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

Saskia Röttges\textsuperscript{1,4}, Jana Roßbach\textsuperscript{2,4}, Christopher F. Hauth\textsuperscript{1,4}, Thomas Biberger\textsuperscript{1,4}, Bernd T. Meyer\textsuperscript{2,4}, Rainer Huber\textsuperscript{3,4}, Jan Rennies\textsuperscript{3,4}, Thomas Brand\textsuperscript{3,4}

\textsuperscript{1}Medizinische Physik, Carl von Ossietzky University, Oldenburg, Germany
\textsuperscript{2}Communication Acoustics, Carl von Ossietzky University, Oldenburg, Germany
\textsuperscript{3}Fraunhofer IDMT, Hearing, Speech and Audio Technology, Oldenburg, Germany
\textsuperscript{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) \cite{1} is based on the latest version of the blind Binaural Speech Intelligibility Model (BSIM20) \cite{2} and the correlation-based version of the Speech Transmission Index (STI) \cite{3}. Former versions of BSIM \cite{4} did not work blindly (i.e., they required separated speech and noise signals) and applied the Speech Intelligibility Index (SII) \cite{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., \cite{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 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 psychacoustics 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 \cite{8} combined with a back-end that also analyses the modulations of the signal \cite{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\cite{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)-bands. The speech transmission index (STI) \cite{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 \cite{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

\[ f(x) = \frac{1}{1 + \exp(4 \cdot s_{50} \cdot (L_{50} - x))} \]

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

\[ \sigma_p = \sqrt{\frac{p(1-p)}{n}} \]

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


