Speech Intelligibility Prediction for Hearing-Impaired Listeners with the LEAP Model - Technical Report Contribution E022

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, bernd.meyer@uni-oldenburg.de,
thomas.brand@uni-oldenburg.de, jan.rennies@idmt.fraunhofer.de

1. Introduction

This contribution (E022) to the first Clarity Prediction Challenge [1] is based on the LEAP model (LEAP: Listening Effort prediction from Acoustic Parameters), introduced by Huber et al. [2] and further developed by Huber et al. [3]. It 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 [3]. 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. During the training period of the challenge, we experimented with different ways of individualizing the predictions for the aided hearing-impaired listeners, including the provided hearing loss simulation. 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. 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. For the open data set, no adaptation of the model was made except for a modified mapping function to predict speech intelligibility scores instead of listening effort ratings.

So far, the LEAP model has only been employed to predict speech perception in monaural or diotic listening conditions. Since the current challenge comprised acoustic scenes with spatially separated sound sources, we experimented with different approaches for taking binaural effects into account (also in conjunction with the blind binaural speech intelligibility model, bBSIM, described in contribution E019 [4]). We observed that the contribution of binaural effects 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. In other words, the contribution of binaural cues beyond head-shadow effects appears to play a minor role only 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. We therefore decided not to incorporate a binaural preprocessing stage in this contribution and employ a better-ear approach instead.

2. Method

The LEAP model is based on a part of 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 structure of the TDNN, its training, and the performance metric will be described in more detail in the following.

2.1. Posteriorgram generation

40-dimensional log-Mel filterbank energies were used as acoustic features 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 were 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 kind 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 [6]. As pointed out in [3], 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 is 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.

2.2. 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 listening effort predictor $\overline{M}$.

2.3. 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 is derived to predict the speech recognition in percent correct by using

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

where L_{50} corresponds to the speech recognition threshold (SRT) at which 50 % of the words are understood correctly [7]. The slope at this point is denoted with s_{50}. The psychometric function is fitted to the training data by minimizing the least squared error. The parameters (L_{50} and s_{50}) that result in the best fitting curve are then used during testing to map the $\overline{M}$ values to intelligibility scores.

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

2.4. Application of the model in the challenge

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. The trimmed stimuli were then used as input to the model. The predictions were made for each ear independently, and the higher score was used as final prediction.

3. 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 modelbasierte rehabilitative Audiologie.

4. References

[1] 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.

[2] 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.

[3] 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.

[4] S. Röttges, J. Roßbach, C. F. Hauth, T. Biberger, R. Huber, J. Rennies, and T. Brand, "Speech Intelligibility Prediction using the bBSIM-STI Model - Technical Report Contribution E019," *Technical report of Clarity Prediction Challenge*, pp. –, 2022.

[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] 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.

[7] 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.