HASA-NET: A NON-INTRUSIVE HEARING-AID SPEECH ASSESSMENT NETWORK

Hsin-Tien Chiang[1], Yi-Chiao Wu[2], Cheng Yu[1], Tomoki Toda[2], Hsin-Min Wang[3], Yih-Chun Hu[4], Yu Tsao[1]

[1]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[2]Information Technology Center, Nagoya University, Japan
[3]Institute of Information Science, Academia Sinica, Taiwan
[4]University of Illinois at Urbana-Champaign, USA

ABSTRACT

Non-intrusive speech assessment methods have garnered significant attention in making objective evaluations, largely because they do not need a clean reference. Recently, deep neural network (DNN) models have been applied to build non-intrusive speech assessment approaches and have been shown to provide promising performance. However, most DNN-based approaches are designed for normal-hearing listeners without considering hearing-loss factors. In this study, we propose a DNN-based hearing aid speech assessment network (HASA-Net), formed by a bidirectional long short-term memory (BLSTM) model, to predict speech quality and intelligibility scores simultaneously based on input speech signals and specified hearing-loss patterns. To the best of our knowledge, HASA-Net is the first work to incorporate quality and intelligibility assessments utilizing a unified DNN-based non-intrusive model for hearing aids. Experimental results show that the speech quality and intelligibility scores predicted by HASA-Net are highly correlated to two well-known intrusive hearing-aid evaluation metrics, hearing aid speech quality index (HASQI) and hearing aid speech perception index (HASPI), respectively.

Index Terms — objective metrics, hearing loss, end-to-end, non-intrusive, multi-task learning

1. INTRODUCTION

Speech quality and intelligibility assessments serve as important tools to many speech-related applications. Speech quality indicates pleasantness or naturalness of a speech signal, and speech intelligibility measures how well the content of the speech can be understood. A straightforward approach to measure speech quality or intelligibility is to conduct subjective listening tests. In such tests, speech signals are played to a group of listeners, who are then asked to score the quality or provide recognized words of the heard speech signals. To avoid potential assessment biases, a large number of subjects is generally required. However, conducting listening tests on many subjects is time-consuming and prohibitively expensive. Therefore, objective speech quality and intelligibility assessments are developed based on such studies and used as surrogates to the subjective listening tests.

Objective speech assessments can be roughly divided into two categories: intrusive and non-intrusive. Intrusive methods [1–6] use the clean speech as reference to compare with the degraded/processed one to give the output index. However, clean speech may not always be available and thus restrict practicality of intrusive methods [7]. On the other hand, non-intrusive methods [8–11] are capable of calculating the output index directly on the degraded/processed speech without reference data. Without the requirement of reference data, non-intrusive methods have been widely adopted in online assessment and real-world applications.

Recently, deep neural network (DNN) models have been used as a fundamental tool for speech quality and intelligibility assessment [12–18]. For speech quality, Quality-Net [13] was proposed as an end-to-end, non-intrusive speech quality evaluation model; Quality-Net is a BLSTM-based model and capable of predicting perceptual evaluation of speech quality (PESQ) scores for noisy/processed speech signals. MOSNet [14] was designed to predict mean opinion score (MOS) for converted speech. Later on, MBNet [18], which consists of a MeanNet and a BiasNet, was proposed as an improved version of MOSNet, considering score variations caused by personal preferences. For speech intelligibility, STOI-Net [17] utilizes CNN-BLSTM with a multiplicative attention mechanism to predict speech intelligibility. A unified model that estimates multiple objective speech quality and intelligibility scores is developed and presented in [16].

Despite the recent attention on utilizing DNN for objective speech evaluation, limited research considers hearing-impaired listeners; that is, users with assistive listening devices, such as hearing aids and cochlear implants. Hearing loss is ranked as the fourth highest cause of disability globally [19] and refers to the total or partial inability to hear. The impacts of hearing loss have profound effects on ability to communicate. Approximately 80 percent of people were living with varying degrees of hearing loss in 2018 [20]. The prevalence of hearing loss in the United States doubles with

every ten year increase in age [19]. In order to broaden the impact of speech research to include the elderly and hearing-impaired, research with hearing impaired listeners are of increased importance. So far, numerous advanced signal processing methods have been tested on scenarios of hearing impaired listeners [21–24]. An effective speech quality and intelligibility assessment system for varying hearing loss patterns thus becomes an essential tool to further develop suitable speech signal processing methods for assistive listening devices.

In this study, we propose a novel deep-learning based end to-end, non-intrusive hearing aid speech assessment model, which we call HASA-Net. HASA-Net is a unified model capable of simultaneously estimating quality and intelligibility given input speech and specified hearing-loss patterns. Specifically, HASA-Net aims to predict HASQI [4] and HASPI [5] scores, which are two well-known evaluation metrics for speech quality and intelligibility designed for hearing aids. HASQI and HASPI are based on an auditory model that incorporates changes due to hearing loss. While both HASQI and HASPI are based on a calculation comparing the noisy signal with its clean reference, HASA-Net predicts scores *without the need of clean reference*. To the best of our knowledge, HASA-Net represents the first step towards an end-to-end, non-intrusive unified model for hearing aids that simultaneously measures quality and intelligibility.

The remainder of this paper is organized as follows. We describe the details of HASA-Net in Section 2. The experimental setup and results are presented in Section 3. Finally, we conclude this work in Section 4.

2. HASA-NET

HASA-Net is an objective speech assessment model adopting multi-task learning and multi-head attention mechanism. We first introduce multi-task learning, multi-head attention mechanism, the overall framework of HASA-Net, and the loss function.

2.1. Multi-task learning

Multi-task learning (MTL) [25] has been widely adopted in speech-related applications [26–29]. The aim of MTL is to simultaneously learn multiple related tasks with a unified model to obtain more robust shared feature representations, thereby improving the overall performance for all tasks. In this paper, we exploit a MTL structure to HASA-Net to jointly predict speech quality and intelligibility scores.

2.2. Multi-head attention mechanism

When considering the significance of each speech frame within the input utterances to the target task, the attention mechanism has shown promising effectiveness in making

models focus on important information of feature maps. For example, silence frames are less informative and should be given less attention (thus giving lighter weights), whereas frames containing speech are more informative and should be given higher attention (thus giving higher weights) for speech emotion recognition [30]. In this study, we use the self-attention mechanism to learn different weighted combinations of all time frames in an input sequence. Specifically, an output of the self-attention layer is the weighted sum of the input sequence by calculating every attention weight based on the dependencies between the current frame and other frames within the input sequence. In contrast to single-head attention, multi-head attention [31] is adopted to capture information from multiple representation subspaces in order to enhance model's expressiveness. In short, the core idea of incorporating multi-head self-attention in HASA-Net is to emphasize different frame-wise features when predicting speech quality and intelligibility scores.

2.3. Network architecture

Fig.1 illustrates the overall architecture of HASA-Net. The input of HASA-Net is composed of concatenated Time-frequency (T-F) features and hearing-loss patterns. T-F features are extracted using a 512-point short time Fourier transform (STFT) with a Hamming window size of 512 points and a hop size of 256 points, resulting in a 257-dimension magnitude spectrum. HASA-Net consists of a stack of one bidirectional LSTM with 100 nodes, followed by one dense layer with 128 rectified linear units (ReLU) nodes. By sharing the bidirectional LSTM and dense layers between all tasks, the remaining layers are split into two separate tasks. One task is for quality estimation and the other is for intelligibility prediction. In each task, a multi-head attention mechanism is first applied to aggregate task-specific information based on the shared features. After the multi-head attention, a dense layer with one node activated with a sigmoid function is applied. The output of the dense layer is the predicted score for each frame. Finally, we obtain the final utterance-level score by calculating the global average pooling operation based on frame scores.

2.4. Objective function

The loss function for each utterance is the summation of utterance-level loss and averaged frame-wise loss. We formulate the loss for quality estimation as follows:

$$L_{Quality} = \frac{1}{N} \sum_{n=1}^{N} [(\hat{Q}_n - Q_n)^2 + \frac{1}{T_n} \sum_{t=1}^{T_n} (\hat{Q}_n - q_{n,t})^2]$$

where $\hat{Q}_n$ and Q_n represent the true and estimated quality scores for the n-th utterance, respectively, while N represents the total training utterances and T_n is the number of frames
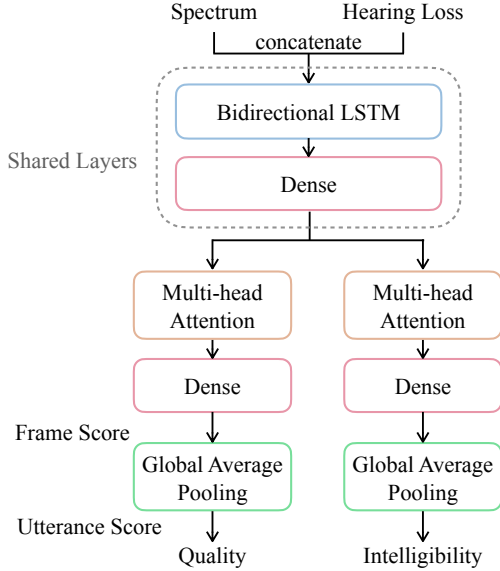
Fig. 1. Architecture of the HASA-Net model.

in utterance n. Meanwhile, $q_{n,t}$ denotes the estimated frame quality score of the t-th frame of utterance n. The loss function for the intelligibility estimator is identical to that of quality. Similarly, the loss function for intelligibility estimation is then given as:

$$L_{Intelligibility} = \frac{1}{N} \sum_{n=1}^{N} [(\hat{I}_n - I_n)^2 + \frac{1}{T_n} \sum_{t=1}^{T_n} (\hat{I}_n - i_{n,t})^2]$$

where $\hat{I}_n$ and I_n denote the true and estimated intelligibility scores for the n-th utterance, and $i_{n,t}$ stands for the estimated frame intelligibility score of the t-th frame of utterance n.

The overall loss is the sum of the losses for the two tasks:

$$L_{Total} = \alpha \times L_{Quality} + \beta \times L_{Intelligibility}$$

where α and β represent the weights for the two tasks. According to our internal evaluations, we empirically set α to 1.0 and β to 1.5.

3. EXPERIMENTS

3.1. Dataset

We conducted experiments using the TIMIT database [32]. All 4620 utterances from the training set of TIMIT were utilized for training. The noisy set was generated by corrupting the utterances with 80 noises combined at seven SNR values: -15, -10, -5, 0, 5, 10 and 15 dB. These noise signals were obtained from the 100 noises dataset [33]. In addition, each utterance was corrupted with one noise type at one SNR level so that all the training data was unparalleled.

For test data, we randomly selected 100 clean speech signals from the test set of the TIMIT database. These selected utterances were then mixed with four unseen noise types (engine, white, street and baby cry) at four SNRs levels:-6,0,6 and 12 dB. In total, the testing data comprises 1200 utterances.

3.2. Hearing-loss patterns

Hearing loss is detected through an audiogram, which is a graphical display showing the degrees of hearing loss at different frequency regions [34]. A threshold at any frequency above 20 dB is considered as a hearing loss. The hearing loss patterns are selected at frequencies of 250, 500, 1000, 2000, 4000, and 6000 Hz. Each audiogram is assigned to one of the six categories flat, sloping, rising, cookie-bite, noise-notched and high-frequency [35]. Different configurations of hearing loss are shown in Fig. 2.

In the flat category (Fig. 2(a)), there are no change in thresholds across frequencies. Audiograms with sloping configurations have increased thresholds from low to high frequencies (Fig. 2(b)). The rising category with thresholds decrease as frequencies increased (Fig. 2(c)). A cookie-bite audiogram has greatest thresholds in the mid frequencies between 500 and 4000 Hz (Fig. 2(d)). A noise-notch configuration reaches a maximum threshold in frequency between 3000 and 6000 Hz (Fig. 2(e)). High-frequency category has hearing loss occurs at frequencies between 2000 and 8000 Hz, while frequencies below 2000 Hz are not affected (Fig. 2(f)).

In total, we have collected 42 hearing loss patterns, where each hearing loss category contained 7 patterns. Every category was further divided into two groups, the larger group with 5 patterns and the smaller one with the remaining 2 patterns.
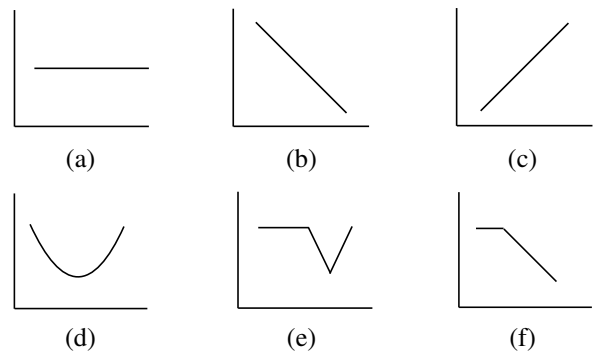

Fig. 2. Hearing loss configurations. (a) Flat (b) Sloping (c) Rising (d) Cookie-bite (e) Noise-notched (f) High-frequency.

3.3. Experimental setup

As mentioned earlier, HASA-Net takes the magnitude spectrum and the hearing-loss patterns as its inputs and outputs

the corresponding quality and intelligibility scores. Given a training utterance, two hearing-loss patterns from the mentioned larger group were selected to form the corresponding training set, which resulted in 4360×2 training utterances in every category.

To evaluate HASA-Net's generalization capability, we considered two scenarios: seen and unseen test sets. The seen test sets utilized the hearing-loss patterns selected from the larger groups and were already presented in the training, whereas the unseen test set contained patterns from the smaller groups that were not used in the training. Each category included two hearing-loss patterns for seen and unseen test sets, and both of them included the same testing utterances. Therefore, there were 14,400 utterances in each of the seen and unseen test sets.

The corresponding ground-truth values for quality and intelligibility of HASA-Net were those calculated by HASQI and HASPI, respectively. HASQI and HASPI are based on an auditory model that incorporates changes due to hearing loss, and then develop mathematical models that match the corresponding quality and intelligibility scores. Both of their metrics are scalars between 0 and 1, where a higher score of HASQI and HASPI represents better speech quality and intelligibility. The stimuli were amplified using the National Acoustics Laboratories revised (NAL-R) [36] linear fitting prescriptive formula based on individual hearing-loss patterns.

We use the RMSprop [37] optimizer with a learning rate of 0.001 and early stopping technique to train HASA-Net. To evaluate the performance, various criteria including mean square error (MSE), linear correlation coefficient (LCC), and Spearman's rank correlation coefficient (SRCC) are selected. In the follows, we first show the effects of the multi-head attention mechanism introduced in Section 2.

3.4. Effects of multi-head attention mechanism

We present the effects of multi-head attention in HASA-Net and set up a baseline model that we call HASA-Net$_{DNN}$. HASA-Net$_{DNN}$ replaces the multi-head attention mechanism of HASA-Net with dense layers. Tables 1 and 2 clearly show that HASA-Net outperforms HASA-Net$_{DNN}$ consistently on the seen and unseen test sets across quality and intelligibility predictions. For the quality estimation, HASA-Net outperforms HASA-Net$_{DNN}$ with notable {LCC, SRCC} improvements of {0.0189, 0.0198} and {0.0206, 0.0197} for the seen and unseen test sets, respectively. In addition, greater improvements are observed in the intelligibility predictions. For instance, HASA-Net yields {LCC, SRCC} improvements of {0.0536, 0.0386} as compared to HASA-Net$_{DNN}$ on the seen test set. For the unseen test set, HASA-Net also outperforms HASA-Net$_{DNN}$ with noticeable {LCC, SRCC} improvements of {0.0469 and 0.026}. The significant performance improvements of HASA-Net confirms the effectiveness of

multi-head attention mechanism for the proposed score predictions.

Table 1. Performance of quality prediction between HASA-Net$_{DNN}$ and HASA-Net on seen and unseen test sets.

	Seen			Unseen		
	MSE	LCC	SRCC	MSE	LCC	SRCC
HASA-Net$_{DNN}$	0.0170	0.9092	0.9074	0.0177	0.9101	0.9097
HASA-Net	**0.0108**	**0.9281**	**0.928**	**0.0108**	**0.9299**	**0.9294**

Table 2. Performance of the intelligibility prediction between HASA-Net$_{DNN}$ and HASA-Net on seen and unseen test sets.

	Seen			Unseen		
	MSE	LCC	SRCC	MSE	LCC	SRCC
HASA-Net$_{DNN}$	0.0422	0.8264	0.8783	0.0444	0.8195	0.8739
HASA-Net	**0.0193**	**0.88**	**0.9169**	**0.022**	**0.8664**	**0.8999**

3.5. Detailed assessment results

Next, we present the detailed assessment results of HASA-Net for quality and intelligibility on the seen and unseen test sets. Figs. 3 and 4 show the scatter plots with the corresponding quality and intelligibility metrics. Each figure shows results from the seen and unseen test sets. The higher the LCC and SRCC, the more accurately HASA-Net predicts the corresponding ground truths, indicating higher prediction accuracies. On average, HASA-Net achieves LCC of 0.9281 and SRCC of 0.9280 on the seen test set for the quality estimation, compared to 0.9299 and 0.9294 on the unseen test set. For the intelligibility prediction, {LCC, SRCC} are {0.88, 0.9169} and {0.8664, 0.8999} on the seen and unseen test sets, respectively. In our experiments, HASA-Net shows a strong correlation between the prediction scores and the ground truths with LCC and SRCC above 0.9 for the quality estimation, and a decrease around 0.86 to 0.92 for the intelligibility estimation.
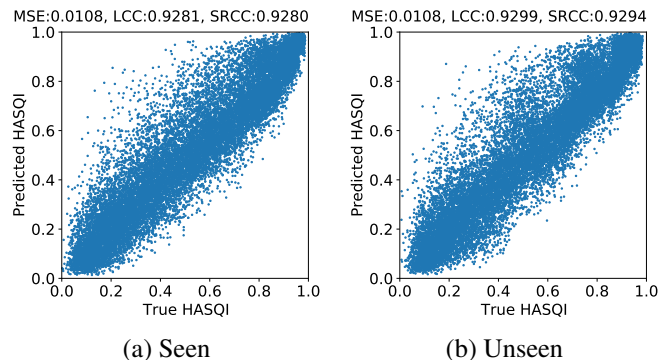

(a) Seen   (b) Unseen

Fig. 3. Scatter plots for quality assessment by HASA-Net.
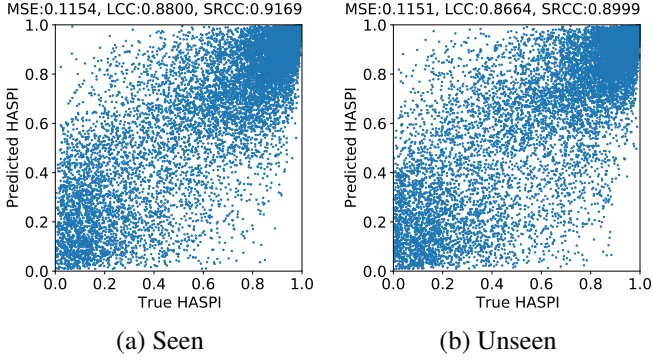
(a) Seen   (b) Unseen

Fig. 4. Scatter plots for the intelligibility assessment by HASA-Net.

The above results show that intelligibility is more difficulty to evaluate than quality, which may be caused by differences in the sensitivity of HASQI and HASPI. The authors in [38] pointed out that HASQI is much more sensitive to noise and show that at 10 dB SNR level, HASPI is 0.99 which is almost close to 1, whereas HASQI is only about 0.3. This is consistent with what we have found in our training set where HASPI scores are much more likely to be close to 1 than HASQI scores. Fig. 5 illustrate the percentage allocation of quality and intelligibility scores in the training set. Provided that the same spectrum and hearing-loss patterns are used, HASQI scores are distributed evenly from 0 to 1, while around 14% and 54% of HASPI scores fall in the range of 0 to 0.1 and 0.9 to 1. This uneven HASPI score distribution of training set brings difficulty to training process. Although HASA-Net achieves slightly lower prediction performance of intelligibility compared to that of quality, it is still promising that HASA-Net has a significant chance to estimate values close to the true intelligibility with average LCC and SRCC of 0.8732 and 0.9084.

Table 3 and 4 provide further details of quality and intelligibility prediction performance based on different types of hearing loss. As mentioned above, HASA-net's prediction of intelligibility is not as good as its prediction of quality as measured by LCC and SRCC on the seen and unseen test sets. However, in the worse case, HASA-Net still achieves {LCC, SRCC} around {0.8148, 0.8471} for the unseen rising hearing-loss patterns, which indicates that the estimated scores still follows the trend of the true scores well. In addition, similar tendencies can be found in the quality and intelligibility estimation. We observe that sloping and high-frequency configurations show higher LCCs and SRCCs, while lower LCCs and SRCCs are found in rising and cookie-bite configurations.

Finally, HASA-Net achieves comparable results in both quality and intelligibility on the seen and unseen test sets across different types of hearing loss. Not surprisingly, the seen test set outperforms the unseen for most types of hearing
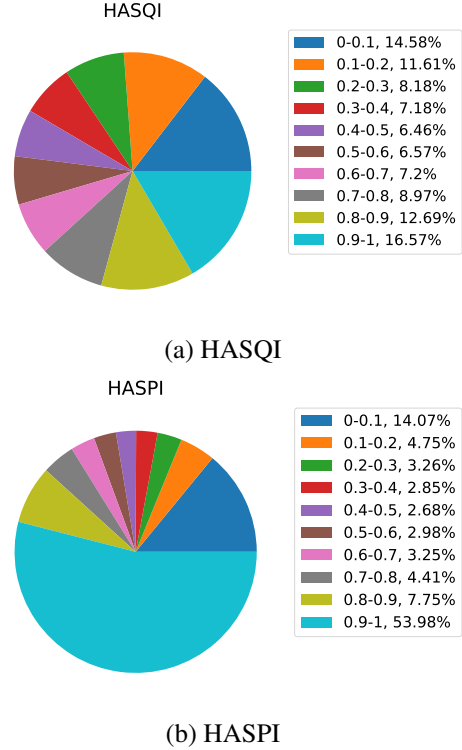

(a) HASQI

(b) HASPI

Fig. 5. Pie charts depicting the percentage allocation of quality and intelligibility scores according to training set.

loss. However, when considering the challenging test conditions where the hearing-loss patterns are totally unseen, the performance of HASA-Net continues to be quite strong, proving the robustness and generalization capability of HASA-Net. In summary, our detailed analysis confirms the effectiveness of HASA-Net.

Table 3. Performance of quality prediction of different types of hearing loss on seen and unseen test sets.

	Seen			Unseen		
Configuration	MSE	LCC	SRCC	MSE	LCC	SRCC
Flat	0.0113	0.9192	0.9155	0.0124	0.9186	0.915
Sloping	0.0067	0.9557	0.9421	0.0076	0.9517	0.9443
Rising	0.0138	0.8965	0.8908	0.0152	0.8897	0.8885
Cookie-bite	0.0156	0.8857	0.8875	0.0135	0.9118	0.9106
Noise-notched	0.0078	0.9434	0.9413	0.0088	0.9325	0.93
High-frequency	0.0094	0.9399	0.9373	0.0073	0.9527	0.9474

3.6. Single-task versus HASA-Net

In this section, we analyze the role of MTL. A single-task model adopts the same architecture of HASA-Net, but outputs only one objective prediction. Table 5 and 6 show the comparison results between the single-task model and HASA-Net. From the tables, we can see that HASA-Net

Table 4. Performance of intelligibility prediction of different types of hearing loss on seen and unseen test sets.

Configuration	Seen			Unseen		
	MSE	LCC	SRCC	MSE	LCC	SRCC
Flat	0.0222	0.8237	0.8996	0.0247	0.8549	0.92
Sloping	0.0152	0.9128	0.889	0.0128	0.9179	0.9023
Rising	0.0338	0.8495	0.8593	0.0369	0.8148	0.8471
Cookie-bite	0.0176	0.8129	0.8947	0.029	0.8284	0.868
Noise-notched	0.0118	0.9163	0.9113	0.0125	0.8763	0.8884
High-frequency	0.0151	0.9068	0.9102	0.0161	0.91	0.9041

produces better results compared to the single-task model. For quality prediction, the results of LCC and SRCC for the single-task model are comparable with that of HASA-Net on both seen and unseen test sets. A possible reason is that the quality prediction already yields good performance, and thus has less room for further improvement. On the other hand, notable improvement from the single-task model to HASA-Net can be noted in the intelligibility estimation. The results show that MTL helps HASA-Net to learn more robust and universal feature representations from different tasks. The MTL scheme not only reduces overfitting to a specific task but also boosts the intelligibility prediction performance. Specifically, HASA-Net produces {LCC, SRCC} improvements of {0.0377, 0.0292} and {0.029, 0.0272} for the seen and unseen test sets, respectively, as compared to the single-task model. In contrast to the single-task model, an additional advantage of HASA-Net is its ability to predict quality and intelligibility scores simultaneously. For applications with limited computation and storage resources, HASA-Net is definitely a better choice than multiple single-task models.

Table 5. Performance of quality prediction between single-task and HASA-Net on seen and unseen test sets.

	Seen			Unseen		
	MSE	LCC	SRCC	MSE	LCC	SRCC
Single-task	0.0132	0.9213	0.9221	0.0133	0.9237	0.9251
HASA-Net	**0.0108**	**0.9281**	**0.928**	**0.0108**	**0.9299**	**0.9294**

Table 6. Performance of intelligibility prediction between single-task and HASA-Net on seen and unseen test sets.

	Seen			Unseen		
	MSE	LCC	SRCC	MSE	LCC	SRCC
Single-task	0.0269	0.8423	0.8877	0.0285	0.8374	0.8727
HASA-Net	**0.0193**	**0.88**	**0.9169**	**0.022**	**0.8664**	**0.8999**

4. CONCLUSION

In this study, we have proposed a DNN-based speech assessment model for listeners using hearing aids, which we call HASA-Net. To the best of our knowledge, this is the first end-to-end, non-intrusive unified model predicting hearing aid speech quality and intelligibility simultaneously. Our experimental results show that HASA-Net's quality and intelligibility estimates are highly correlated with HASQI and HASPI scores. We also demonstrated the superior performance of HASA-Net as compared to the single-task model. The end-to-end framework gave HASA-Net the capability to directly combine with DNN-based speech signal processing models for joint optimization. HASA-Net provides a significant building block towards tackling various complex challenges for the hearing-impaired, such as speech enhancement, dereverberation, and separation. Future work includes listening tests with hearing-impaired listeners to investigate correlation of HASA-Net and human ratings.

5. REFERENCES

[1] Antony W Rix, John G Beerends, Michael P Hollier, and Andries 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)*. IEEE, 2001, vol. 2, pp. 749–752.

[2] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

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

[4] James M Kates and Kathryn 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.

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

[6] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey, "Sdr–half-baked or well done?," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[7] Md Rafidul Islam, Md Ashequr Rahman, Md Numan Hasan, ANM Shahriyar Hossain, Ahmed Nazim Uddin, and Mohammad Ariful Haque, "Non-intrusive objective

evaluation of speech quality in noisy condition," in *2016 9th International Conference on Electrical and Computer Engineering (ICECE)*. IEEE, 2016, pp. 586–589.

[8] Ludovic Malfait, Jens Berger, and Martin Kastner, "P. 563—the itu-t standard for single-ended speech quality assessment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1924–1934, 2006.

[9] Doh-Suk Kim and Ahmed Tarraf, "Anique+: A new american national standard for non-intrusive estimation of narrowband speech quality," *Bell Labs Technical Journal*, vol. 12, no. 1, pp. 221–236, 2007.

[10] Tiago H Falk, Chenxi Zheng, and Wai-Yip 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.

[11] Charlotte Sørensen, Angeliki Xenaki, Jesper B Boldt, and Mads G Christensen, "Pitch-based non-intrusive objective intelligibility prediction," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 386–390.

[12] Asger Heidemann Andersen, Jan Mark De Haan, Zheng-Hua Tan, and Jesper 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.

[13] Szu-Wei Fu, Yu Tsao, Hsin-Te Hwang, and Hsin-Min Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on blstm," *arXiv preprint arXiv:1808.05344*, 2018.

[14] Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang, "Mosnet: Deep learning based objective assessment for voice conversion," *arXiv preprint arXiv:1904.08352*, 2019.

[15] Xuan Dong and Donald S Williamson, "A classification-aided framework for non-intrusive speech quality assessment," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 100–104.

[16] Xuan Dong and Donald S Williamson, "An attention enhanced multi-task model for objective speech assessment in real-world environments," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 911–915.

[17] Ryandhimas E Zezario, Szu-Wei Fu, Chiou-Shann Fuh, Yu Tsao, and Hsin-Min 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)*. IEEE, 2020, pp. 482–486.

[18] Yichong Leng, Xu Tan, Sheng Zhao, Frank Soong, Xiang-Yang Li, and Tao Qin, "Mbnet: Mos prediction for synthesized speech with mean-bias network," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 391–395.

[19] Lisa L Cunningham and Debara L Tucci, "Hearing loss in adults," *New England Journal of Medicine*, vol. 377, no. 25, pp. 2465–2473, 2017.

[20] Bolajoko O Olusanya, Adrian C Davis, and Howard J Hoffman, "Hearing loss grades and the international classification of functioning, disability and health," *Bulletin of the World Health Organization*, vol. 97, no. 10, pp. 725, 2019.

[21] Ying-Hui Lai, Fei Chen, Syu-Siang Wang, Xugang Lu, Yu Tsao, and Chin-Hui Lee, "A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1568–1578, 2016.

[22] Natalie Yu-Hsien Wang, Hsiao-Lan Sharon Wang, Tao-Wei Wang, Szu-Wei Fu, Xugan Lu, Hsin-Min Wang, and Yu Tsao, "Improving the intelligibility of speech for simulated electric and acoustic stimulation using fully convolutional neural networks," *IEEE Transactions on Neural Systems and Rehabilitation Engineering: a Publication of the IEEE Engineering in Medicine and Biology Society*, 2020.

[23] Eric W Healy, Jordan L Vasko, and DeLiang Wang, "The optimal threshold for removing noise from speech is similar across normal and impaired hearing—a time-frequency masking study," *The Journal of the Acoustical Society of America*, vol. 145, no. 6, pp. EL581–EL586, 2019.

[24] Soha A Nossier, MRM Rizk, Nancy Diaa Moussa, and Saleh el Shehaby, "Enhanced smart hearing aid using deep neural networks," *Alexandria Engineering Journal*, vol. 58, no. 2, pp. 539–550, 2019.

[25] Rich Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[26] Zhizheng Wu, Cassia Valentini-Botinhao, Oliver Watts, and Simon King, "Deep neural networks employing multi-task learning and stacked bottleneck features

for speech synthesis," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 4460–4464.

[27] Zhuo Chen, Shinji Watanabe, Hakan Erdogan, and John R Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[28] Yusuke Shinohara, "Adversarial multi-task learning of deep neural networks for robust speech recognition.," in *Interspeech*. San Francisco, CA, USA, 2016, pp. 2369–2372.

[29] Suyoun Kim, Takaaki Hori, and Shinji Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4835–4839.

[30] Kyoung Ju Noh, Chi Yoon Jeong, Jiyoun Lim, Seungeun Chung, Gague Kim, Jeong Mook Lim, and Hyuntae Jeong, "Multi-path and group-loss-based network for speech emotion recognition in multi-domain datasets," *Sensors*, vol. 21, no. 5, pp. 1579, 2021.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[32] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett, "Getting started with the darpa timit cd-rom: An acoustic phonetic continuous speech database," *National Institute of Standards and Technology (NIST), Gaithersburgh, MD*, vol. 107, pp. 16, 1988.

[33] Yuxuan Wang and DeLiang Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[34] Roy D Patterson, Ian Nimmo-Smith, Daniel L Weber, and Robert Milroy, "The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold," *The Journal of the Acoustical Society of America*, vol. 72, no. 6, pp. 1788–1803, 1982.

[35] Waleed B Alshuaib, Jasem M Al-Kandari, and Sonia M Hasan, "Classification of hearing loss," in *Update On Hearing Loss*. IntechOpen, 2015.

[36] Denis Byrne and Harvey Dillon, "The national acoustic laboratories'(nal) new procedure for selecting the gain and frequency response of a hearing aid," *Ear and hearing*, vol. 7, no. 4, pp. 257–265, 1986.

[37] Tijmen Tieleman and Geoffrey Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.

[38] James M Kates, Kathryn H Arehart, Melinda C Anderson, Ramesh Kumar Muralimanohar, and Lewis O Harvey Jr, "Using objective metrics to measure hearing-aid performance," *Ear and hearing*, vol. 39, no. 6, pp. 1165, 2018.