Speech Intelligibility Prediction for Hearing-Impaired Listeners with Phoneme Classifiers based on Deep Learning

Jana Roßbach[1,4], *Rainer Huber*[2,4], *Saskia Röttges*[3,4], *Christopher F. Hauth*[3,4], *Thomas Biberger*[3,4], *Thomas Brand*[3,4], *Bernd T. Meyer*[1,4], *Jan Rennies*[2,4]

[1]Communication Acoustics, Carl von Ossietzky University, Oldenburg, Germany
[2]Fraunhofer IDMT, Hearing, Speech and Audio Technology, Oldenburg, Germany
[3]Medizinische Physik, Carl von Ossietzky University, Oldenburg, Germany
[4]Cluster of Excellence Hearing4all, Germany

jana.rossbach@uni-oldenburg.de, rainer.huber@idmt.fraunhofer.de,
saskia.roettges@uni-oldenburg.de, christopher.hauth@uni-oldenburg.de,
thomas.biberger@uni-oldenburg.de, thomas.brand@uni-oldenburg.de,
bernd.meyer@uni-oldenburg.de, jan.rennies@idmt.fraunhofer.de

Abstract

The prediction of speech recognition is an important tool for the optimization of speech enhancement algorithms. The first Clarity Prediction Challenge was organized to find the most accurate prediction models for hearing-impaired listeners and stimuli processed by different speech enhancement algorithms. This paper describes a contribution to the challenge which is based on a listening effort prediction model. The predictions are performed non-intrusively using only the output signals from the hearing aid processors. The challenge is split into a closed data set where all listeners and enhancement algorithms are included in training and testing, and an open data set where some listeners and one algorithm are missing in the training set. For the closed data set, an individual mapping from the model output to speech intelligibility scores is used whereas for the open data set the same mapping is applied for all data points. The model achieves a prediction accuracy of 25.88 % root mean squared error (RMSE) (MBSTOI: 28.52 %) and a correlation of 0.70 for the closed set. The open set results in an RMSE of 32.07 % (MBSTOI: 36.52 %) and a correlation of 0.54. The proposed non-intrusive model outperforms the intrusive baseline model MBSTOI for both data sets.

Index Terms: speech recognition, clarity challenge, non-intrusive prediction, hearing impairment

1. Introduction

Understanding speech is often challenging in the presence of competing sound sources, especially for listeners with hearing loss. For diagnostic purposes as well as for the development and evaluation of assistive listening devices, methods for speech intelligibility assessment are required. While the most reliable way of measuring speech intelligibility is to conduct formal listening experiments, this is often too time-consuming or can only been done for a limited set of parameters or listening conditions. It would therefore be desirable to have instrumental measures which can predict speech intelligibility. However, to date there is no generally accepted model for predicting the intelligibility of speech processed by arbitrary signal enhancement strategies as measured in hearing-impaired listeners.

The first Clarity Prediction Challenge [1] was designed to benchmark different model approaches on a common data set, in which word recognition rates were measured in hearing-impaired listeners. The stimuli consisted of processed target speech recorded in different realistic acoustic scenes with spatially distributed sound sources. The signal processing had been performed by several research groups during the Clarity Enhancement Challenge [2] and, in general, comprised a combination of individual hearing loss compensation and noise reduction.

This paper describes a contribution to the prediction challenge (entry ID E022) based on the LEAP model (LEAP: Listening Effort prediction from Acoustic Parameters), introduced by Huber et al. [3] and further developed by Huber et al. [4]. LEAP is a fully blind model which derives its predictions solely based on audio signals containing speech degraded by noise, reverberation, or distortions. The model has also been successfully used for predicting the benefit of non-linear speech enhancement strategies [4]. The model was originally developed to predict ratings of perceived listening effort obtained from normal-hearing listeners, i.e., it does not comprise adaptations to introduce individual factors such as increased hearing thresholds. A further limitation of the LEAP model is that it has so far only been employed to predict speech perception in monaural or diotic listening conditions, i.e., it does not comprise a stage for predicting effects of binaural listening. On the one hand, the lack of individualization and the lack of binaural processing could limit the model's prediction accuracy for the challenge. On the other hand, the stimuli of the present challenge generally contained some sort of individualized audibility loss compensation, which may have reduced the impact of hearing loss on the collected experimental data. Similarly, informal listening to the stimuli of the challenge suggested that binaural effects likely did not play a major role, possibly because the speech enhancement schemes reduced the binaural information contained in the unprocessed stimuli (see also discussion section). We therefore decided not to extend the LEAP model by additional stages incorporating individual hearing loss or binaural hearing. Such stages would have substantially increased the model complexity, while the potential increase in prediction accuracy for the present set of conditions was unclear. Instead, a simple better-ear approach was implemented to account for possible perceptual advantages due to spatially separated sound sources in a minimalist way.

2. Method

The LEAP model (as illustrated in Fig. 1) is based on a part of

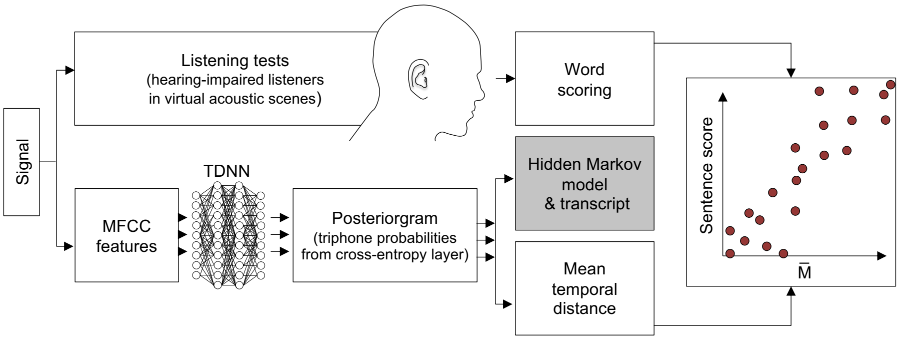

Figure 1: *Illustration of the LEAP model.*

an automatic speech recognition (ASR) system that computes triphone posterior probabilities (or ”posteriorgrams”) by means of a deep time-delay neural network (TDNN). Posteriorgrams are spanned by the dimensions time and triphones and represent the ASR system's certainty for having recognized a certain triphone at a certain point in time. Speech deterioration caused by, e.g., additive noise, reverberation, or distortions lead to an increased recognition uncertainty of the ASR system, which is reflected by a ”smearing” of the graphical representation of the posteriorgram along the time axis. The degree of smearing is quantified by a performance metric, which is the ”mean temporal distance” ($\overline{M}$), proposed by Hermansky et al. [5]. When predicting listening effort ratings, $\overline{M}$ is linearly mapped to the subjective listening effort scale. In the present application, however, a sigmoidal mapping to the speech intelligibility scores is employed. The mapping function was derived empirically using the training data set of the Clarity Prediction Challenge. The database, the structure of the TDNN, its training, the performance metric, and the baseline model will be described in more detail in the following.

2.1. Database

The challenge database contains multiple audio signals [6] and different types of listener data [2]. Not all data are needed for the non-intrusive model of this study, so only the data used are described here. To create a spatial acoustic scene, unique target utterances and unique noise segments are convolved with binaural room impulse responses (BRIRs) and mixed together. These audio signals and the audiograms of 27 hearing-impaired listeners were used as input to the hearing aid processors of the first enhancement challenge [2]. The outputs of the hearing aid processors were used for listening experiments with the corresponding hearing-impaired listeners. The challenge database is split into an open and a closed data set, and each data set is again split into training and testing. In the closed data set, all 27 listeners and ten enhancement algorithms are included in both the training and test sets. The training data of the open data set contains only the audiograms of 22 listeners and the outcomes of nine enhancement algorithms. However, the missing listeners and algorithms are included in the test data of the open data set. The measured intelligibility scores of all training signals from both data sets were provided for the challenge participants to optimize the models. The intelligibility scores of both test sets were provided after the submission of the predicted scores.

2.2. Posteriorgram generation

The stimuli provided in the challenge were preprocessed by removing the first 2 seconds and the last 1 s, which were known to contain noise only. 40-dimensional log-Mel filterbank energies were calculated from the trimmed stimuli and used as acoustic feature input to the TDNN. The length of each feature frame was 10 ms. Apart from the current feature frame, a context of +/-15 feature frames was used as input to the TDNN. The input layer was followed by seven hidden layers with 700 rectified linear units (ReLU) each. The dimensionality of the output layer was 6448, i.e., one neuron per triphone. The ASR was trained with about 1000 hours of clean German speech of an in-house training data set, expanded to about 8000 hours by mixing the speech with different kinds of noises and also convolving it with different room impulse responses. The network was trained with the lattice-free maximum mutual information (LF-MMI) criterion [7]. As pointed out in [4], the TDNN used here had two output layers during training, one that followed the LF-MMI objective function and one that followed a cross-entropy (CE) objective function. The latter one is usually used to regularize training only, while the former one is used for ASR purposes. However, here the CE output layer was used for generating posteriorgrams, due to better results in terms of higher correlations between subjective listening effort ratings and corresponding predictions by the LEAP model regarding earlier experiments.

2.3. Performance metric

The measure $\overline{M}$ computes the average difference between two vectors of triphone posteriors $p_{t-\Delta t}$ and p_t (i.e., two columns of the posteriorgram) with a temporal distance Δt:

$$M(\Delta t) = \frac{1}{T - \Delta t} \sum_{t=\Delta t}^{T} D(p_{t-\Delta t}, p_t).$$

T is the temporal length of the analyzed posteriorgram (which is equal to the length of the analyzed speech file), and D is the symmetric Kullback-Leibler divergence between two vectors x and y with components $x(i)$ and $y(i)$:

$$D(x, y) = \sum_{i=1}^{N} x(i) log(\frac{x(i)}{y(i)}) + \sum_{i=1}^{N} y(i) log(\frac{y(i)}{x(i)}).$$

N equals the dimensionality of the TDNN output layer (6448) and $M(\Delta t)$ is computed for Δt = 350 to 800 ms (in 50 ms steps) and averaged to the final speech recognition predictor $\overline{M}$.

The predictions are done for each ear independently, and the higher score is used as final prediction.

2.4. Mapping from $\overline{M}$ to speech recognition

The measure $\overline{M}$ is an entropy-based scalar and needs to be mapped to a perceptual scale according to the experiment at hand based on a reference condition. In this challenge, the mapping was derived to predict the speech recognition in percent correct by using

$$f(x) = \frac{1}{1 + exp(4 * s_{50} * (L_{50} - x))} \tag{1}$$

where L_{50} corresponds to the speech recognition threshold (SRT) at which 50 % of the words are understood correctly [8]. The slope at this point is denoted with s_{50}. The psychometric function was fitted to the training data by minimizing the least squared error. The parameters (L_{50} and s_{50}) that resulted in the best fitting curve were then used during testing to map the $\overline{M}$ values to intelligibility scores.

The parameters that fitted best to all points of the open training data were used to map the $\overline{M}$ values of the open test set. The mapping for the closed data set was a bit different, because it was done individually for each of the listeners. The training data was divided into 27 data sets, one set for each listener. For each of the listener data sets, the optimal mapping parameters were calculated and stored with the corresponding listener ID. The $\overline{M}$ values of the closed test set were mapped by using the individual parameters of each listener.

2.5. Baseline model

A baseline model is necessary to evaluate the model performance. The challenge organizers decided to use the modified binaural short-time objective intelligibility (MBSTOI) [9] for this purpose. The MBSTOI uses the cross-correlation between the envelope of a clean speech signal and a degraded signal [10]. The binaural processing is performed with an equalization and cancellation (EC) stage before the envelope extraction. Since the MBSTOI is based on correlation, the model output is a value between 0 and 1 and needs to be mapped to intelligibility scores. This is done using a sigmoidal fit that minimizes the root mean squared error (RMSE) of the training data set.

The MBSTOI does not contain its own hearing loss simulation. For this reason, the signals from the enhancement algorithms must be sent through an hearing loss model before they can be used as input to the MBSTOI. The challenge organizers used a hearing loss model [11] that simulates reduced audibility, dynamic range, time resolution, and frequency resolution.

3. Results and discussion

To determine the prediction accuracy of the proposed model, the RMSEs and the correlations between the measured and predicted intelligibility scores were calculated. Table 1 shows the results of LEAP and the baseline model MBSTOI. For both the closed and open data set, LEAP achieved a lower RMSE and a higher correlation than MBSTOI. In the closed data set, the RMSE of LEAP was 2.6 % lower than the RMSE of MBSTOI. This difference was larger in the open data set (4.5 %). For both models, the RMSE was lower and the correlation was higher for the closed set than for the open set.

The better predictions of LEAP compared to MBSTOI are notable because MBSTOI is an intrusive model receiving clean speech as input in addition to the output of the hearing aid processors, and because of the fact that binaural processing was explicitly modeled by MBSTOI using an EC mechanism, while a simple better-ear approach was used for LEAP in this study.

Table 1: *Root mean squared errors (RMSEs) and correlations of LEAP and MBSTOI for the closed and open data set.*

data set	model	**RMSE**	**std err**	ρ
closed	LEAP	25.88 %	0.53 %	0.70
	MBSTOI	28.52 %	0.58 %	0.62
open	LEAP	32.07 %	1.21 %	0.54
	MBSTOI	36.52 %	1.35 %	0.53

A possible explanation for the differences between the two prediction models for the closed data set could be that our model approach used individual mapping functions of model output values to intelligibility scores. Instead of integrating a hearing loss model into LEAP, individual listener-dependent mapping parameters were used for the closed data set, which can be interpreted as indirect integration of hearing loss. Effectively, this corresponds to taking into account the general trend of an individual listener to have "relatively better" or "relatively worse" performance in the tested group of listeners. In contrast, in MBSTOI the hearing losses are integrated in the processing steps and the same mapping is applied to all listeners. The advantage of the individual mapping is that not only the audiogram but also other factors such as supra-threshold deficits or cognitive aspects which affect speech recognition are taken into account. During the training period of the challenge, we experimented with different ways of individualizing the predictions for the aided hearing-impaired listeners, including the hearing loss simulation provided by the challenge organizers. We did not, however, find significant improvements of individualized predictions in comparison to using the generic model framework. Hence, the only individualization we included was to derive individual mapping functions for the closed data set based on the available training data. The good prediction scores of the model compared to the baseline appear to support the use of individualized mapping functions when possible. This is in line with a recent study employing a different model framework and different acoustic conditions [12], which also reported increased prediction accuracy when using individual mappings compared to integrating individual threshold-simulating noise into the model's processing stages.

The disadvantage of this approach is that this can only be done for listeners included in training data. Some listeners were not included in the training data of the open set, so for consistency the identical mapping was applied to all listeners of this data set and, thus, no individual hearing loss integration was performed. However, even without individual mapping, LEAP performed better than MBSTOI, and thus mapping alone cannot explain why LEAP provided better predictions than MBSTOI for the open data set.

A further reason why the auxiliary information used by MBSTOI did not produce more accurate predictions could be the limited role of binaural information. In principle, interaural level and time differences could have significantly impacted the perceptual data of the challenge since the acoustic scenes comprised spatially separated sound sources. However, during the training phase of the challenge, we experimented with different approaches for taking binaural effects into account. In general, we observed that the contribution of binaural effects

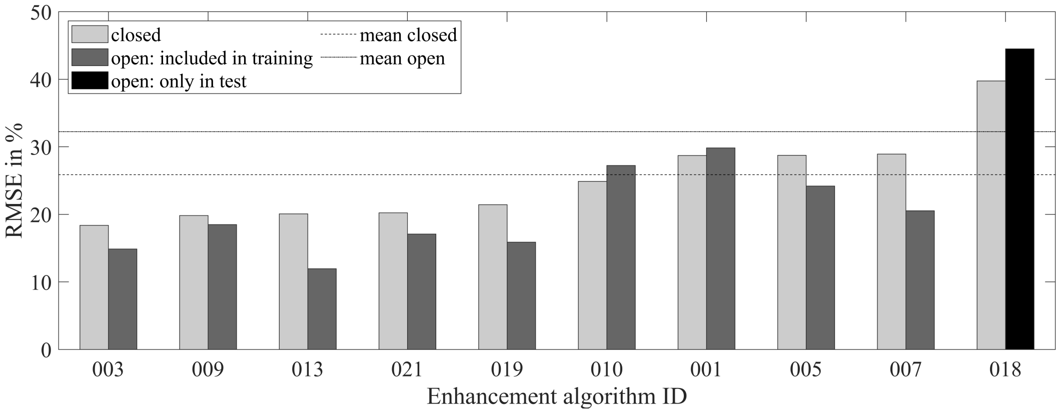

Figure 2: *Root mean squared errors for each enhancement algorithm. The algorithms are sorted by ascending RMSE of the closed data set*

to the predicted intelligibility scores could be largely captured by employing a simple better-ear approach, i.e., by computing the predictions independently for each ear and then taking the higher score as prediction score. Specifically, including a blind EC processing stage (as proposed in [13]) did not improve prediction accuracy for the provided training data. In other words, the contribution of binaural cues beyond head-shadow effects appears to have played only a minor role for the current test set. It is possible that this is due to the algorithmic modifications, which may have reduced the binaural cues and focused on SNR enhancement. Another potential reason for the limited contribution of binaural cues could be that the hearing-impaired listeners were not able to exploit such cues due to their hearing loss.

The difference between the open and closed data sets was not only that some listeners were missing in the training, but also that one of the enhancement algorithms was missing. To further investigate the influence of the enhancement algorithms on the prediction accuracy, the RMSEs for each algorithm are shown in Figure 2. For the closed data set, the RMSEs are sorted in ascending order for better readability. One RMSE of the open data set is colored black to highlight the algorithm E018 that was not included in the training. E018 is the algorithm that has the highest RMSE. The figure also illustrates that predictions for each of the other algorithms were more similar in RMSE for the closed and open data sets. This result suggests that the RMSEs depend more on the enhancement algorithms than on the listeners. With an individual mapping per enhancement algorithm a higher prediction accuracy could probably be achieved, but this was not allowed in this challenge.

4. Conclusion

The aim of the first Clarity Prediction Challenge is find the model that most accurately predicts the speech recognition for a set of experimental data that was collected for a group of hearing-impaired listeners and stimuli processed by different types of speech enhancement algorithms. For the closed data set, the proposed non-intrusive LEAP model was extended by an individual mapping from model output to speech intelligibility scores, while for the open data set a general, listener-independent mapping was used. The LEAP model outperforms the intrusive baseline model MBSTOI for both data sets, i.e., better prediction results are achieved despite the fact that the model does not require target speech signals to be available separately.

5. Acknowledgement

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2177/1 – Project ID 390895286, Project ID 352015383 – SFB 1330 A1/A2 and Project ID 325439187 – Multilinguale modellbasierte rehabilitative Audiologie.

6. References

[1] J. Barker, M. Akeroyd, T. J. Cox, J. F. Culling, J. Firth, S. Graetzer, H. Griffiths, L. Harris, G. Naylor, Z. Podwinska, E. Porter, and R. Viveros Munoz, "The 1st Clarity Prediction Challenge: A machine learning challenge for hearing aid intelligibility prediction." submitted to Interspeech, 2022.

[2] S. Graetzer, J. Barker, T. J. Cox, M. Akeroyd, J. F. Culling, G. Naylor, E. Porter, and R. V. Muñoz, "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2, pp. 1181–1185, 2021.

[3] R. Huber, C. Spille, and B. T. Meyer, "Single-ended prediction of listening effort based on automatic speech recognition," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2017-Augus, pp. 1168–1172, 2017.

[4] R. Huber, A. Pusch, N. Moritz, J. Rennies, H. Schepker, and B. T. Meyer, "Objective assessment of a speech enhancement scheme with an automatic speech recognition-based system," *Speech Communication; 13th ITG-Symposium, Oldenburg, Germany*, p. 86–90, 2018.

[5] H. Hermansky, E. Variani, and V. Peddinti, "Mean temporal distance: Predicting ASR error from temporal properties of speech signal," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 7423–7426, 2013.

[6] I. Demirsahin, O. Kjartansson, A. Gutkin, and C. Rivera, "Open-source Multi-speaker Corpora of the English Accents in the British Isles," in *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*. Marseille, France: European Language Resources Association

(ELRA), May 2020, pp. 6532–6541. [Online]. Available: https://www.aclweb.org/anthology/2020.lrec-1.804

[7] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2751–2755, 2016.

[8] T. Brand and B. Kollmeier, "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," *The Journal of the Acoustical Society of America*, vol. 111, no. 6, pp. 2801–2810, 2002.

[9] A. H. Andersen, J. M. de Haan, Z. H. Tan, and J. Jensen, "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," *Speech Communication*, vol. 102, pp. 1–13, 2018. [Online]. Available: https://doi.org/10.1016/j.specom.2018.06.001

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

[11] Y. Nejime and B. C. Moore, "Simulation of the effect of threshold elevation and loudness recruitment combined with reduced frequency selectivity on the intelligibility of speech in noise," *The Journal of the Acoustical Society of America*, vol. 102, no. 1, p. 603–615, 1997.

[12] A. M. Kubiak, J. Rennies, S. D. Ewert, and B. Kollmeier, "Prediction of individual speech recognition performance in complex listening conditions," *The Journal of the Acoustical Society of America*, vol. 147, no. 3, pp. 1379–1391, 2020.

[13] C. F. Hauth, S. C. Berning, B. Kollmeier, and T. Brand, "Modeling Binaural Unmasking of Speech Using a Blind Binaural Processing Stage," *Trends in Hearing*, vol. 24, 2020.