Speech Intelligibility Prediction using the bBSIM-STI Model - Technical Report Contribution E019

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

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

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

1. Introduction

This contribution (E019) to the first Clarity Prediction Challenge (CPC1) [1] is based on the latest version of the blind Binaural Speech Intelligibility Model (BSIM20) [2] and the correlation-based version of the Speech Transmission Index (STI) [3]. Former versions of BSIM [4] did not work blindly (i.e., they required separated speech and noise signals) and applied the Speech Intelligibility Index (SII) [5] as back-end. In this contribution we use the blind front-end of the BSIM20 which is called bBSIM in the following. bBSIM produced equal results as the non-blind version but requires less auxiliary information about the target speech and the masking noise, so that it can be combined with arbitrary back-ends predicting speech recognition scores (see, e.g., [6, 7]).

The use of bBSIM helps to understand, how relevant the binaural information in the CPC1 is for speech understanding. In this contribution, we use the correlation based STI as back-end, as it is takes reverberation effects into account and produced the best predictions for the test set of CPC1 compared to other back-ends we tried. This back-end is not blind as it requires target speech and interfering noise separately and thus the combination of bBSIM and STI is a hybrid model. Note, that in this contribution no machine learning is applied but two classic approaches from psychoacoustics are combined that are very easy to compute. In this respect this contribution is very close to the baseline model of the Clarity challenge which used a very similar binaural front-end [8] combined with a back-end that also analyses the modulations of the signal [9]. In this respect this contribution can be seen as an alternative baseline model that shows how far we (the authors) were able to get without machine learning and training to the test data.

2. Method

2.1. bBSIM

The bBSIM[2] receives the mixed target speech and interferer signals at the left and the right ear as input. The stimuli provided in the challenge were preprocessed by removing the first 2 seconds and the last 1 second that were known to contain only noise. After this, noisy frames of the signal were still detected. We decided to additionally apply an rms based voice activity detection to remove silent frames. To simulate the frequency selectivity of the human auditory system, the input signals are split into 30 Equivalent Rectangular Bandwidth-(ERB-)[10] spaced frequency bands by using a gammatone filterbank [11] ranging from 150 Hz to 8000 Hz. Based on the individual pure tone audiograms, two internal threshold simulating noises were added to the left and right input signals to simulate the hearing loss. The left and right threshold simulating noises were generated as uncorrelated signals, so that the EC stage of bBSIM cannot cancel them out. For frequencies up to 1500 Hz, binaural processing is realized as blind equalization-cancellation (EC) [2] mechanism, where the differences in interaural time differences (ITDs) and interaural level differences (ILDs) between target and interfering signal can be used to improve the signal-to-noise ratio. For frequencies above 1500 Hz, the better ear is selected blindly. In the equalization step the two ear channels of each gammatone filter channel are equalized in level and phase. Then, the cancellation step is applied, which uses two different strategies: 1) a minimization of the output power and 2) a maximization of the output power. While the first strategy can be assumed to be the better strategy at negative SNRs, because it attenuates the interfering signal, the second strategy can be assumed to be better at positive SNRs, because the power of the target signal is increased. To choose the best of both strategies in each frequency channel, the speech-to-reverberation modulation energy ratio (SRMR) [12] is used. SRMR describes the ratio between speech-like and non speech-like modulations by calculating a ratio between the energy in modulation frequency channels below 16 Hz and above 16 Hz. The SRMR is calculated for both strategies and both ear channels and, subsequently, the EC channel and the ear channel with the higher SRMR are combined to produce a single channel signal with enhanced SNR. Due to its simple calculation SRMR can be applied independently to each ERB channel.

2.2. Speech Transmission Index

The speech transmission index (STI) [13] receives bBSIM's output signals of the clean target speech and degraded speech as input. The calculation of the separate target and interfering signals is possible as bBSIMs processing is linear with respect to the signals, so that speech and noise can be processed separately using the EC parameters determined by the blind model (see [2] for details). The STI analyzes the modulation transfer function by comparing the envelopes of the input signals to calculate the modulation transmission index for each frequency band. Here, the normalized covariance method [3] was applied: The covariance between the envelopes of the target speech and the degraded speech were calculated and then normalized with the individual variances of the target speech and the degraded speech. The weighted average of the transfer index of all frequency bands gives the STI and is very similar to the later proposed short-time objective intelligibility (STOI) measure [8].

2.3. Mapping from STI to speech recognition

The STI is an index value ranging from 0 to 1 and needs to be mapped to a perceptual scale according to the experiment based on a reference condition. In this challenge, the mapping is derived to predict the speech recognition in percent correct by using

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

where L_{50} corresponds to the speech recognition threshold (SRT) at which 50 % of the words are understood correctly [14]. 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 fit best to all points of the open training data have been used to map the STI index values of the open test set. The mapping for the closed data set has been done individually for each listener: The training data is divided into 27 data sets, one set for each listener. For each listener, the optimal mapping parameters are calculated and stored with the corresponding listener ID. The STI index values of the closed test set are mapped by using the individual parameters of each listener.

3. Discussion

We observed that the model's binaural processing did not generate relevant spatial or binaural unmasking. This indicates that the signals do not provide usable binaural information. To evaluate this, the model has to be applied to the unprocessed signals. A further reason for the missing unmasking might be that the applied signal enhancement algorithms have destroyed binaural information.

Furthermore we observed that the listener's individual hearing loss as expressed by the pure tone audiogram was not important for the accuracy of the model predictions. This finding might mirror the fact that the listener's adjusted the overall level themselves and that consequently audibility did not play an important role in these measurements and that suprathreshold hearing deficits are not well described by the pure tone audiogram. For that reason we did not use the pure tone audiogram at all in our second submission (E022).

For the interpretation of the results of this challenge it has to be taken into account that the human recognition data is binomially distributed and that consequently the standard error of each measured recognition score is given by

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}}, \tag{2}$$

with p denoting the recognition score of the sentence (with values from 0 to 1) and n denoting the number of words tested in this sentence. If, for example, a sentence with six words is tested and three of them have been repeated correctly by the listener, the standard error of the p estimate equals approximately 20%. In other words, even a perfect model that predicts p exactly will achieve an average standard error not better than 20%. Considering this helps to interpret the results of this challenge. We recommend to predict average recognition scores in the next round of this challenge so that differences between the participating prediction models are not blurred due to the statistics of the ground truth data.

4. Acknowledgement

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

5. References

[1] S. Graetzer, J. Barker, T. J. Cox, M. Akeroyd, J. F. Culling, G. Naylor, E. Porter, and R. Viveros Muñoz, "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing," *in Proceeding of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2021, Brno, Czech Republic, 2021.*

[2] C. F. Hauth, S. C. Berning, B. Kollmeier, and T. Brand, "Modelling binaural unmasking of speech using a blind binaural processing stage," *Trends in Hearing*, vol. 24, 2020.

[3] I. Holube and B. Kollmeier, "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," *The Journal of the Acoustical Society of America*, vol. 100, no. 3, pp. 1703–1716, 1996.

[4] R. Beutelmann, T. Brand, and B. Kollmeier, "Revision, extension, and evaluation of a binaural speech intelligibility model," *The Journal of the Acoustical Society of America*, vol. 127, no. 4, pp. 2479–2497, 2010.

[5] ANSI, "ANSI S3.5-1997, American national standard methods for calculation of the speech intelligibility index," *Am. Natl. Stand. Institute, New York*, 1997.

[6] D. Hülsmeier, C. F. Hauth, S. Röttges, K. P., J. Roßbach, M. R. Schädler, B. T. Meyer, A. Warzybok, and T. Brand, "Towards Non-Intrusive Prediction of Speech Recognition Thresholds in Binaural Conditions, in Proc. Conference on Speech Community (ITG)," 2021.

[7] S. Röttges, C. Hauth, J. Rennies, and T. Brand, "Using a blind EC mechanism for modelling the interaction between binaural and temporal speech processing," *Acta Acustica united with Acustica*, Accepted by Acta Acustica united with Acustica 2022.

[8] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time – Frequency Weighted Noisy Speech," *IEEE Transaction on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[9] A. Andersen, J. M. de Haan, Z. Tan, and J. J., "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] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *Journal of the Acoustical Society of America*, vol. 74, no. 3, pp. 750–753, 1983.

[11] V. Hohmann, "Frequency analysis and synthesis using a Gammatone filterbank," *Acta Acustica united with Acustica*, vol. 88, no. 3, p. 433–442, 2002.

[12] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech," *2014 14th International Workshop on Acoustic Signal Enhancement, IWAENC 2014*, pp. 55–59, 2014.

[13] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *The Journal of the Acoustical Society of America*, vol. 318, no. 1980, 1979.

[14] 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.