N-MTTL SI Model: Non-Intrusive Multi-Task Transfer Learning-Based Speech Intelligibility Prediction Model with Scenery Classification

INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

N-MTTL SI Model: Non-intrusive Multi-task Transfer Learning-Based Speech Intelligibility Prediction Model with Scenery Classification

Ľuboš Marcinek[1], *Michael Stone*[2], *Rebecca Millman*[2], *Patrick Gaydecki*[1]

[1]Department of Electrical and Electronic Engineering, University of Manchester, UK
[2]Manchester Centre for Audiology and Deafness, University of Manchester, UK

lubos.marcinek@postgrad.manchester.ac.uk,
{michael.stone, rebecca.millman, patrick.gaydecki}@manchester.ac.uk

Abstract

The application of speech enhancement algorithms for hearing aids may not always be beneficial to increasing speech intelligibility. Therefore, a prior environment classification could be important. However, previous speech intelligibility models do not provide any additional information regarding the reason for a decrease in speech intelligibility. We propose a unique non-intrusive multi-task transfer learning-based speech intelligibility prediction model with scenery classification (N-MTTL SI model). The solution combines a Mel-spectrogram analysis of the degraded speech signal with transfer learning and multi-task learning to provide simultaneous speech intelligibility prediction (task 1) and scenery classification of ten real-world noise conditions (task 2). The model utilises a pre-trained ResNet architecture as an encoder for feature extraction. The prediction accuracy of the N-MTTL SI model for both tasks is high. Specifically, RMSE of speech intelligibility predictions for seen and unseen conditions is 3.76% and 4.06%. The classification accuracy is 98%. In addition, the proposed solution demonstrates the potential of using pre-trained deep learning models in the domain of speech intelligibility prediction.

Index Terms: speech intelligibility prediction, environmental sound classification, multi-task learning, transfer learning

1. Introduction

Speech perception decline is often associated with hearing impairment. People who wear a hearing aid (HA) experience difficulties understanding speech, especially in adverse acoustic scenarios containing various types of environmental noise [1]. The major complaints expressed by HA users are the insufficient reduction of environmental noise and the unexpected simultaneous amplification of noise with speech [2]. This problem has been approached by speech enhancement combined with noise-reduction algorithms for HAs. However, the benefits of speech enhancement algorithms may be perceived only in some acoustic environments. In other words, the application of the same speech enhancement algorithm can be beneficial in one environment but result in a negative impact on speech intelligibility and quality in a different environment. Diverse speech enhancement algorithms must be applied to different noisy environments to reduce noise adequately [2, 3]. Therefore, the automatic classification of acoustic environments in which speech enhancement would benefit a HA user is vital [4, 5]. In principle, this could be achieved by an objective speech intelligibility prediction metric running online in HAs [6].

Work connected to speech intelligibility prediction dates back several decades. Since then, many speech intelligibility models have been proposed based on different principles starting with the models such as Articulation Index (AI) [7, 8] and its successor SII [9] or STI [10]. Later on, more sophisticated models incorporated further elements of human auditory functioning (STOI [11], mr-sEPSM [12], HASPI [13]). The aforementioned objective models are intrusive – they require a reference signal to determine predictions of speech intelligibility, which is not available in real-world scenarios. On the contrary, non-intrusive speech intelligibility models (ModA [14], SRMR [15], NIC-STOI [6]) require only a degraded speech signal for speech intelligibility prediction. This advantage makes non-intrusive models more suitable for real-world applications [5]. Recently, rapid progress in the deep learning domain has also been utilised in the field of speech intelligibility prediction. Several models have been introduced employing DNN [16], CNN [17, 18] or U-Net [19]. However, none of the speech intelligibility models proposed so far provide any additional information regarding the cause of the speech intelligibility degradation that could be further used to fine-tune speech enhancement algorithms.

Therefore, this paper proposes a novel non-intrusive multi-task transfer learning-based speech intelligibility model (N-MTTL SI model) which provides speech intelligibility prediction along with scenery classification at the same time. N-MTTL SI determines speech intelligibility and classifies environmental noise by combining two common techniques from deep learning (multi-task learning and transfer learning). This unique model is based on a pre-trained ResNet architecture and uses Mel-spectrograms of degraded speech signals as an input. Since there is a scarcity of suitable speech intelligibility datasets with subjective scores, we used an objective speech intelligibility metric (STOI [11]) to label the dataset with intelligibility predictions. We therefore emphasize that the proposed solution is proof-of-concept, rather than a direct comparison with pre-existing speech intelligibility models.

2. Non-intrusive multi-task transfer learning-based speech intelligibility model

Multi-task learning (MTL) [20] is an approach in deep learning when the model performs at least two tasks. MTL has been successfully applied in various fields [20] including speech (e.g., speech recognition [21], speech enhancement [22], or objective speech assessment in real-world environments by generating several objective intelligibility and quality scores [23]).

http://dx.doi.org/10.21437/Interspeech.2021-1878

Sharing representations between related tasks in MTL can sometimes result in better performance than is observed in the single-task model. The most common type of multi-task learning is hard parameter sharing of hidden layers, which has been shown to reduce overfitting [20]. This approach is adopted in our N-MTTL SI model. The hidden layers are shared between all the tasks while preserving task-specific output layers. Our N-MTTL SI model (Figure 1) utilises pre-trained ResNet architecture on the ImageNet dataset, which is used as an encoder for feature extraction. Here, the transfer learning element occurs when the architecture (ResNet) trained on one dataset, and performing well on one task, is used to solve a task in a different domain. Such transfer learning has been explored in urban sound classification [24] but is novel when it comes to predictions of speech intelligibility. The ResNet architecture is adjusted (see 3.3) to perform two specific tasks: 1). speech intelligibility (SI) prediction and 2). classification of noise present in the degraded speech signal, i.e., scenery classification.

2.1. ResNet (Residual Networks)

The ResNet is one of the most popular and powerful architectures that scored very well in the ImageNet challenge [25] for image classification: It solved the gradient vanishing problem by introducing residual blocks with skip connection, the objective of which is to perform identity mapping. The entire architecture comprises several residual blocks stacked on top of each other, depending on the ResNet version. Each residual block contains a pair of convolutional layers (3x3) with the same number of filters. The number of output filters increases twofold with every two residual blocks. The residual block also contains a batch normalization layer and Rectified Linear Unit (ReLU) activation function, which is used after every convolutional layer [26]. Previous work [17] has explored the importance of the convolution layer to extract spectro-temporal patterns in the input signal related to speech intelligibility. Therefore, we expect the convolutional layers of ResNet to be beneficial for both our specific tasks in our N-MTTL SI model.

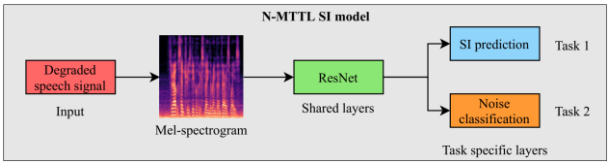

Figure 1: *Non-intrusive multi-task transfer learning-based speech intelligibility model.*

3. Experimental design

In this section, the dataset used for training, validation and testing of the N-MTTL SI model is described including the preprocessing performed. Additionally, we provide the details of the training procedure.

3.1. Dataset (Training, Validation and Test set)

The dataset [27, 28] consists of 28 speakers of British English (14 females and 14 males). It comprises ten noisy conditions: 1. speech-shaped noise, 2. babble (six speakers), 3. kitchen, 4. meeting room, 5. cafeteria, 6. restaurant, 7. subway station, 8. car, 9. metro, 10. busy traffic intersection. The conditions (3. - 10.) represent the real-world noise e.g., domestic noise (3.), office noise (4.), public spaces noise (5., 6.), transportation noise (7. - 9.), street noise (10.). These noisy conditions were produced at four different signal-to-noise ratios (SNRs): 0, 5, 10, 15 dB, resulting in 40 different conditions (10 noises x 4 SNRs) and approximately 10 different sentences per speaker in each condition. Overall, there are approximately 400 sentences per speaker. The dataset contains 11572 sentences that were divided into a training and a validation set using a ratio of 80% (9282 sentences) to 20% (2290 sentences). Sentences range in duration from 1 - 15 seconds. However, the vast majority of sentences (93%, 10776) are 2 - 4 seconds in duration. A separate test set consisting of 824 sentences spoken by two British English speakers (a female and a male) was used to test the proposed N-MTTL SI model on unseen conditions, which differ from conditions used for training the N-MTTL SI model. The unique set of noisy conditions in this set comprises 1. living room, 2. office space, 3. bus, 4. cafeteria (open area), 5. public square. These types of noise also represent real-world conditions: domestic noise (1.), office noise (2.), transport noise (3.) and street noise (4., 5.). Furthermore, the test set differs from the training and validation sets in terms of SNRs (2.5, 7.5, 12.5 and 17.5 dB). Overall, this makes for 20 conditions (5 types of noise x 4 SNRs) and approximately 20 sentences per speaker and condition.

3.2. Dataset preprocessing

During the preprocessing stage, we generated speech intelligibility labels using the STOI model based on clean and corresponding degraded sentences. STOI was chosen because its predictions correlate very well with actual intelligibility in noisy conditions resembling ones in the used dataset [11]. This model and its extended version (ESTOI [29]) were also used in previous studies to label speech data [23, 30]. The input data for our N-MTTL SI model were generated by the conversion of degraded sentences into Mel-spectrograms, which provided necessary visual representation for the model. The spectrograms were generated employing a Short-Time Fourier Transform (STFT) of 2048-length Hann-windowed speech signals with a 22.05 kHz sampling rate. Subsequently, windowed speech samples were extracted with a hop length of 512. Mel-scale mapping was done using Mel bands to obtain the Mel-spectrogram [31]. The resulting spectrograms were stored as 224 x 224 pixel images.

3.3. Training the N-MTTL SI model

Modifications to the ResNet architecture were performed before commencing training. Specifically, two task-related heads were added to the model so as to perform: 1. Regression to predict speech intelligibility and 2. Classification of background noise into one of ten classes (see noisy conditions in section 3.1.). In addition, the loss functions were defined for intelligibility prediction (Mean Square Error – MSE) and noise classification (Cross-Entropy). The weights for both specific tasks were learned and determined by the model considering the homoscedastic uncertainty of each task [32]. We used the Cyclical Learning Rate [33] technique to determine an efficient learning rate (LR) for training. The first three epochs, with maximum LR = 0.1, were trained using a frozen ResNet model when only added layers were trained. Subsequently, after unfreezing, ten more epochs with a discriminative learning rate (LR = 0.003 for early layers and LR = 0.03 for last layers) were run to fine-tune the pre-trained ResNet. During the training, a dropout rate of 25% was applied and the ADAM optimizer was used.

4. Experimental results

4.1. Speech intelligibility prediction

Results (Table 1) were obtained by running two versions of the N-MTTL SI model (ResNet-18 MT, ResNet-34 MT) using them as the feature encoder, compared with the single-task models (ResNet-18 ST, ResNet-34 ST) predicting only speech intelligibility based on the same architectures (ResNet-18, ResNet-34). Performance of the models is similar: We do not observe better prediction performance of the single-task models (ResNet-18 ST, ResNet-34 ST) in comparison to its multi-task counterparts (ResNet-18 MT, ResNet-34 MT). Considering the number of layers in ResNet-18 MT versus ResNet-34 MT, as well as the prediction accuracy of these models, we consider the ResNet-18 MT version of the N-MTTL SI model as the best solution. Specifically, the Root Mean Square Error (RMSE) values for the validation and test set obtained by ResNet-18 MT are similar (3.76% and 4.06%), which indicates that the model provides comparable speech intelligibility prediction for seen and unseen conditions. The linear relationship between predicted (ResNet-18 MT) and STOI-labelled intelligibility, as measured by Pearson correlation (r) is 0.93 for seen conditions and 0.86 for unseen (Figure 2). Spearman rank correlation (ρ) confirms monotonicity between the estimated and labelled intelligibility, which is 0.93/0.85 for the seen/unseen conditions, respectively. Lastly, Kendall's rank correlation coefficient, as used in the speech intelligibility prediction literature [17, 18], also expresses the degree of monotonicity in the relation between measurements and predictions. Kendall's (τ) values obtained by the N-MTTL SI model (ResNet-18 MT) are 0.80 and 0.69 for seen and unseen conditions, respectively. Overall, the accuracy measures of the model are comparable to the literature [17], where similar values are observed for unseen conditions. However, conditions in our dataset are not identical to [17], therefore direct comparison is not possible.

Table 1: *SI prediction of N-MTTL SI model (variants).*

N-MTTL SI model (variants)	Speech intelligibility prediction accuracy							
	Seen conditions				Unseen conditions			
	RMSE	r	ρ	τ	RMSE	r	ρ	τ
ResNet-18 ST	3.78	0.93	0.93	0.79	4.15	0.87	0.85	0.67
ResNet-18 MT	3.76	0.93	0.93	0.80	4.06	0.86	0.85	0.69
ResNet 34 ST	4.05	0.92	0.92	0.77	4.26	0.84	0.82	0.65
ResNet-34 MT	4.03	0.93	0.93	0.78	3.99	0.87	0.86	0.70

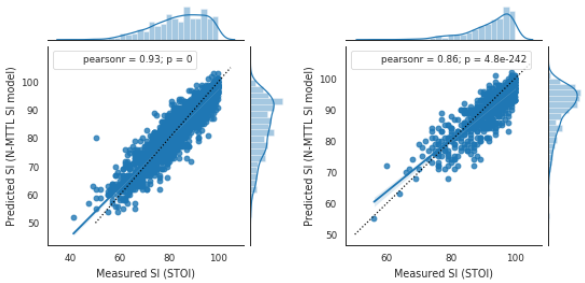

Figure 2: *Pearson correlation and data distribution for seen conditions (left) and unseen conditions (right). SI predictions by N-MTTL SI (ResNet-18 MT).*

4.1.1. Speech intelligibility prediction in seen conditions

We further examine the prediction performance achieved by the N-MTTL SI model (ResNet-18 MT) for individual conditions in the validation set and test set. Figure 3 shows the RMSE calculated for different noise conditions and SNRs in the validation set. The training and validation sets contain identical conditions. The largest cumulative RMSE can be observed in the babble condition, consistent with previous work [17], showing that babble is challenging for speech intelligibility predictions. We suspect that a component of informational masking (IM) from the six-talker babble may contribute to this result: IM is more complex in its operation than an energetic masker. The smallest masking effect is observed in the car noise condition, which is primarily low-frequency and where human hearing sensitivity is low. Car noise is a less effective masker than babble, which shares the spectro-temporal characteristics of speech.

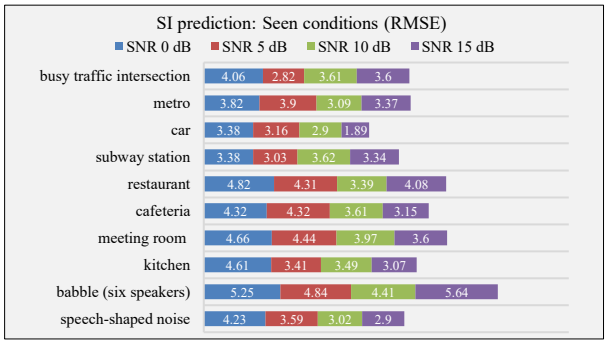

Figure 3: *Speech intelligibility predictions for seen conditions (validation set) obtained by N-MTTL SI model.*

4.1.2. Speech intelligibility prediction in unseen conditions

The same relative variability due to the nature of the background noise is observed in RMSE values calculated for unseen conditions in the test set (Figure 4). Specifically, the largest cumulative RMSE is also present in the environment where is it very likely that speech of other people occurs, e.g., a cafeteria (open space) and the spectro-temporal characteristics of single or multi-talker babble make them very effective maskers. On the other hand, an office space environment primarily includes noise resembling printer and paper noise, which are not very effective sources of masking. This condition shows the smallest RMSE and predictions are more precise.

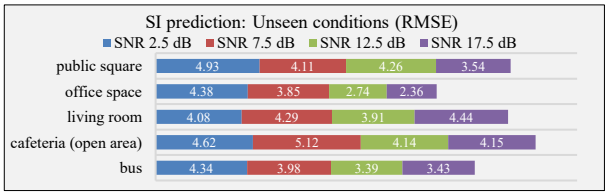

Figure 4: *Speech intelligibility predictions for unseen conditions (test set) obtained by N-MTTL SI model.*

4.2. Scenery classification

4.2.1. Scenery classification in seen conditions

A class was assigned to each of the ten different types of noise that were present in degraded speech signals used in the

training and validation set (3.1). The proposed N-MTTL SI model (ResNet-18 MT) can classify the type of noise into one of these classes with 98% accuracy. Several common performance measures (precision, recall, f1 score) are also very high (0.98). More details regarding the classification can be observed in the confusion matrix (Figure 5) which provides information about the number of correctly and incorrectly classified cases for the validation set. The number of incorrectly classified cases is 45 out of a total of 2290 cases (sentences). Specifically, the most frequently misclassified condition was metro noise with 12 incorrect classifications. This condition is followed by cafeteria (9 misclassifications), traffic and meeting room (6 misclassifications) conditions. On the other hand, speech-spectrum shaped noise (ssn) was 100% correctly classified.

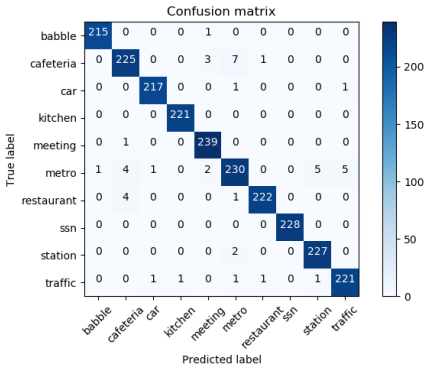

Figure 5: *Confusion matrix for the scenery classification task performed by N-MTTL SI model in the validation set.*

4.2.2. Scenery classification in unseen conditions

The classification performance of the N-MTTL SI model cannot be properly evaluated for the test set because the training and validation set do not contain identical noisy conditions. However, noise types in the test set originate from the same noise categories (office noise, transportation noise, street noise). Therefore, we visualised the actual labels for 824 sentences in the test set and the corresponding predicted labels (classes) in the form of a Sankey diagram (Figure 6) to explore if the model was able to pick up the characteristics of the similar but unseen types of noise as those used during the training of the model. Results show that sentences within the actual class public square (166 sentences) were classified into seven different types of noise from the training set. This effect might be caused by the variability of noises presented in the public square environment. Most sentences were classified as the metro (78) or traffic noise (58). Office space (164) was not classified with similar environments. Only a few sentences were predicted as being from a meeting room (20) and remaining sentences as, traffic (71), car noise (56), metro (13). In terms of the living room class (166), the vast majority of sentences were predicted as being from classes in a similar category, such as a meeting room (108). The outside cafeteria (open area) (164) was mostly classified as similar, namely cafeteria (117). Lastly, sentences containing bus noise (164) were mainly classified as car (76), metro noise (33) and traffic (30). These are all similar types of noise comprising transport noise. It is impressive, that the proposed N-MTTL SI model (ResNet-18 MT) can correctly assign similar classes for many sentences containing unseen types of noise (classes).

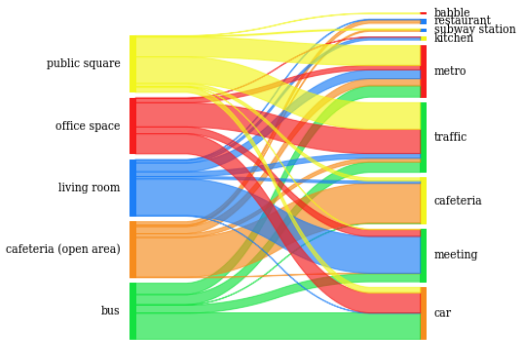

Figure 6: *Sankey diagram visualizing the classification performance for unseen noise conditions in the test set. The left side represents the actual classes and the right side predicted classes.*

5. Conclusion

The objective of this work was to develop a new model that would provide speech intelligibility prediction and information regarding a factor (type of noise – scenery classification) that contributes to degraded speech intelligibility. The proposed solution (N-MTTL SI model) combines transfer learning with multi-task learning and uses the Mel-spectrograms of degraded speech signals as the input. The N-MTTL SI model based on a pre-trained ResNet architecture achieved very high accuracy for simultaneous speech intelligibility predictions (RMSE 3.76%/4.06% for seen/unseen conditions) and noise classification (98%) considering the duration range of sentences (1 - 15 seconds) and 60 different conditions (noise types and SNRs). The classification accuracy is comparable to, or better than, seen in the best models providing either only environmental sound classifications [34], or sound classification in HAs [35].

Due to the scarcity of suitable speech intelligibility datasets with subjective scores, the STOI model was used to label the dataset with speech intelligibility estimations. This approach allowed the labelling of a large quantity of speech without the necessity to conduct time-consuming and expensive listening tests. Such labelling does not allow a direct comparison with existing speech intelligibility models. However, the scope of our work was proof-of-concept and comparisons would be possible if a suitable dataset of sufficient size becomes available in the future. Future work could also involve the extension of the model by including additional measures e.g., predictions based on speech reception thresholds, or both the speech envelope and temporal fine structure.

Finally, considering the continual development of technologies associated with HA devices, we believe that our solution could be deployed in the future. The accurate classification of acoustic scenes and the speech intelligibility predictions provided by the N-MTTL SI model would help to select a more suitable speech enhancement algorithm for HAs, leading to improved speech perception and potentially greater HA satisfaction.

6. Acknowledgements

This work is funded by the Medical Research Council (Grant No. MR/N013751/1, ref: 1916490) and the NIHR Manchester Biomedical Research Centre.

7. References

[1] J. M. Festen and R. Plomp, "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," *J. Acoust. Soc. Am.*, vol. 88, no. 4, pp. 1725–1736, 1990.

[2] G. Park, W. Cho, K. S. Kim, and S. Lee, "Speech enhancement for hearing aids with deep learning on environmental noises," *Appl. Sci.*, vol. 10, no. 17, 2020.

[3] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Commun.*, vol. 49, no. 7–8, pp. 588–601, 2007.

[4] V. Hamacher *et al.*, "Signal Processing in High-End Hearing Aids: State of the Art, Challenges, and Future Trends," *EURASIP J. Adv. Signal Process.*, vol. 2005, no. 18, p. 152674, 2005.

[5] T. H. Falk *et al.*, "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 114–124, Mar. 2015.

[6] C. Sørensen, M. S. Kavalekalam, A. Xenaki, J. B. Boldt, and M. G. Christensen, "Non-intrusive codebook-based intelligibility prediction," *Speech Commun.*, vol. 101, no. June, pp. 85–93, 2018.

[7] H. Fletcher and R. H. Galt, "The Perception of Speech and Its Relation to Telephony," *J. Acoust. Soc. Am.*, vol. 22, no. 2, pp. 89–151, 1950.

[8] N. R. French and J. C. Steinberg, "Factors Governing the Intelligibility of Speech Sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, 1947.

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

[10] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, 1980.

[11] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[12] S. Jørgensen, S. D. Ewert, and T. Dau, "A multi-resolution envelope-power based model for speech intelligibility," *J. Acoust. Soc. Am.*, vol. 134, no. 1, pp. 436–446, 2013.

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

[14] F. Chen, O. Hazrati, and P. C. Loizou, "Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure," *Biomed. Signal Process. Control*, vol. 8, no. 3, pp. 311–314, May 2013.

[15] T. H. Falk, C. Zheng, and W. Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 18, no. 7, pp. 1766–1774, 2010.

[16] C. Spille, S. D. Ewert, B. Kollmeier, and B. T. Meyer, "Predicting speech intelligibility with deep neural networks," *Comput. Speech Lang.*, vol. 48, no. October 2017, pp. 51–66, 2018.

[17] A. H. Andersen, J. M. De Haan, Z. H. Tan, and J. Jensen, "Nonintrusive Speech Intelligibility Prediction Using Convolutional Neural Networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 10, pp. 1925–1939, 2018.

[18] M. B. Pedersen, A. Heidemann Andersen, S. H. Jensen, and J. Jensen, "A Neural Network for Monaural Intrusive Speech Intelligibility Prediction," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 336–340.

[19] M. B. Pedersen, M. Kolbæk, A. H. Andersen, S. H. Jensen, and J. Jensen, "End-to-end speech intelligibility prediction using time-domain fully convolutional neural networks," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2020-Octob, pp. 1151–1155, 2020.

[20] S. Ruder, "An Overview of Multi-Task Learning in Deep Neural Networks ∗ arXiv : 1706 . 05098v1 [cs . LG] 15 Jun 2017," no. May, 2017.

[21] M. L. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 6965–6969, 2013.

[22] S. W. Fu, Y. Tsao, and X. Lu, "SNR-aware convolutional neural network modeling for speech enhancement," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 08-12-Sept, pp. 3768–3772, 2016.

[23] X. Dong and D. 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)*, 2020, pp. 911–915.

[24] K. Palanisamy, D. Singhania, and A. Yao, "Rethinking CNN Models for Audio Classification," *arXiv*, 2020.

[25] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 770–778, 2016.

[27] C. Valentini-Botinhao, "Noisy speech database for training speech enhancement algorithms and TTS models," 2017.

[28] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 08-12-Sept, pp. 352–356, 2016.

[29] J. Jensen and C. H. Taal, "An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[30] D. Sharma, Y. Wang, P. A. Naylor, and M. Brookes, "A data-driven non-intrusive measure of speech quality and intelligibility," *Speech Commun.*, vol. 80, pp. 84–94, 2016.

[31] S. S. Stevens, J. Volkmann, and E. B. Newman, "A Scale for the Measurement of the Psychological Magnitude Pitch," *J. Acoust. Soc. Am.*, vol. 8, no. 3, pp. 185–190, 1937.

[32] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 7482–7491, 2018.

[33] L. N. Smith, "Cyclical learning rates for training neural networks," *Proc. - 2017 IEEE Winter Conf. Appl. Comput. Vision, WACV 2017*, no. April, pp. 464–472, 2017.

[34] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "ESResNet: Environmental sound classification based on visual domain models," *arXiv*, 2020.

[35] G. Park and S. Lee, "Environmental noise classification using convolutional neural networks with input transform for hearing aids," *Int. J. Environ. Res. Public Health*, vol. 17, no. 7, 2020.