and wavLM, we extracted the output of the intermediate layer for each second of the speech signal. For instance, if a signal has a 5-second length, then we could extract 1024 $\times$ five dimensional feature. We utilized these features in three ways by 'mean value', 'concatenate', and 'concatenate + feature selection (512)'. The 'mean value' is performed by averaging the extracted feature based on the time stamp. Hence, we only have a vector with 1024 dimensions. The 'concatenate' means that we use all the extracted features, so a vector with 1024 $\times$ number of seconds (#sec.) dimension can be obtained. Lastly, the 'concatenate + feature selection (512)' means after the concatenation, we perform a univariate feature selection by selecting the 512 highest-scoring features based on the training data.

In most cases, the results in the table II have a higher correlation, lower RMSE, and lower SE, which indicates that the proposed method could achieve better results than the baseline MBSTOI. In addition to the higher correlation and lower RMSE, the proposed method could be regarded as a blind or non-intrusive method because the reference speech signal is not required for predicting speech intelligibility. Although the GeMAPS only consists of 62 dimensions, it could be utilized to achieve higher performance compared to the baseline method. The additional 26 extended parameters also positively contribute to improving the prediction accuracy. Comparing the performance of these low dimensional features to the recent state-of-the-art features, the wavLM could achieve the highest performance almost in all evaluation metrics used in the experiment. This finding aligns well with the prior research, which also showed that the wavLM performs better than wav2vec2 and HuBERT on all downstream speech processing tasks [25].

Observing the experiment results we conducted to reduce the feature dimension, we found that averaging the vectors obtained from each speech segment could achieve a higher

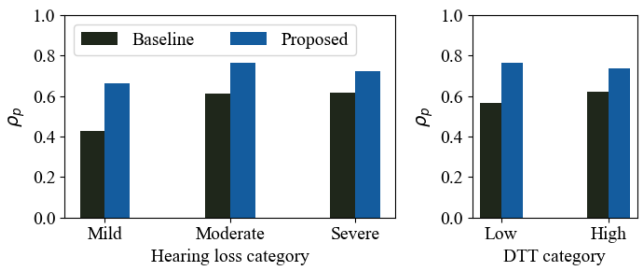

Fig. 3. Correlation analysis on the actual correctness and prediction results based on the listener characteristics: (Left) audiogram, (Right) digit triplet test (DTT)

prediction accuracy. For instance, by using the wavLM feature, the ρ_p of the proposed method (mean value) is 5% higher than the proposed method (concatenate). Similarly, the RMSE is smaller, and the R^2 is higher. The feature selection method in our experiment could not improve the prediction performance. For instance, the prediction by the feature selection using wav2vec2 and HuBERT caused a reduction in the correlation. At the same time, the results show no significant difference ($\rho > 0.05$) compared to those without feature selection.

An in-depth understanding of the hearing status allows for assessing the degree and type of hearing loss. An audiogram is a standard assessment tool that provides detailed information about hearing loss by plotting an individual's hearing levels at different frequencies and volume levels. Meanwhile, the SRT obtained from the DTT provides a direct and operational indicator of an individual's level of speech intelligibility in complex listening environments. Figure 3 shows the results grouped by the audiogram and DTT results. These results indicate that the proposed method could more accurately predict speech intelligibility in all categories of hearing loss conditions. Compared to the analysis of the results by audiogram and the DTT categories, the prediction using the DTT category shows a more stable speech intelligibility prediction for listeners with mild levels of hearing loss than the prediction grouped by the hearing loss category from audiogram.

While this research provides valuable insights into speech intelligibility prediction, we acknowledge certain limitations inherent in the study. First, the dataset used for training and evaluation may be limited and not fully represent the wide range of real-world listener conditions and noisy scene scenarios. To our knowledge, the CPC1 dataset is the only data that includes the DTT results. In the future, it is important to carry out more data collection for additional listeners on a broader range of scenes. Second, the study primarily focuses on the features used in common speech processing, such as in automatic speech recognition or emotion recognition, overlooking potential contributions from other important acoustic features or cues in speech intelligibility perception. For instance, the binaural hearing cues naturally utilized by human ears to recognize speech have yet to be considered. Future research should explore the incorporation of these cues to enhance the accuracy and robustness of speech intelligibility prediction models.

V. Conclusion

We proposed a method incorporating low-dimensional and state-of-the-art acoustic features for speech processing to predict speech intelligibility in noise for hearing aids. The prediction method was developed based on a stack regressor using various traditional machine learning regressors, including linear regressor, support vector regressor, and random forest regressor. Based on the overall results, the proposed method outperformed the previous method, particularly when utilizing wavLM features. The better performance is achieved with the help of the DTT input that is categorized based on the calculation of the average SRT of all listeners, representing speech intelligibility level. Since the proposed method does not require high computational power due to the low dimensional feature and relatively simple machine learning model, it can be considered as a baseline method before going through a further experiment using extensive machine learning models and sophisticated features for speech intelligibility prediction. In future work, we plan to further analyze the phenomena in hearing loss that both successfully and unsuccessfully restored by hearing aids and the effect on speech intelligibility perception.

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), and the Japan Society for the Promotion of Science (JSPS) KAKENHI grant (No. 22K21304).

References

[1] D. M. Jayakody, O. P. Almeida, C. P. Speelman, *et al.*, "Association between speech and high-frequency hearing loss and depression, anxiety and stress in older adults," *Maturitas*, vol. 110, pp. 86–91, 2018.

[2] T. L. G. Health, "Amplifying the global issue of hearing loss," *The Lancet. Global health*, vol. 10(10), e1360, 2022.

[3] J. Löhler, M. Cebulla, W. Shehata-Dieler, S. Volkenstein, C. Völter, and L. E. Walther, "Hearing impairment in old age," *Deutsches Ärzteblatt International*, vol. 116, no. 17, pp. 301–310, 2019.

[4] H. Levitt, "Noise reduction in hearing aids: A review," *Journal of Rehabilitation Research and Development*, vol. 38, no. 1, pp. 111–121, 2001.

[5] L. Gwilliams and M. H. Davis, "Extracting language content from speech sounds: The information theoretic approach," in *Speech perception*, vol. 74, 2022, pp. 113–139.

[6] H. Meister, S. Rählmann, M. Walger, S. Margolf-Hackl, and J. Kießling, "Hearing aid fitting in older persons with hearing impairment: The influence of cognitive function, age, and hearing loss on hearing aid benefit," *Clinical Interventions in Aging*, pp. 435–443, 2015.

[7] R. M. Cox, G. C. Alexander, and I. M. Rivera, "Comparison of objective and subjective measures of speech intelligibility in elderly hearing-impaired listeners," *Journal of Speech, Language, and Hearing Research*, vol. 34, no. 4, pp. 904–915, 1991.

[8] Y. Feng and F. Chen, "Nonintrusive objective measurement of speech intelligibility: A review of methodology," *Biomedical Signal Processing and Control*, vol. 71, p. 103 204, 2022.

[9] D. Byrne and N. Murray, "Predictability of the required frequency response characteristic of a hearing aid from the pure-tone-audiogram," *Ear and hearing*, vol. 7, no. 2, pp. 63–70, 1986.

[10] E. Van 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.

[11] A. Heinrich, H. Henshaw, and M. A. 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, p. 782, 2015.

[12] M. S. Vlaming, R. C. MacKinnon, M. Jansen, and D. R. Moore, "Automated screening for high-frequency hearing loss," *Ear and hearing*, vol. 35, no. 6, p. 667, 2014.

[13] D. Ertmer, "Relationships between speech intelligibility and word articulation scores in children with hearing loss," *Journal of speech, language, and hearing research*, vol. 53, pp. 1075–86, 2010.

[14] D. Kewley-Port, T. Z. Burkle, and J. H. Lee, "Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners," *Journal of the Acoustical Society of America*, vol. 122, no. 4, pp. 2365–2375, 2007.

[15] A. J. Bosmana and G. F. Smoorenburg, "Intelligibility of dutch cvc syllables and sentences for listeners with normal hearing and with three types of hearing impairment," *Audiology*, vol. 34, no. 5, pp. 260–284, 1995.

[16] S. H. Ferguson and D. Kewley-Port, "Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners," *Journal of the Acoustical Society of America*, vol. 112, no. 1, pp. 259–271, 2002.

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

[18] A. H. Andersen, J. M. d. Haan, Z.-H. Tan, and J. Jensen, "A binaural short time objective intelligibility measure for noisy and enhanced speech," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[19] P. Guiraud, A. H. Moore, R. R. Vos, P. A. Naylor, and M. Brookes, "Machine learning for parameter estimation in the mbstoi binaural intelligibility metric," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, IEEE, 2022, pp. 1–5.

[20] T. H. Falk, V. Parsa, J. F. Santos, *et al.*, "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *IEEE signal processing magazine*, vol. 32, no. 2, pp. 114–124, 2015.

[21] F. Eyben, K. R. Scherer, B. W. Schuller, *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[22] B. Schuller and G. Rigoll, "Recognising interest in conversational speech-comparing bag of frames and supra-segmental features," in *Interspeech 2009*, 2009, pp. 1999–2002.

[23] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[24] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[25] S. Chen, C. Wang, Z. Chen, *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[26] S. Graetzer, J. Barker, T. J. Cox, *et al.*, "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing," in *Interspeech 2021*, vol. 2, 2021, pp. 686–690.

[27] S. Graetzer, M. A. Akeroyd, J. Barker, *et al.*, "Dataset of british english speech recordings for psychoacoustics and speech processing research: The clarity speech corpus," *Data in brief*, vol. 41, p. 107 951, 2022.

[28] T. Cox, M. Akeroyd, J. Barker, *et al.*, "Predicting speech intelligibility for people with a hearing loss: The clarity challenges," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 265, 2023, pp. 4599–4606.

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

[30] Y. Nejime and B. C. 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, no. 1, pp. 603–615, 1997.