Review of the Binaural Speech Intelligibility Model (BSIM)

Thomas Brand[1], Christopher F. Hauth[1], Saskia Röttges[1], Jan Rennies[2]
[1]*Medical Physics, Universität Oldenburg and Cluster of Excellence Hearing4all, 26129 Oldenburg, Germany, E-Mail: thomas.brand@uni-oldenburg.de*
[2]*Fraunhofer IDMT, Hearing, Speech and Audio Technology and Cluster of Excellence Hearing4all, Oldenburg, Germany*

Introduction

The Binaural Speech Intelligibility Model (BSIM) predicts speech intelligibility for spatially separated target speech and interferers. Several model revisions and extensions have been introduced adapting BSIM to modulated noises, reverberant rooms, and temporally varying binaural conditions. Furthermore a blind version of BSIM's binaural preprocessing was introduced, which requires only a mixture of target speech and interferers and does not require them as clean signals. This blind front-end can be combined with arbitrary back-ends predicting speech intelligibility.

If BSIM is used for planning and control of measurement conditions in lab studies, auxiliary information about the measuring conditions is available and non-blind versions of BSIM can be used. The same holds for evaluating the hearing abilities of individual listeners with hearing loss in controlled test conditions. On the other hand, if BSIM should be used, for instance, for controlling the processing of hearing instruments in unknown and variable conditions the model has to perform blindly and in real time.

This review discusses different BSIM settings with respect to the required auxiliary information and to real-time applicability.

BSIM front-ends

All versions of BSIM consist of a front-end and a back-end. In the back-end the selection of the better ear and the binaural processing is modelled and a prediction of the spatial and binaural release from masking is made. The back-end uses the front-end's output for calculating the final intelligibility estimate or Speech Reception Threshold (SRT, i.e. the signal-to-noise ration where 50% of the words are recognized correctly).

All front-end versions of BSIM base on the Equalization-Cancellation (EC) model [1] which assumes that the listener's internal signal-to-noise ratio can be improved by equalizing the Interaural Level Difference (ITD) and the Interaural-Time-Difference (ITD) and subsequently subtracting the left and right ear signal which cancels parts of the noise signal. This processing takes place almost independently within in each frequency channel [2]. The amount of the improvement depends on the listening condition and of internal processing errors (jitters) as proposed by [3] that fit the model's accuracy to human performance.

Non-blind front-ends

In the first version of BSIM [4] these processing errors have been implemented using Monte-Carlo Simulations in which the level and time jitters in the equalization stage have been realized by detuning (jittering) the optimal level and time equalization factors. The jitters are given by random variables drawn repetitively from normal distributions with standard deviations given by [3] and averaging the EC output. An advantage of this method is that a signal is calculated that can directly be used arbitrary back-ends as well as for listening tests (see below). The disadvantage is that this method is computationally slow (see Figure 3). This problem was solved in a revision of BSIM [5] (called BSIM2010 in the following) which does not require Monte-Caro simulations, but which calculates the SNR for each frequency channel analytically in a very fast way. Furthermore in this revision a short-term processing (stBSIM) has been introduced which allows to apply BSIM to modulated interferers.

Blind front-end

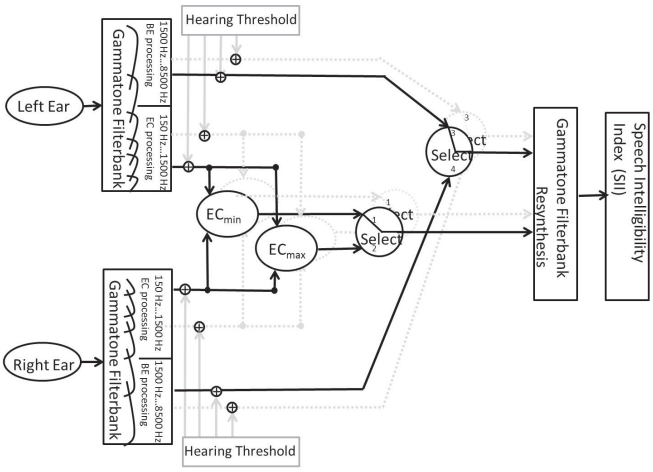

Figure 1: Hybrid version of BSIM using the blind binaural front-end (bBSIM) and the non-blind SII as back-end. From [6].

[6] presented a blind front-end of BSIM (called bBSIM in the following) which does not require any auxiliary information about the clean target or interferer signals, but requires only the mixed signals. The equalization parameters are now controlled by the Speech-to-Reverberant-Modulation-energy Ratio (SRMR) [7], which is a blind measure of speech likeness and which acts here as a preliminary back-end. The SRMR analyzes the ratio between modulation energy below 16 Hz and above 16 Hz and can easily be calculated independently for each frequency band. For frequencies above 1500 Hz the SRMR is used to select the better ear and below 1500 Hz the SRMR is used to select between to alternative EC processing strategies. One strategy is the minimization of the EC output, which is the optimal strategy for negative SNRs, when the noise has to be suppressed by destructive interference of the two ear signals. The other strategy is the maximization of the EC output, which is the optimal strategy for negative SNRs, when the speech has to be enhanced by constructive interference of

the two ear signals. Figure 1 shows a sketch of bBSIM using the SII as back-end.

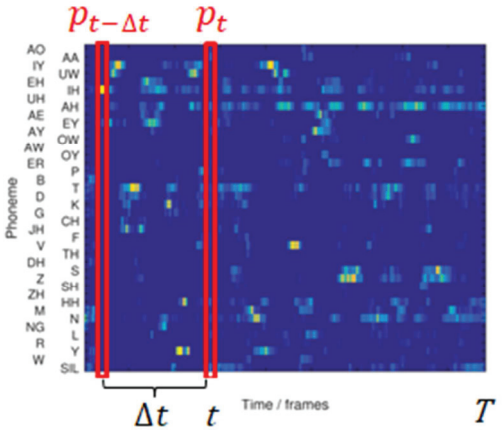

Figure 2: Example of phoneme recognizer output used for calculating the Mean-Temporal-Distance in [19].

Non-blind BSIM back-ends

Some of the back-ends used in for BSIM use auxiliary information about the clean target speech signal and the clean interferer. This information is not available to the real listener and therefore these back-ends are called "intrusive" or "non-blind".

Speech-intelligibility Index (SII)

The Speech-intelligibility-Index (SII) [8] which has been used in most BSIM studies so far [e.g. 2, 4, 5, 9, 10, 11] analyzes the signal-to-noise ratio from -15 to 15 dB in different frequency band in order to calculate an index that can then be transformed to a recognition score or a SRT value. It is very convenient to combine the analytic front-end of BSIM [5] with the SII as this front-end directly predicts the frequency dependent SNR as needed by the SII. In general the SII is not well suited for predicting intelligibility in reverberant conditions as long as also the reverberant speech is assumed to be useful. However, if the SII is used together with a useful-to-detrimental separation that analyses the Binaural-Room-Impulse-Respond (BRIR) the model yielded its best prediction accuracy [9]. Note, that in this mode BSIM not only requires auxiliary information about the clean speech and noise signals but also about the BRIR. Other non-blind back-ends that have been evaluated successfully using BSIM are the Speech-Transmission-Index [12] in the version proposed in [13] which is very similar to the Short-Term-Objective-Intelligibility (STOI) [14].

Blind BSIM back-ends

NI-STOI

[15] proposed a non-intrusive (that means blind) version of STOI, which does not require separate target speech and interferers, but only receives a mixture of them. This non-intrusive STOI (NI-STOI) predicts intelligibility based on the correlation between the envelopes of clean and degraded target speech calculated in 1/3-octave frequency bands, which is very similar to [13]. However, for estimating the clean speech envelopes from the degraded speech envelopes, NI-STOI uses a statistical model, which requires training with clean speech.

FADE

Another blind front-end evaluated in combination with bBSIM was the Framework for Auditory Discrimination Experiments (FADE) [16] which requires speech and noise signals mixed at different SNRs to predict SRTs. FADE applies separable Gabor filterbank features (SGBFB) [17] which are extracted from logarithmically scaled mel spectrograms (LogMS) of the signals. These features are used for training and testing Gaussian Mixture Model-HMM (GMM-HMM) ASR-systems at different SNRs. In order to yield optimal predictions FADE's ASR systems are trained for each acoustic environment and spatial configuration separately. After this training FADE's recognition rates are determined by comparing the recognized words with the presented words. In other words FADE acts like a human listener in an intelligibility test.

Mean Temporal Distance (MTD)

Another ASR based back-end motivated by [18] was evaluated together with bBSIM in [19]. Here the output of the bBSIM front-end is used as input for the ASR-based back-end that is combined with an entropy measure applied to phoneme recognition probabilities. A mel-spectrum with 40 frequency channels is calculated from which a DNN calculates phoneme probabilities. The DNN is trained by minimizing the word error rates based on the phoneme probabilities which are mapped to word transcriptions using a Hidden Markov model (HMM). Note, that this HMM is only required for the training. In the final test, recognition rates are estimated directly based on the DNN output by calculating the mean temporal distance (MTD) from the posteriorgrams [20]. Consequently no transcription of the speech is needed. Instead, like for the other index based back-ends (SII, STI, NI-STOI) mentioned before, the MTD is mapped to the SRT using a reference MTD value, i.e. the MTD at the SRT for a reference condition (anechoic room and collocated speech and noise).

Tabelle 1: Comparison between predicted and measured SRTs of [4] given as squared correlation coefficient R^2 and root mean squared error RMSE. The different models are: BSIM2010 (non-blind reference model) [5], BAPSI (Binaural ASR-based Prediction of Speech Intelligibility, bBSIM and phoneme recognizer with MTD) [19], niSTOI (bBSIM and non-intrusive STOI), F-KAIN (FADE based on monaural and binaural features), and F-bBSIM (bBSIM front-end with FADE back-end). From [21].

Model	R^2	RMSE/dB
BSIM06	0.94	1.1
BAPSI	0.89	1.3
NI-STOI	0.65	2.2
F-KAIN	0.93	1.1
F-bBSIM	0.92	1.6

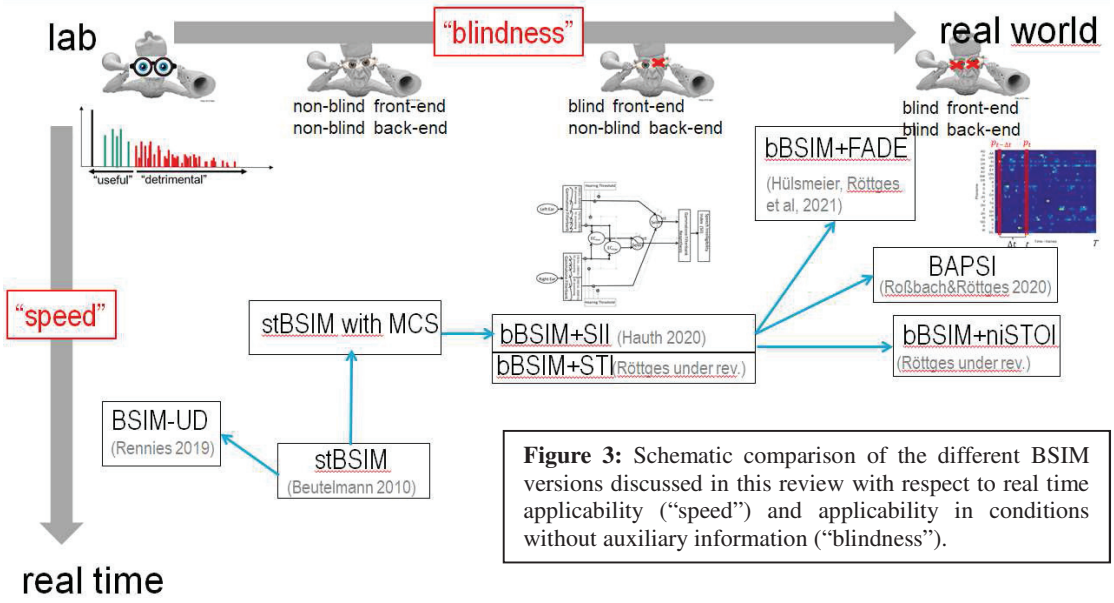

Figure 3: Schematic comparison of the different BSIM versions discussed in this review with respect to real time applicability ("speed") and applicability in conditions without auxiliary information ("blindness").

Discussion

The majority of the studies using the BSIM approach used the quite simple SII as back-end which analyses the SNR. This worked astonishing well even in conditions with reverberation and echoes as long as the front-end applies a useful/detrimental separation [9, 10]. This approach is non-blind not only with respect to the required auxiliary knowledge about the clean speech and noise signals but also with respect to auxiliary knowledge about the binaural room impulse response (See upper left inlay of Figure 3, the model is non-blind and has additional glasses.). In order to apply BSIM also to situations where this auxiliary information is not available the SII back-end has to be replaced by a blind back-end. This requires so far the use of Monte-Carlo simulations (indicated as "stBSIM MCS" in Figure 3) making the model slower.

Different back-ends that analyze the front-ends output either with respect to the modulation content of the signal [7, 12, 13, 14, 15] or that apply ASR methods have been evaluated and were found to work quite well for many conditions.

When comparing different back-ends that are referred to as being "blind" or "non-intrusive", it has to be carefully considered that different authors use these expressions in different manners:

NI-STOI analyses the speech-likeness of the signal modulations. This makes NI-STOI universally applicable to arbitrary unknown speech. This comes at the cost of sometimes imprecise predictions when the signal contains speech-like modulations without being intelligible speech [15]. Nevertheless, NI-STOI is a very promising candidate for real-time applications, because it can be calculated very fast and without any auxiliary information. That means it can be applied to arbitrary unknown signals.

FADE does not require a reference index, because this value is found by FADE itself during training. However, it has to be taken into account that this requires a specific training to each tested condition and that this training requires that speech and noise are mixed at different SNRs. In other words FADE requires auxiliary information during training. Furthermore, it has to be taken into account that FADE requires a transcript of the speech during testing, as the words recognized by FADE have to be compared to the presented words in order to calculate the intelligibility. That means FADE is blind in the way that it can be applied to mixed signals during testing but in its current form FADE cannot be applied to arbitrary unknown speech.

The MTD back-end uses an ASR approach but interestingly it does not require a transcript of the test speech, because no recognition scores are calculated but the probability that phonemes are presented. In this sense the MTD distance is – like the NI-STOI - blind in the sense that no transcript of the speech is required. As the MTD estimate is calculated based on phonemes without analyzing the semantic or syntactic context, it can be recalculated very quickly compared to other ASR based methods.

Summary and Outlook

- The non-blind BSIM version with the simple SII back-end and with useful-detrimental of the separation of the BRIR in the front-end gives optimal predictions even for conditions with complex binaural and temporal interactions. However this model requires not only auxiliary information about the clean speech and noise signals but also about the BRIR.
- If BSIM should be applied blindly, the front-end has to be replaced by a blind front-end that analyses only the mixed signal without any auxiliary information about the BRIR. This model is not able to do a useful/detrimental separation of the reverberated speech signal.
- The relatively simple NI-STOI, the MTD and FADE have been evaluated as blind back-ends and reached in some conditions nearly the same accuracy in combination with bBSIM as the non-blind reference model.
- In principle real-time versions of BSIM are possible that update the estimate of intelligibility within fractions of a

second and which can be used for example for controlling the binaural processing strategies of hearing aids.

- So far the blind BSIM approach described here is not applicable to speech-in-speech conditions, because it assumes that the interfering signal has modulations that are not speech-like.
- This problem has to be addressed in future studies by including localization models into the front-end and by using back-ends that are able to estimate the intelligibility of several concurrent talkers.

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 352015383 – SFB 1330 A 1.

References

[1] Durlach, N.I.: Equalization and Cancellation Theory of Binaural Masking-Level Differences. The Journal of the Acoustical Society of America, 35 (1963), 1206–1218

[2] Beutelmann, R., Brand, T., Kollmeier, B.: Prediction of binaural speech intelligibility with frequency-dependent interaural phase differences. The Journal Acoustical Society of America, 126 (2009), 1359-68

[3] vom Hövel, H. Zur Bedeutung der Übertragungseigenschaften des Außenohrs sowie des binauralen Hörsystems bei gestörter Sprachübertragung. RWTH Aachen, Aachen, 1984

[4] Beutelmann, R. & Brand, T.: Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. The Journal of the Acoustical Society of America, 120 (2006), 331–342

[5] 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, 127 (2010), 2479–2497

[6] Hauth, C.F., Berning, S.C., Kollmeier, B., & Brand, T.: Modelling binaural unmasking of speech using a blind binaural processing stage. Trends in Hearing, 24 (2020), 1-16

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

[8] ANSI. ANSI S3.5-1997, American national standard methods for calculation of the speech intelligibility index. Am. Natl. Stand. Institute, Ney York (1997)

[9] Rennies, J., Warzybok, A., Brand, T., & Kollmeier, B.: Modeling the effects of a single reflection on binaural speech intelligibility. The Journal of the Acoustical Society of America, 135(2014), 1556– 1567

[10] Rennies, R., Warzybok, A., Brand, T., & Kollmeier, B.: Measurement and Prediction of Binaural-Temporal Integration of Speech Reflections. Trends in Hearing, 23 (2019), 1–22

[11] Hauth, C.F., & Brand, T.: Modeling sluggishness in binaural unmasking of speech for maskers with time-varying interaural phase differences. Trends in Hearing, 22 (2018), 1–10

[12] Steeneken H.J.M. & Houtgast, T.: A physical method for measuring speech-transmission quality. The Journal of the Acoustical Society of America, 67 (1980), 318-326

[13] Holube, I. & Kollmeier, B.: Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. The Journal of the Acoustical Society of America, 100 (1996), 1703–1716

[14] Taal, C.H., Hendriks, R.C., Heusdens, R., & Jensen, J.: An Algorithm for Intelligibility Prediction of Time – Frequency Weighted Noisy Speech. IEEE Transaction on Audio, Speech, and Language Processing, 19 (2011), 2125–2136

[15] Andersen, A.H., de Haan, J.M., Tan Z.-H., Jensen, J.: A non-intrusive short-time objective intelligibility measure. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) New Orleans, United States (2017) 5085–5089

[16] Schädler, M.R., Hülsmeier, D., Warzybok, A., Hochmuth, S., Kollmeier, B. Microscopic Multilingual Matrix Test Predictions Using an ASR-Based Speech Recognition Model. Interspeech 2016, (2016)

[17] Schädler, M.R. & Kollmeier B.: Separable spectro-temporal Gabor filter bank features: Reducing the complexity of robust features for automatic speech recognition. The Journal of the Acoustical Society of America, 137 (2015), 2047–2059

[18] Huber, R., Krüger, M., & Meyer B.T.: Single-ended prediction of listening effort using deep neural networks. Hearing Research, 359 (2018), 40-49

[19] Roßbach, J., Röttges S., Hauth, F.C., Brand, T., & Meyer, B.T.: Non-intrusive binaural prediction of speech intelligibility based on phoneme classification, ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings (2021)

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

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