MTI-Net: A Multi-Target Speech Intelligibility Prediction Model

Ryandhimas E. Zezario[1 2], Szu-wei Fu,[3], Fei Chen[4], Chiou-Shann Fuh[1], Hsin-Min Wang[2], Yu Tsao[2]

[1]National Taiwan University, [2]Academia Sinica, [3]Microsoft Corporation,
[4] Southern University of Science and Technology of China

{ryandhimas, yu.tsao}@citi.sinica.edu.tw

Abstract

Recently, deep learning (DL)-based non-intrusive speech assessment models have attracted great attention. Many studies report that these DL-based models yield satisfactory assessment performance and good flexibility, but their performance in unseen environments remains a challenge. Furthermore, compared to quality scores, fewer studies elaborate deep learning models to estimate intelligibility scores. This study proposes a multi-task speech intelligibility prediction model, called MTI-Net, for simultaneously predicting human and machine intelligibility measures. Specifically, given a speech utterance, MTI-Net is designed to predict human subjective listening test results and word error rate (WER) scores. We also investigate several methods that can improve the prediction performance of MTI-Net. First, we compare different features (including low-level features and embeddings from self-supervised learning (SSL) models) and prediction targets of MTI-Net. Second, we explore the effect of transfer learning and multi-tasking learning on training MTI-Net. Finally, we examine the potential advantages of fine-tuning SSL embeddings. Experimental results demonstrate the effectiveness of using cross-domain features, multi-task learning, and fine-tuning SSL embeddings. Furthermore, it is confirmed that the intelligibility and WER scores predicted by MTI-Net are highly correlated with the ground-truth scores.

Index Terms: Subjective listening tests, WER, STOI, speech intelligibility prediction, self-supervised learning

1. Introduction

For many speech-related applications, such as hearing aids, telecommunications, and automatic speech recognition (ASR), speech intelligibility is a key metric of user satisfaction. The metric measures the ratio of correctly recognized words to total words in a listening test, in which participants listen to a set of speech and answer what they hear. Since the measurements are based on human hearing tests, collecting data from a sufficient number of listeners is critical for unbiased measurements. However, conducting large-scale hearing tests is prohibitive. To overcome this problem, many signal processing-based intelligibility metrics have been proposed as surrogate metrics for human listening behavior, such as articulation index (AI) [1], speech intelligibility index (SII) [2], extended SII (ESII) [3], speech transmission index (STI) [4], and short-time objective intelligibility (STOI) [5]. These signal processing-based intelligibility metrics can be roughly divided into two categories, intrusive and non-intrusive metrics. Intrusive metrics [1, 2, 3, 4, 5] require the corresponding clean speech as a reference to measure intelligibility scores, while non-intrusive metrics do not need a reference [6, 7]. Compared to intrusive metrics, non-intrusive metrics have better flexibility but generally provide lower prediction accuracy.

Recently, deep learning (DL) models have been introduced into speech intelligibility prediction systems. For these systems, a DL model is used as a regression function to predict the intelligibility score given a speech signal. Based on the types of ground-truth labels, these DL-based prediction systems can be divided into two categories: one that predicts objective evaluation scores, such as STI, and STOI [8, 9, 10], and the other that predicts human subjective ratings of human listening tests [11, 12, 13]. Due to differences in the listening abilities of the listeners, the range of human subjective ratings is larger than that of objective scores, as reported in [13]. Therefore, training an intelligibility model that predicts human subjective ratings is more difficult than training a model that predicts objective scores.

In our previous work, we proposed a multi-objective speech assessment model, called MOSA-Net [14]. MOSA-Net uses cross-domain features (spectral and temporal features) and latent representations from a SSL model [15] to predict objective quality and intelligibility scores simultaneously. As reported in [14], based on the cross-domain features and multi-task learning criterion, MOSA-Net can accurately predict objective quality (PESQ) [16] and intelligibility (STOI) [5] scores. In this paper, we propose an improved version of MOSA-Net, called the Multi-Target Speech Intelligibility Prediction Model (MTI-Net), for simultaneously predicting human and machine intelligibility scores. More specifically, machine intelligibility scores are word error rate (WER) scores from automatic speech recognition (ASR); and human intelligibility scores include: (1) subjective listening test results, and (2) objective evaluation scores (STOI in this study). MTI-Net is formed by a convolutional bidirectional long short-term memory (CNN-BLSTM) architecture with a multiplicative attention mechanism. To improve the prediction power, we further fine-tune the SSL model and use the fine-tuned embeddings (latent representations) as the input features for MTI-Net. Experimental results confirm that MTI-Net can predict human and machine speech intelligibility assessment scores well. With such multi-task learning criteria (simultaneous prediction of subjective listening test results, WER, and STOI), the prediction accuracy of individual outcomes can be improved. Furthermore, we observe that the fine-tuned SSL embeddings help MTI-Net achieve better results than SSL embeddings without fine-tuning.

The rest of this paper is organized as follows. Section II reviews related work. Section III presents the proposed MTI-Net. Section IV describes experimental setup and results. Finally, the conclusions and future work are presented in Section V.

2. Related Work

2.1. DL-based intelligibility prediction models

Most DL-based speech assessment models aim to predict quality scores, and fewer work has focused on predicting intelli-

gibility scores. In [12] and [11], intrusive and non-intrusive speech intelligibility prediction models are proposed; both employ a CNN model as the main model architecture and take compressed spectral features as input to predict measured intelligibility from multiple listening experiments. In [10] and [14], the CNN-BLSTM model architecture and attention mechanism are used to predict the objective intelligibility score (STOI). In [17], the residuals of speech enhancement (SE) processed speech are incorporated to facilitate evaluating models to predict objective evaluation metrics (PESQ and ESTOI) without the need for a clean reference. Recently, a non-intrusive hearing aid speech assessment network (HASA-Net) [18] was proposed. Compared to other assessment models, HASA-Net takes the hearing loss pattern as an additional input to predict well-known hearing-aid evaluation metrics, namely the hearing aid speech quality index (HASQI) [19] and the hearing aid speech perception index (HASPI) [20].

2.2. Representations of SSL models

Recently, the audio SSL model has attracted a lot of attention. Audio SSL models are trained from large-scale unlabeled data to learn representative embeddings [21]. It has been shown that SSL embeddings can be used as input features for various downstream tasks and yield promising performance [22, 23, 24, 21]. In [25], it was reported that fine-grained acoustic information may not be fully characterized by SSL embeddings. To further improve SSL embeddings to specific target tasks, some studies propose to take additional raw data or low-level features as input [26, 14, 27]. Meanwhile, fine-tuning SSL on the target downstream task also shows significant improvements. For examples, fine-tuning an SSL model improves three recognition tasks (speech emotion recognition, speaker verification, and spoken language understanding) [28], end-to-end speech translation [27], and MOS prediction [29]. The above studies confirmed the effectiveness of combining cross-domain features with SSL embeddings and fine-tuning the SSL model to improve performance.

3. Multi-Target Speech Intelligibility Prediction Model

The overall architecture of MTI-Net is shown in Fig. 1. As shown in the figure, MTI-Net takes input from two branches. In the first branch, given a speech waveform **X**, the short-time Fourier transform (STFT) and learnable filter banks (LFB) are applied to generate two types of acoustic features, which are concatenated and sent to the convolutional layer. In the second branch, the speech waveform **X** is processed by an SSL model to obtain SSL embeddings. These two branches of features are concatenated and fed into a bidirectional layer and a fully connected layer. Finally, the output of the fully connected layer is mapped to three assessment metrics, namely the intelligibility, WER, and STOI scores. For each metric, an attention layer, a fully connected layer, and a global average pooling layer are applied to generate the final prediction output. To improve training stability, the objective function for training MTI-Net is a combined frame-level and utterance-level score, defined as follows:

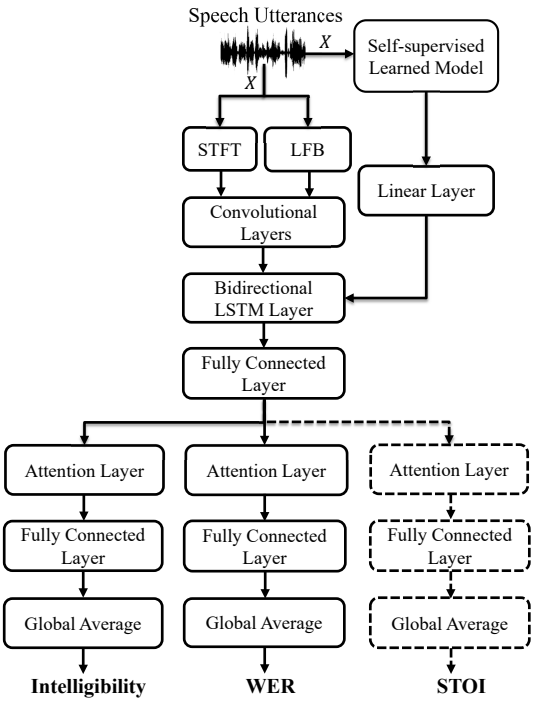

Figure 1: *Architecture of the MTI-Net model.*

$$
\begin{aligned}
O &= L_I + L_W + L_S \\
L_I &= \tfrac{1}{U}\sum_{u=1}^{U}[(I_u - \hat{I}_u)^2 + \tfrac{\alpha_I}{F_u}\sum_{f=1}^{F_u}(I_u - \hat{i}_f)^2] \\
L_W &= \tfrac{1}{U}\sum_{u=1}^{U}[(W_u - \hat{W}_u)^2 + \tfrac{\alpha_W}{F_u}\sum_{f=1}^{F_u}(W_u - \hat{w}_f)^2] \\
L_S &= \tfrac{1}{U}\sum_{u=1}^{U}[(S_u - \hat{S}_u)^2 + \tfrac{\alpha_S}{F_u}\sum_{f=1}^{F_u}(S_u - \hat{s}_f)^2]
\end{aligned}
\tag{1}
$$

where $\{I_u, \hat{I}_u\}$, $\{W_u, \hat{W}_u\}$, and $\{S_u, \hat{S}_u\}$ are the true and predicted utterance-level scores of intelligibility, WER, and STOI, respectively; U denotes the total number of training utterances; F_u denotes the number of frames in the u-th training utterance; $\hat{i}_f$, $\hat{w}_f$, and $\hat{s}_f$ are the predicted frame-level scores of Intelligibility, WER, and STOI of the f-th frame, respectively; α_I, α_W, and α_S are the weights between utterance-level and frame-level losses.

To further improve the prediction accuracy of MTI-Net, we fine-tune the SSL model for the intelligibility prediction task. The process of fine-tuning the SSL model is defined as follows:

$$
\begin{aligned}
\textbf{Res} &= SSLModel_\theta\,(\mathbf{X}) \\
\textbf{Mean} &= MeanPooling\,(\textbf{Res}) \\
EstimatedScore &= LinearLayer_\theta\,(\textbf{Mean})
\end{aligned}
\tag{2}
$$

where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_u \ldots, \mathbf{x}_U]$ denotes the speech waveforms used for model training, **Res** denotes the embeddings from the SSL Model, **Mean** denotes the mean pooling of the embeddings. The parameter set θ of the SSL model is fine-tuned by minimizing the loss function based on the mean squared error (MSE), i.e., the first part of L_I in Eq. (1). Note that during the training of the MTI-Net with SSL features, the parameters of the pre-trained or fine-tuned SSL model are fixed. Therefore,

the training objective function of the MTI-Net with fine-tuned SSL features is still Eq. (1).

In this study, we focus on the following two points: (1) As reported in [29], fine-tuning the SSL model can already achieve satisfactory MOS prediction results. We intend to explore whether fine-tuning SSL embeddings in combination with low-level (but fine-grained) features can yield further improvements in predicting speech intelligibility. (2) We intend to investigate whether the multi-task learning criterion for predicting scorers of subjective listening tests, ASR results, and objective STOI scores can improve prediction performance.

4. Experiments

4.1. Experimental setup

We evaluated the proposed MTI-Net model on the Taiwan Mandarin Hearing In Noise test - Quality & Intelligibility (TMHINT-QI) [13] dataset. The dataset includes clean, noisy, and enhanced utterances from five different SE systems, including Karhunen-Loeve transform (KLT) [30], minimum-mean squared error (MMSE) [31], fully convolutional network (FCN) [32], deep denoising autoencoder (DDAE) [33], and transformer-based SE [34]). There are 226 subjects participated in the listening test [1], and each listener was requested to rate the quality and answer what she/he heard for 108 utterances. For more details on the TMHINT-QI dataset, please see [13]. In this study, we selected the intelligibility score as one target. Another target, WER, was obtained by performing recognition on each utterance using the open-source Google ASR system [35]. Each subjective intelligibility score ranges from zero to one and represents the percentage of correctly recognized characters. Similarly, each WER score ranges from zero to one, with lower scores indicating higher ASR accuracy.

To prepare the training set, we selected 15,000 utterances, each of which was evaluated by one listener. To prepare the test set, we selected 1,900 utterances, each of which was evaluated by 2 to 3 listeners, and the average score of these listeners was used as the ground-truth score for each utterance. Notably, the training and test utterances did not overlap. We used three evaluation metrics, namely MSE, linear correlation coefficient (LCC), and Spearman's rank correlation coefficient (SRCC) [36] to evaluate the performance of MTI-Net. A lower MSE value indicates that the predicted scores are closer to the ground-truth scores (lower is better), while higher LCC and SRCC scores indicate a higher correlation between predicted scores and ground-truth scores (higher is better).

4.2. MTI-Net with different features and targets

In the first experiment, we aim to compare the performance of MTI-Net using different features and targets. We tested performance using power spectral features, features extracted by learnable filter banks, and the SSL embeddings, denoted as PS, LBF, and SSL, respectively. The HuBERT model [15] was used to extract the SSL embeddings. The cross-domain features (hereafter referred to as CS) stand for the use of three features: PS, LBF, and SSL. For a fair comparison, the same model architecture (CNN-BLSTM with attention) [14] was used to test the performance of different features. We further tested two prediction targets: one only predicting listening test results (denoted as I), and the other one predicting both listening test results and WER scores (denoted as I+W). The experimental results are summarized in Table 1. From the table, we first note that MTI-Net with cross-domain features can achieve the best performance for both intelligibility and WER score prediction, confirming the advantage of combining acoustic information from different features. Next, when comparing the results of single-target (I) and two-target (I+W), the (I+W) system always outperform its (I) counterpart. The results demonstrate that multi-task learning enables MTI-Net to achieve better speech intelligibility and WER predictions. In the following, we will adopt MTI-Net using cross-domain features and two-target (I+W) as the base system for further comparison.

Table 1: *LCC, SRCC, and MSE results of MTI-Net using different features and targets. PS, LBF, and SSL, respectively, denote power spectral features, features extracted by learnable filter banks, and the SSL embeddings. CS denotes the cross-domain features. I and W denote the intelligibility and WER scores, respectively, which are the prediction targets of MTI-Net.*

Feature	Target	LCC	SRCC	MSE
Intelligibility Score Prediction				
PS	I	0.543	0.483	0.034
PS	I+W	0.585	0.570	0.031
LBF	I	0.386	0.371	0.041
LBF	I+W	0.588	0.536	0.032
SSL	I	0.630	0.610	0.003
SSL	I+W	0.640	0.625	0.029
CS	I	0.731	0.685	0.022
CS	I+W	**0.761**	**0.708**	**0.020**
WER Score Prediction				
PS	W	0.661	0.644	0.057
PS	I+W	0.611	0.615	0.065
LBF	W	0.576	0.514	0.066
LBF	I+W	0.583	0.542	0.066
SSL	W	0.734	0.725	0.047
SSL	I+W	0.690	0.687	0.052
CS	W	0.774	0.764	0.040
CS	I+W	**0.779**	**0.766**	**0.040**

4.3. MTI-Net with knowledge transfer (KT) and multi-task learning (MTL) methods

The main task of MTI-Net is to predict intelligibility and WER scores. We examine whether the STOI metric provides MTI-Net useful information for better prediction ability. Therefore, in this experiment, we investigate two methods for integrating STOI information into MTI-Net. The first method is based on knowledge transfer (KT), and the second method is based on multi-task learning (MTL). For the KT method, we adopted the pretrained network STOI-Net [10] as the seed model and fine-tuned the model to suit the prediction task of this study. For the MTL method, we simply used the STOI score as an additional target, as shown in Fig. 1. The best system in Table 1, i.e., CS(I+W), was used as the base model (denoted as Base in Table 2). The base models refined by the KT and MTL methods are denoted as KT and MTL, respectively, in Table 2. Table 2 lists the results of KT and MTL. The results of CS(I+W) in Table 1, are also listed and termed Base in Table 2 for comparison. As can been seen from Table 2, KT achieves better performance than Base in both intelligibility and WER prediction,

[1] Written informed consent approved by the Academia Sinica Institutional Review Board for this study was obtained from each participant before conducting the experiment.

Table 2: *LCC, SRCC, and MSE results of MTI-Net with knowledge transfer (KT) and multi-task learning (MTL).*

Methods	Target	LCC	SRCC	MSE
Intelligibility Score Prediction				
Base	I+W	0.761	0.708	0.020
KT	I+W	0.781	0.713	0.019
MTL	I+W+S	**0.789**	**0.722**	**0.018**
WER Score Prediction				
Base	I+W	0.779	0.766	0.040
KT	I+W	0.798	0.780	0.036
MTL	I+W+S	**0.811**	**0.802**	**0.034**

Table 3: *LCC, SRCC, and MSE results of MTI-Net without and with fine-tuning SSL embeddings, termed MTI-Net and MTI-Net(FT-SSL), respectively.*

Systems	Target	LCC	SRCC	MSE
Intelligibility Score Prediction				
FT-SSL	I+W+S	0.795	0.678	0.024
MTI-Net	I+W+S	0.789	0.722	0.018
MTI-Net(FT-SSL)	I+W+S	**0.823**	**0.735**	**0.017**
WER Score Prediction				
FT-SSL	I+W+S	0.824	0.815	0.036
MTI-Net	I+W+S	0.811	0.802	0.034
MTI-Net(FT-SSL)	I+W+S	**0.834**	**0.822**	**0.031**

confirming the advantages of the KT method. Furthermore, MTL achieves the best performance among the three methods, indicating that incorporating the STOI score at the target can effectively improve the prediction accuracy.

4.4. MTI-Net with fine-tuning SSL embeddings

Next, we examine the effectiveness of fine-tuning SSL embeddings in MTI-Net. Table 3 shows the prediction results of two MTI-Net systems: (1) MTI-Net: the best system from Table 2, which uses cross-domain features without fine-tuning SSL embeddings; and (2) MTI-Net(FT-SSL): MTI-Net using fine-tuned SSL embeddings in cross-domain features. The results of directly fine-tuning the SSL model to predict intelligibility, WER and STOI scored are also listed for comparison, denoted as FT-SSL in Table 3. We followed Eq. (2) and [29] and used the script[2] to implement FT-SSL. Note that compared to the fine-tuning process of the SSL model described in Section III, a slight modification was applied to enable multi-task learning. Specifically, the mean pooling of the embeddings from the SSL model was fed to three linear layers corresponding to intelligibility, WER, and STOI, respectively, instead of one linear layer for predicting intelligibility. From Table 3, we can see that MTI-Net using pre-trained SSL embeddings (without fine-tuning the SSL model during MTI-Net training) can already achieve promising performance, comparable to FT-SSL in all evaluation metrics. However, MTI-Net(FT-SSL) yields the best performance, confirming the effectiveness of fine-tuning SSL embeddings during MTI-Net training.

We also present the scatter plots of the predicted targets of MTI-Net(FT-SSL), MTI-Net, and FT-SSL in Fig. 2. The figure again confirms that MTI-Net(FT-SSL) can achieve more accurate intelligibility and WER predictions than MTI-Net and FT-SSL. More specifically, the points of MTI-Net(FT-SSL) are more densely distributed on the diagonal than those of MTI-Net and FT-SSL. The results in Tables 1-3 and Fig. 2 confirm the benefits of fine-tuning SSL embeddings, using cross-domain features, and the multi-task learning criterion of MTI-Net.

[2]https://github.com/nii-yamagishilab/mos-finetune-ssl

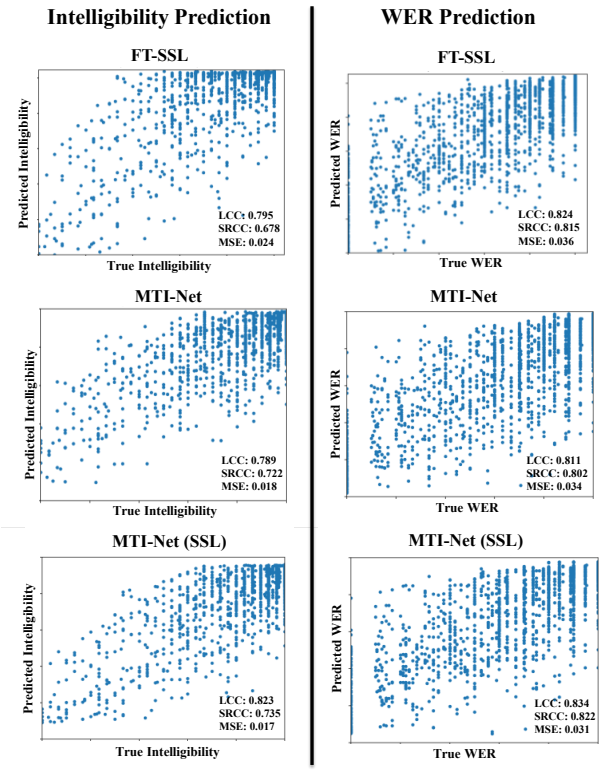

Figure 2: *Scatter plots of three speech intelligibility prediction models, including FT-SSL [29], MTI-Net, and MTI-Net(FT-SSL).*

5. Conclusions

In this paper, we have proposed MTI-Net that aims to predict human speech intelligibility and machine WER. MTI-Net adopts cross-domain features and a CNN-BLSTM model architecture with attention mechanism, and is trained by a multi-task learning criterion to predict both subjective listening test and WER scores simultaneously. In experiments, we first demonstrated the effectiveness of combining low-level (and fine-grained) features with SSL embeddings. Next, we confirmed the advantages of multi-task learning. Finally, we verified the positive results of fine-tuning SSL embeddings. To our knowledge, this is the first work that aimed to directly predict WER with high correlation scores. Our experimental results also confirmed that by combining information from subjective listening test and WER scores, along with objective STOI metrics, MTI-Net can more accurately predict intelligibility and WER scores. In the future, we will investigate using MTI-Net as a learnable objective function to guide SE model training.

6. References

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

[2] A. S. S. 1997, "Methods for calculation of the speech intelligibility index," in *Acoustical Society of America*, 1997.

[3] T. Houtgast and H. . M. Steeneken, "Evaluation of speech transmission channels by using artificial signals," *Acustica*, vol. 25, no. 6, pp. 355–367, 1971.

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

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

[6] T. H. Falk, C. Zheng, and W. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[7] F. Chen, O. Hazrati, and P. C. Loizou, "Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure," *Biomedical Signal Processing and Control*, vol. 8, no. 3, pp. 311–314, 2012.

[8] X. Jia and D. Li, "A deep learning-based time-domain approach for non-intrusive speech quality assessment," in *Proc. APSIPA ASC*, 2020, pp. 477–481.

[9] X. Dong and D. S. Williamson, "An attention enhanced multi-task model for objective speech assessment in real-world environments," in *Proc. ICASSP*, 2020, pp. 911–915.

[10] 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 *Proc. APSIPA ASC*, 2020, pp. 482–486.

[11] A. H. Andersen, J. M. D. Haan, Z. H. Tan, and J. Jensen, "Nonintrusive speech intelligibility prediction using convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1925–1939, 2018.

[12] M. B. Pedersen, A. H. Andersen, S. H. Jensen, and J. Jensen, "A neural network for monaural intrusive speech intelligibility prediction," in *Proc. ICASSP*, 2020, pp. 336–340.

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

[14] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features," *arXiv:2110.02635*, 2022.

[15] 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*, 2021.

[16] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[17] X. Dong and D. S. Williamson, "Towards real-world objective speech quality and intelligibility assessment using speech-enhancement residuals and convolutional long short-term memory networks," *The Journal of the Acoustical Society of America*, vol. 148, no. 5, pp. 3348–3359, 2020.

[18] H.-T. Chiang, Y.-C. Wu, C. Yu, T. Toda, H.-M. Wang, Y.-C. Hu, and Y. Tsao, "Hasa-net: A non-intrusive hearing-aid speech assessment network," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 907–913.

[19] J. M. Kates and K. H. Arehart, "The hearing-aid speech quality index (hasqi) version 2," *Journal of the Audio Engineering Society*, vol. 62, no. 3, pp. 99–117, 2014.

[20] ——, "The hearing-aid speech perception index (haspi)," *Speech Communication*, vol. 65, pp. 75–93, 2014.

[21] S. Liu, A. Mallol-Ragolta, E. Parada-Cabeleiro, K. Qian, X. Jing, A. Kathan, B. Hu, and B. W. Schuller, "Audio self-supervised learning: A survey," *arXiv preprint arXiv:2203.01205*, 2022.

[22] E. Dunbar, M. Bernard, N. Hamilakis, T. A. Nguyen, M. de Seyssel, P. Rozé, M. Rivière, E. Kharitonov, and E. Dupoux, "The Zero Resource Speech Challenge 2021: Spoken Language Modelling," in *Proc. INTERSPEECH*, 2021, pp. 1574–1578.

[23] S. Evain, H. Nguyen, H. Le, M. Z. Boito, S. Mdhaffar, S. Alisamir, Z. Tong, N. Tomashenko, M. Dinarelli, T. Parcollet, A. Allauzen, Y. Estève, B. Lecouteux, F. Portet, S. Rossato, F. Ringeval, D. Schwab, and L. Besacier, " LeBenchmark: A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech," in *Proc. INTERSPEECH*, 2021, pp. 1439–1443.

[24] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin *et al.*, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. INTERSPEECH*, 2021, pp. 1194–1198.

[25] Z. Huang, S. Watanabe, S.-w. Yang, P. Garcia, and S. Khudanpur, "Investigating self-supervised learning for speech enhancement and separation," *arXiv preprint arXiv:2203.07960*, 2022.

[26] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *arXiv preprint arXiv:2110.13900*, 2021.

[27] H. Nguyen, F. Bougares, N. Tomashenko, Y. Estève, and L. Besacier, "Investigating self-supervised pre-training for end-to-end speech translation," in *Proc. INTERSPEECH*, 2020, pp. 1466–1490.

[28] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[29] E. Cooper, W.-H. Huang, T. Toda, and J. Yamagishi, "Generalization ability of mos prediction networks," in *Proc. ICASSP*, 2022.

[30] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 2, pp. 87–95, 2001.

[31] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[32] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *Proc. APSIPA ASC*, 2017.

[33] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. INTERSPEECH*, 2013, pp. 436–440.

[34] J. Kim, M. El-Khamy, and J. Lee, "T-GSA: Transformer with Gaussian-weighted self-attention for speech enhancement," in *Proc. ICASSP*, 2020, pp. 6649–6653.

[35] A. Zhang, "Speech recognition (version 3.6) [software], available: https://github.com/uberi/speech_recognition#readme," in *Proc. ICCC*, 2017.

[36] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.