Objective Prediction of Hearing Aid Benefit Across Listener Groups Using Machine Learning: Speech Recognition Performance With Binaural Noise-Reduction Algorithms

Marc R. Schädler¹, Anna Warzybok¹, and Birger Kollmeier¹

Abstract
The simulation framework for auditory discrimination experiments (FADE) was adopted and validated to predict the individual speech-in-noise recognition performance of listeners with normal and impaired hearing with and without a given hearing-aid algorithm. FADE uses a simple automatic speech recognizer (ASR) to estimate the lowest achievable speech reception thresholds (SRTs) from simulated speech recognition experiments in an objective way, independent from any empirical reference data. Empirical data from the literature were used to evaluate the model in terms of predicted SRTs and benefits in SRT with the German matrix sentence recognition test when using eight single- and multichannel binaural noise-reduction algorithms. To allow individual predictions of SRTs in binaural conditions, the model was extended with a simple better ear approach and individualized by taking audiograms into account. In a realistic binaural cafeteria condition, FADE explained about 90% of the variance of the empirical SRTs for a group of normal-hearing listeners and predicted the corresponding benefits with a root-mean-square prediction error of 0.6 dB. This highlights the potential of the approach for the objective assessment of benefits in SRT without prior knowledge about the empirical data. The predictions for the group of listeners with impaired hearing explained 75% of the empirical variance, while the individual predictions explained less than 25%. Possibly, additional individual factors should be considered for more accurate predictions with impaired hearing. A competing talker condition clearly showed one limitation of current ASR technology, as the empirical performance with SRTs lower than −20 dB could not be predicted.

Keywords
speech recognition, speech perception modeling, hearing impairment, hearing aid

Date received: 21 June 2017; revised: 5 January 2018; accepted: 22 February 2018

Introduction
With worldwide increasing life expectancy and (age-related) hearing disorders, a steadily increasing demand for high-performance hearing aids (HA) with effective algorithms that improve the everyday communication will occur. To make the development of HA algorithms more efficient and to improve the prescription of a certain HA processing scheme to the individual hearing-impaired (HI) user, valid and reliable objective methods for the prediction of the individual benefit are highly desirable. The current article therefore proposes and evaluates a new method for pursuing this task by extending the recently proposed framework for auditory discrimination experiment (FADE) prediction (Schädler, Warzybok, Ewert, & Kollmeier, 2016) applied to the individual aided speech-in-noise recognition performance with and without a given HA algorithm in realistic binaural listening conditions. This extension includes the simulation of the respective HA processing for the individual user, an approximate simulation of the effect of the individual hearing loss

¹Medizinische Physik and Cluster of Excellence “Hearing4all,” Carl von Ossietzky Universität Oldenburg, Germany

Corresponding author:
Marc R. Schädler. Medizinische Physik and Cluster of Excellence “Hearing4all,” Carl von Ossietzky Universität Oldenburg, Küpkersweg 74, D-26129 Oldenburg, Germany.
Email: marc.r.schaedler@uni-oldenburg.de
(HL), and an approximation of the binaural hearing performance by considering the better ear listening effect.

Many signal processing schemes have been proposed with the aim of improving speech recognition in noisy environments, with the most promising ones using multiple microphone input signals (as opposed to single-channel, monaural processing schemes, see Kollmeier & Kiessling, 2016 for a review). Because all HA algorithms operate best under certain assumptions about the desired target signal and undesired interfering sound components, the real-life performance of these algorithms with arbitrary acoustic situations is hard to predict, especially for real users with impaired hearing. To assess the utility of a hearing device, comprehensive performance tests have to be carried out in a variety of acoustical situations and a variety of individual users with a hearing impairment both in aided and unaided conditions. This makes the determination of the optimum HA processing and, even more so, the development of new and more effective hearing devices, very time and effort consuming. Consequently, suitable and accurate models of the binaural speech recognition for listeners with impaired hearing that are applicable to unaided and aided realistic conditions are desired to make the development-evaluation cycle of HA algorithms more efficient.

As outlined in more detail in Table 1, numerous models that can be used to predict the performance of human listeners in speech recognition tests in noise consider either the effect of hearing impairment or (non-linear) signal preprocessing (e.g., by HAs). Fewer approaches were proposed which consider a combination of, and hence, the interplay between both. Many models were not designed to be used with fluctuating noise maskers, which are prevalent in many realistic communication environments. Finally, only few models can work natively on binaural signals.

### Table 1. Classification of Models That Can Be Used to Predict the Speech Recognition Performance in Noise of Human Listeners.

| Model       | Authors                  | (hearing impairment) | (aided hearing) | (fluctuating interferer) | (binaural hearing) | (intrusive classification) | (reference-based) |
|-------------|--------------------------|----------------------|----------------|--------------------------|-------------------|--------------------------|------------------|
| PEMO + DTW 1| Holube and Kollmeier (1996) | x                    | x              | –                        | –                 | –                        | –                |
| PEMO        | Holube et al. (1997)     | x                    | –              | x                        | –                 | 1                        | x                |
| SII         | ANSI 1997 (1997)         | x                    | –              | –                        | 1                 | x                        | –                |
| STI         | IEC (1998)               | –                    | –              | –                        | –                 | 1                        | x                |
| CSII        | Kates and Arehart (2005) | x                    | x              | –                        | –                 | 2                        | x                |
| Glimpsing   | Cooke (2006)             | –                    | –              | x                        | –                 | 1                        | x                |
| HMM         | Stadler et al. (2007)    | x                    | –              | x                        | –                 | 1                        | x                |
| PEMO + DTW 2| Jürgens and Brand (2009) | x                    | x              | –                        | 2                 | x                        | –                |
| SII         | Rhebergen et al. (2010)  | x                    | –              | x                        | –                 | 1                        | x                |
| BSIM (revised)| Beutelmann et al. (2010)| x                    | x              | –                        | 1                 | x                        | –                |
| STOI        | Taal et al. (2011)       | –                    | x              | –                        | –                 | 2                        | x                |
| mr-sEPSM    | Jørgensen et al. (2013)  | –                    | –              | x                        | –                 | 1                        | x                |
| SRMR        | Falk et al. (2013)       | x                    | x              | –                        | –                 | x                        | –                |
| HASPI       | Kates and Arehart (2014) | x                    | x              | –                        | –                 | 2                        | x                |
| ESTOI       | Jensen and Taal (2016)   | –                    | x              | x                        | –                 | 2                        | x                |
| mr-GPSM     | Biberger and Ewert (2017)| –                    | x              | x                        | –                 | 2                        | x                |
| Original FADE| Schädler et al. (2016a) | –                    | x              | x                        | –                 | –                        | –                |
| FADE        | Current                  | x                    | x              | x                        | –                 | –                        | –                |

Note. “HI” indicates that the model was designed to take individual impaired hearing into account. “HA” indicates that the model can take nonlinear signal preprocessing, for example, by hearing-aid algorithms, without modifications into account. “F” indicates that the model is designed to be used with stationary and fluctuating noise maskers. “B” indicates that the model can work on binaural signals. “I” indicates whether the models use intrusive classification and require the separated/unmixed speech or noise signals after (1) or before (2) potential nonlinear signal processing. “R” indicates that the output of the model depends on empirical data from speech recognition experiments with human listeners.

FADE = framework for auditory discrimination experiment; PEMO = perception model; DTW = dynamic-time-warp; STI = Speech Transmission Index; SII = Speech Intelligibility Index; ESII = extended SII; CSII = coherence SII; HMM = Hidden Markov Model; mr-sEPSM = multiresolution speech-based Envelope Power Spectrum Model; mr-GPSM = multiresolution Generalized Power Spectrum Model; BSIM = Binaural Speech Intelligibility Model; STOI = Short-Time Objective Intelligibility; ESTOI = extended STOI; HASPI = Hearing-Aid Speech Perception Index; SRMR = Speech-to-reverberation Modulation Energy Ratio.
In addition to possible use cases, models can be characterized by their assumptions and the underlying a priori information entering the prediction of speech reception thresholds (SRTs) that are quite diverse and can limit their applicability. Two distinct categories of assumptions can be identified that have far-reaching implications:

- **Superhuman a priori knowledge of various degrees**, such as access to separated speech or noise signals in any other way than it would be required to perform the speech test, that is, to generate mixed signals with a defined signal-to-noise ratios (SNRs). Such a priori knowledge could include, for example, frozen signal parts, ideal binary masks, or calculation of time/frequency-dependent SNRs. We will refer to this category as models with intrusive classification.

- **Outcome calibration using an empirical reference to predict the outcome of a speech recognition test**, which generally means that empirical speech recognition data are used to set parameters of the model. This is accompanied by the assumption that the model correctly reflects desired properties of the real world in the reference condition because the prediction in the reference condition is by definition “correct.” We will refer to this category as (empirical) reference-based models.

Here, we define single-ended classification as nonintrusive and double-ended classification as intrusive, in close analogy to the case of empirical measurements of SRTs. Classification in this context refers to the decision for the recognized items, for example, words, or any mapping to a scale in the evaluation of the recognized items, for example, word recognition rates. By this definition, measurements of empirical SRTs are nonintrusive, even if the listeners heard the clean sentences before performing the speech recognition test. Figure 1 illustrates the definition with some examples. Hence, the best example for nonintrusive classification is the measurement of SRTs with human listeners. To perform the speech test, separate speech and noise signals are required to mix them at a desired SNR and optionally process them with some arbitrary algorithm before presenting the mixed and processed signal to the listener. During measurements of SRTs, listeners have access only to the mixed and processed signals. Prior to the measurements with the matrix sentence test, a few training lists usually allow the listeners to adapt to the speech material; and even clean speech material might be presented. The measurement itself is intrusive because the separate signals are required to perform it, but the classification (here the human listener) is not because she or he does not rely on signals prior to mixing and processing to solve the speech recognition task, for example, by listening to the clean speech or noise signal on one ear. An example for a possibly nonintrusive measurement of an SRT with human listeners would be to ask them to estimate the SRT.

Models with intrusive classification solve a different task than the human listeners, as they have access to more information about the solution. Please note, however, that access to the separate signals does not reveal the empirical SRT. In other words, knowing the exact testing conditions is different information than knowing the empirical outcome of the corresponding test when it is performed with human listeners.

The category of models with intrusive classification can be subdivided into those models that require (parts of) the separate signals (1) after or (2) before a potential nonlinear signal processing (e.g., compression, clipping, or frequency shifts by HAs). The first subdivision allows for a combination with arbitrary signal processing algorithms, for example, to predict the benefit of HAs, while the second subdivision does not directly provide this...
option because adding signal and noise and the nonlinear processing is generally not permutative. A couple of modeling approaches that can be used to predict the speech recognition performance in noise of human listeners are presented in Table 1 together with the corresponding classification whether

- they were designed to take hearing-impairment into account (HI),
- they can be used with nonlinear hearing-aid algorithms (HA),
- they were designed to be used with fluctuating noise maskers (F),
- they can be used in binaural conditions (B),
- they use intrusive classification (I),
- or reference-based (R).

Please note that Table 1 attempts a coarse technical classification of existing modeling approaches, reflecting the necessary requirements for aided predictions in realistic binaural listening conditions and contains by no means statements about the extensibility nor prediction accuracy for any particular model or task.

The overview in Table 1 reveals that many models are designed for rather specific use cases, and no approach fulfilled all the technical requirements for aided predictions in realistic binaural listening conditions. SNR-based models, such as the Speech Transmission Index (IEC, 1998; Steeneken & Houtgast, 1980, 2002), the speech intelligibility-weighted signal-to-noise ratio (iSNR; Greenberg, Peterson, & Zurek, 1993), the Speech Intelligibility Index (SII; ANSI, 1997), and similar or derived methods, such as the extended SII (Rhebergen, Versfeld, de Laat, & Dreschler, 2010), the revised Binaural Speech Intelligibility Model (BSIM; Beutelmann, Brand, & Kollmeier, 2010), the multiresolution speech-based Envelope Power Spectrum Model (Jørgensen, Ewert, & Dau, 2013), or the multiresolution Generalized Power Spectrum Model (Biberger & Ewert, 2017), derive an estimate of the SNR or an internal SNR in different frequency or modulation bands. These approaches aim to represent signal properties that are highly correlated with speech intelligibility and hence with speech recognition performance. To calculate these properties, they require separate speech and noise signal parts after possible signal processing. This is why they cannot be used with nonlinear signal processing algorithms without significant modifications, such as using the separation method from Hagerman and Olofsson (2004; cf. Figure 1). Hence, their mapping, that is, classification, is Intrusive (1). Some of these models specifically make assumptions about the speech material, such as the SII with the band importance functions that were determined on empirical speech recognition results.

The Glimpsing Model from Cooke (2006) calculates (binary) SNR masks from time-frequency bins of a spectrotemporal representation of the clean speech and noise signals, and the Hidden Markov Model (HMM)-based approach from Stadler, Leijon, and Hagerman (2007) builds descriptive models from the separate clean speech and noise signals. The way they require the separate signals in their presented configuration does not allow using them without modifications with nonlinear signal processing algorithms, similar to the first group of models. Hence, these approaches also use Intrusive (1) classification. In contrast to the first group, these approaches train automatic speech recognizer (ASR) models and (could) perform a speech recognition task. They have the potential to make predictions of SRTs without relying on empirical reference data, which, however, was not shown.

Correlation or coherence-based models, such as the coherence SII (Kates & Arehart, 2005), the Short-Time Objective Intelligibility (STOI) measure (Taal, Hendriks, Heudsens, & Jensen, 2011), the extended STOI (Jensen & Taal, 2016), or the Hearing-Aid Speech Perception Index (Kates & Arehart, 2014) quantify the presence of properties (often certain modulation patterns) of the clean input speech signal in the noisy processed output signal. Similar to the first group, they aim to represent signal properties that are highly correlated with speech intelligibility. In contrast to the first group, they do not require the separate signal parts after but before signal processing. Hence, they can be used in combination with nonlinear signal processing algorithms without further modification, which was also one of their design goals. Because their classification still requires additional signals apart from the noisy processed speech signal, it is Intrusive (2). Further, the signals before and after processing need to be carefully aligned in time, which excludes all signal processing algorithms that do not allow a temporal alignment, such as accelerated speech. STOI and extended STOI do not take individual impaired hearing into account. Further, coherence SII and Hearing-Aid Speech Perception Index were designed to consider individual impaired hearing but were not shown predict the effect of fluctuating noise maskers.

The only measure that does not use intrusive classification is the speech-to-reverberation modulation energy ratio (Falk, Cosentino, Santos, Suelzle, & Parsa, 2013), which is calculated solely from the noisy processed signal. It is even able to predict an SRT without any knowledge about the clean speech or noise signal, but still not able to predict speech recognition scores or SRTs without empirical reference data. Furthermore, this approach was not designed for modulated maskers and binaural conditions.

All these models do not predict the outcome of a speech recognition test directly—although some
could—but a (possibly highly) correlated objective measure. For a direct comparison with empirical data, these objective measures need to be mapped to the outcome of a speech-in-noise recognition experiment, which is either a recognition rate depending on the SNR or directly an SRT. This mapping is generally calibrated with or fitted to empirical data from speech recognition experiments and hence, reference-based.

There are fundamental disadvantages of models with reference-based mappings or intrusive classification. The disadvantage of exploiting information that is not accessible to the human listener, who must perform the task with noisy processed stimuli, is a potential uncontrolled bias of the prediction towards unrealistically high performance levels. For human listeners, the limited information results in a desired natural lower boundary for their performance in speech recognition tests, which is commonly quantified with the SRT. Models with intrusive classification intentionally circumvent the difficulties associated with solving the task, possibly at the cost of universality. The disadvantage of using empirical reference data is that a part of the model is undetermined by design and needs to be filled in with empirical data that compensates for any unexplained detail. Apart from requiring empirical measurements prior to predictions, the selection of a specific empirical reference makes implicit assumptions about the generalization ability towards unknown conditions and the model construction. These assumptions might eventually be violated if the predicted conditions differ from the reference condition (e.g., different noise type, different SNR, different vocabulary, different speaker, or different nonlinear channel characteristics, etc.). If the prediction accuracy depends on the selection of the reference (which often is the case), the universality of the model is clearly limited. Further, the models are not strictly objective when empirical reference data from speech recognition experiments influence the outcome of the predictions. An unpleasant “synergistic” effect in models that use reference-based mapping and intrusive classification is that the former conceals any indication of wrong assumptions of the latter. In other words, a bias due to superhuman knowledge can neither be detected nor quantified.

A first step to address these drawbacks was already taken by Holube and Kollmeier (1996), who used a dynamic-time-warp (DTW)-based speech recognizer with a modified version of the perception model (PEMO) from Dau, Püschel, and Kohlrausch (1996) as a front-end to estimate word recognition rates depending on the SNR, hence measuring the psychometric function of what can be considered a very simple ASR system. In its initially presented form, it uses nonintrusive classification because it performs the speech recognition task (on a word level) comparing noisy test words with noisy response alternatives. Therefore, it could be even used with nonlinear signal processing. Further, it is the only discussed approach that does not need empirical reference data to predict the outcome (recognition probability depending on the SNR). The approach in the current study shares many properties and goals with the modeling approach from Holube and Kollmeier (1996). However, to achieve an acceptable prediction accuracy on a logatome recognition task, Jürgens and Brand (2009) found that the approach required knowledge about the clean speech samples when performing the speech recognition task, that is, using intrusive classification. Its use with fluctuating noise maskers did not provide satisfactory prediction accuracies, and there are reasonable doubts that the approach in its presented form would generalize to realistic fluctuating noise conditions (Holube, Wesselkamp, Dreschler, and Kollmeier, 1997). In other words, the system was not sufficiently robust against fluctuating noise maskers.

In accordance with this observation, Schädler, Hülsmeier, Warzybok, Hochmuth, and Kollmeier (2016a) recently found that independent frequency bands (as assumed, e.g., with the DTW + PEMO approach) for speech recognition in fluctuating noise masker seem to be an untenable assumption. In the domain of ASR, and much in contrast to many models of auditory signal processing, integration across frequency bands is regarded an essential feature to solve the task of speech recognition in noise, and particularly, in fluctuating noise conditions. In other words, the currently used front-ends in ASR and the vast majority of speech intelligibility models do not agree on which information of an audio signal is most relevant for robust speech recognition. To improve the DTW-based approach from Holube and Kollmeier (1996), it seems appropriate to use a robust ASR system to improve the recognition performance and achieve better predictions of human speech recognition performance at the same time.

It is important to differentiate the aims of auditory-motivated feature extraction (like in computational auditory scene analysis [CASA]), robust automatic speech recognition, and the prediction of human speech recognition performance. The aim of CASA is to extract high-level information from an audio signal, where high-level information here means anything that makes sense to us humans. The solutions can be technical and don’t—need to—adhere to what is commonly understood as auditory principles, which is why it is also referred to as computational acoustic scene analysis. However, many of these high-level information extraction tasks can be performed extremely well by human listeners, which is why they serve as a “model” as long as they perform better than their technical counterparts. Hence, the interest there is limited to understand the principles of—or even just mimic—the model solution to better solve the task of
information extraction. The same is true for any form of bionics. From this perspective, robust ASR is a special problem in CASA, which addresses the transcription of recordings of spoken speech, an extremely challenging and useful high-level information extraction task. Still, the criterion for a good solution is not adherence to any auditory signal processing principles but the raw recognition performance in relevant conditions. Coincidentally, some parts of robust ASR systems implement basic principles of auditory signal processing, such as a frequency analysis on a Mel or Bark scale, the compression of amplitude values with a logarithm and the extraction of spectral (e.g., Mel-frequency cepstral coefficients) and—less prominent—temporal (e.g., temporal derivatives) modulation patterns. Again, this is the current state because it is the best technical solution in terms of recognition performance; still no generally superior solution has been found. Despite recent advances in the field of ASR (primarily due to the use of deep neural networks, efficient training algorithms, hardware acceleration, and large databases), robust speech recognition of spontaneous speech in realistic multitalker environments is still a stronghold of human listeners, turning humans into models for robust speech recognition. Because of the many nonlinearly interacting components of ASR systems, it is still unknown why human listeners perform better, that is, which components of an ASR system must be improved to put it on a par with humans in the listening experiments, that is, the recognition of words/sentences presented in the presence of noise at different SNRs. The desired outcome, for example, an SRT at 50% correct word recognition rate, can be calculated directly from the evaluation of the transcripts of the ASR system, without using any empirical data. Because the SRT predictions do not depend on empirical reference data, they are objective. The nonintrusive classification of noisy signals with FADE puts a natural lower bound on the recognition performance that strongly depends on the signal representation and was found to be in good agreement with empirical measurements of listeners with normal hearing (Schädler et al., 2015). The approach was shown to predict the outcome of the matrix sentence test in different languages in stationary and fluctuating noise conditions (Schädler et al., 2016a). The auditory signal representation, which was originally designed for robust automatic speech recognition, was extended by common signal processing deficiencies to model impaired hearing (Kollmeier, Schädler, Warzybok, Meyer, & Brand, 2015, 2016). The impairment can be individualized using the individual absolute hearing threshold.

In the current work, FADE is evaluated for the first time as an objective method for SRT predictions in aided conditions for normal-hearing and HI listeners in binaural realistic listening conditions. The extent to which this model can explain the human behavior is assessed with the empirical data from Völker, Warzybok, and Ernst (2015) who measured SRTs of 10 listeners with normal hearing and 12 listeners with aided impaired hearing with eight different noise-reduction algorithms in three binaural acoustic conditions. There, the empirical data were modeled with the SII-based BSIM from Beutelmann et al. (2010). BSIM only predicts SRTs relative to a reference condition, which is why its predictions are always biased to some extent by the empirical reference data. Because this model requires separate speech and noise signals, which are not available anymore after the nonlinear processing of the algorithms, Völker et al. (2015) did not perform individual aided predictions. This limitation could be circumvented by using a technique called shadow filtering, which attempts to separate the signal parts by applying the nonlinear processing to the speech and noise signals separately. However, shadow
filtering would have introduced additional assumptions about the linearity of the signal processing.

To enable objective predictions of SRTs in the cafeteria condition for listeners with normal hearing, the FADE approach is extended towards binaural conditions by implementing a simple better ear listening strategy. Further, to enable individual predictions of SRTs, the model is individualized by taking into account the absolute hearing thresholds (left and right) given by the audiograms and the configuration of the hearing devices (left and right). The individual and group predictions for unaided and aided conditions for listeners with normal hearing and impaired hearing are compared with the empirical data in terms of SRTs and benefit in SRT, that is, the improvement in SRT due to activation of noise suppression scheme compared with the reference condition (with deactivated noise suppression).

Methods

Experiment

The stimuli and the empirical SRTs were taken from Völker et al. (2015). There, SRTs of 10 listeners with normal hearing and 12 listeners with aided impaired hearing were adaptively measured with the German matrix sentence test in a binaural cafeteria environment that was simulated with head-related impulse responses from Kayser et al. (2009) with eight different binaural preprocessing strategies. Three different noises were employed to determine SRTs in a condition where the target speaker was located in front of the listener: (a) A 20-talker babble noise (20T), (b) the ambient sound in the cafeteria in which the head-related impulse responses were recorded (cafeteria ambient noise [CAN]), and (c) a single interfering talker (SIT). In this study, the focus was on the natural, recorded ambient sound cafeteria, which was assumed to represent the most realistic condition. The preprocessing algorithms included adaptive differential microphones (ADM), a coherence filter, single-channel noise reduction (SCNR), and binaural beamformers and are listed in Section Binaural Preprocessing Algorithms. In addition, a condition with no preprocessing (NoPre) served as a reference. To provide listeners with impaired hearing with audible signals in all conditions, the multiband dynamic compressor of the Master Hearing Aid (MHA) was used to compensate their HL with the nonlinear fitting procedure by the National Acoustic Laboratories (Byrne, Dillon, Ching, Katsch, & Keidser, 2001), even in the reference condition with no binaural preprocessing. For the group of listeners with normal hearing, no amplification was provided. The group of listeners with impaired hearing consisted of eight male and four female listeners with ages ranging from 21 to 66 years. Their hearing thresholds were 43.9 dB higher on average over normal-hearing listeners between frequencies of 500 Hz and 4 kHz and ranged from 20 dB to 55 dB HL. For more details on the audiometric data and the compensation of HL, please refer to Völker et al. (2015).

Prediction of SRTs With FADE

The outcome of the German matrix test was predicted with FADE (Schädler et al., 2016b). The condition here is defined by the task (matrix sentence recognition), the speech signals, the noise signal, the individual signal processing, and the individual audibility. Therefore, the ASR system is trained and tested for each condition separately, that is, there is no interdependence regarding the conditions. It is particularly noteworthy that no prior information about any empirical SRT is required to obtain a predicted SRT. In other words, no reference to another condition is required, and an SRT can be predicted for an isolated condition.

The procedure to obtain the predicted outcome in one condition with FADE is as follows:

Gaussian Mixture Model (GMM) and HMM-based ASR systems are trained on all available matrix sentences; one for each considered SNR (−24 dB to 6 dB in 3-dB steps). Therefore, random portions of the noise signal were chosen to generate the training data. The mixed recordings were then processed with the (individual) HA using the MHA (Grimm, Herzke, Berg, & Hohmann, 2006) with the algorithm that corresponds to the condition. From the processed signals, features were extracted that incorporate the individual hearing thresholds as explained in Section Model of Impaired Hearing. The feature extraction is based on separable Gabor filter bank (SGBFB) features with mean-and-variance normalization. The SGBFB feature set was shown to robustly extract information of speech signals from audio recordings (Schädler & Kollmeier, 2015) and to be suitable for modeling tone detection and speech recognition experiments with normal-hearing listeners (Schädler et al., 2016b). SGBFB features encode a range of spectrotemporal modulation patterns, where mean-and-variance normalization removes information about the absolute level and the overall dynamic range of the recording. For the acoustic model, word models with six states were used. One GMM/HMM word model was trained for each word and SNR. For each SNR, the 50 word models were combined with a language model that describes the (language-independent) syntax of the matrix test; that is, 5 words with 10 options each.

The ASR systems are then used to recognize all available matrix sentences at all considered SNRs (−24 dB to 6 dB in 3-dB steps). Like for the training data, random portions of the noise signal were chosen to generate the testing data. The mixed recordings were, again like for
the training data, processed with the (individual) HA using MHA with the algorithm that corresponds to the condition. From the processed signals, features that incorporate the individual hearing thresholds were extracted, like for the training data. This training/testing scheme is also known as matched training, an ideal training condition. It is employed to get an estimate of the lowest achievable SRT in the given condition. Recognition performance of the ASR system is subject to the same limitations that trained human listeners encounter due to the task, the speech material, the noise signal, the individual signal processing, and the individual audibility. From the transcripts, the word recognition rates in percent are determined for each combination of training and testing SNR and compiled into a square matrix, with training SNR in one dimension and testing SNR in the other dimension. This matrix is referred to as the recognition result map (cf. Schädler et al., 2016b). Usually the testing SRT depends on the training SNR. The lowest testing SNR at which 50% of the words were correctly recognized was interpolated between the 3 dB-spaced samples and taken as the predicted SRT. For further details about the simulations, please refer to the original publications (Schädler et al., 2015, 2016b). The concept of how the predictions with FADE were individualized is illustrated in Figure 2.

![Figure 2. Illustration of the individualization of predictions with FADE.](image)

The feature extraction and HAs were individualized based on the audiogram, which is explained in Sections “Model of Impaired Hearing” and “Model of Binaural Hearing.” FADE is free software and available online.1

### Model of Impaired Hearing

The feature extraction of the original publication from Schädler et al. (2016b) was modified to implement an absolute hearing threshold. The implementation of the hearing threshold is identical to the one from Kollmeier et al. (2016), where the absolute hearing threshold and a suprathreshold parameter were successfully used to predict individual outcomes of the German matrix sentence test in noise for listeners with impaired hearing. In this work, only the hearing thresholds were considered because no suitable empirical data existed to determine individual estimates of the suprathreshold component. To account for the loss of sensitivity, which is usually assessed by the audiogram, the individual hearing threshold was applied to the spectrotemporal representation, the log Mel-spectrogram, by setting any input spectrogram level to the audiogram level if it was below this value. The effect of this implementation of the hearing threshold on the log Mel-spectrogram of a noisy speech sample is illustrated in Figure 3. Applying an individual threshold has the effect that any patterns below have no effect on the feature vector anymore. Hence, part of the speech information is concealed from the recognizer, which naturally results in individually lower recognition performance and increased SRTs.

### Model of Binaural Hearing

The measurements in Völker et al. (2015) were dichotic, which to model required a stage that integrated the two acoustic signals from both ears. Different models of binaural hearing and signal processing were proposed in the literature (e.g., Beutelmann et al., 2010); however, due to using intrusive classification, none of them works blindly. In other words, these models require information that is not accessible in the current setup, where only noisy, processed signals are available, and hence could not be used. Here, to limit the complexity of the approach, independent feature vectors were calculated for each ear, which were then concatenated to form one binaural feature vector. This very simple binaural model assumes a late, that is, cortical, integration of binaural information in speech processing. Because the recognizer can learn a better ear during the training stage, we denominate it as automatic better ear listening. An implementation of the feature extraction is available online.2

---

1. FADE is free software and available online.
2. An implementation of the feature extraction is available online.
Binaural Preprocessing Algorithms

The eight binaural signal preprocessing strategies were taken from Völker et al. (2015) and included (1) No preprocessing (NoPre), (2) ADM (Elko & Pong, 1995), (3) ADM with coherence-based noise reduction (ADM + coh; Grimm, Hohmann, & Kollmeier, 2009), (4) SCNR (e.g., Gerkmann, Breithaupt, & Martin, 2008), (5) fixed binaural minimum variance distortionless response (MVDR) beamformer (e.g., Van Veen & Buckley, 1988), (6) adaptive binaural MVDR beamformer (e.g., Griffiths & Jim, 1982), (7) common postfilter (com PF) based on fixed MVDR (e.g., Simmer, Bitzer, & Marro, 2001), (8) com PF based on adaptive MVDR (e.g., Simmer et al., 2001), and (9) individual postfilter (ind PF) based on adaptive MVDR (e.g., Simmer et al., 2001). The algorithms were described and instrumentally evaluated by Baumgärtel et al. (2015) in the same experimental conditions. ADM and SCNR are monaural algorithms and do not require a binaural link, that is, they work independently on both ears. The remaining six algorithms are binaural preprocessing strategies, where the signal input on one side can influence the output on the other side. All algorithms aim to improve speech intelligibility; however, they achieve that goal only to a varying degree (Völker et al., 2015). The goal of the current study is to predict this benefit, which can be measured in improvement in SRT. For details on the individual algorithms please refer to Völker et al. (2015), Baumgärtel et al. (2015), or the corresponding original publications. It is assumed that this selection provides a good representation of modern binaural speech and audio signal processing algorithms, and hence also reflects typical effects, nonlinearities, and artifacts involved in such processing. In the current study, the signal processing was performed by using the MHA, which was also used by Völker et al. (2015) to carry out the measurement of the empirical data. For the individual predictions, the same individual configuration that was used for the empirical measurements was used to process the noisy training and testing data (cf. Section “Prediction of Speech Reception Thresholds With FADE”) before individual feature vectors that considered the individual hearing thresholds were extracted (cf. Section “Model of Impaired Hearing”).

Results

Predicted SRTs

The predicted SRTs are plotted against the empirical SRTs in Figure 4 for the 20T condition and the CAN condition, and median SRTs for the groups of listeners with normal hearing and aided impaired hearing in all conditions are presented in numerical form in Table 2. It should be noted that the FADE approach in its current form is not suitable to predict the recognition thresholds for the conditions with an SIT. The model seems to be unable to differentiate between the two competing speakers, and no a priori information was provided to the algorithm which speaker is the target and which is the interferer. Without a cue to differentiate the speakers, the FADE model can only perform the recognition task at much higher SNRs than achieved by human listeners. This is reflected by the finding that the predicted SRTs for listeners with normal hearing in the SIT condition were about 10 dB higher than the empirical ones. Reasonable predictions for the generally very homogeneous group of listeners with normal hearing can be considered a prerequisite to performing reasonable predictions for the generally less homogeneous group of listeners with impaired hearing, which is why no predictions were performed for listeners with impaired hearing in this condition. Hence, in the following, the extreme (SRTs below −20 dB) SIT condition is not considered, and the analysis focuses on the artificial 20T condition and, primarily, on the realistic CAN condition. However, for completeness, the available data for the SIT condition are provided in numerical form.
For the 20T and CAN conditions, the median empirical SRTs of both groups (normal and impaired hearing) show an effect size of about 5 dB due to the different binaural preprocessing algorithms. The median predicted SRTs show a very similar effect size. The median empirical SRTs for listeners with normal hearing are lower than the ones for listeners with impaired hearing, which is also correctly predicted by the model. The median predictions for both groups (black symbols) were mostly underestimated, that is, the predicted SRTs were higher than the empirical ones. This deviation is more pronounced in the CAN than in the 20T condition. However, the predicted values seem to be scattered around a parallel to the diagonal, which indicates an
Predicted Benefits in SRT

The benefit in SRT was calculated as the inverted difference between the SRT in a condition and the SRT in the reference (NoPre) condition and reflects the improvement due to using an algorithm (higher values are better). Note that this calculation was performed independently for the predicted and the simulated data, and each individual. Hence, the predictions of benefits did not depend on empirical data (apart from the individual audiograms), and the median values are calculated from differences of SRTs, that is, they are not differences of median SRTs. While the analysis in this section focuses on the realistic CAN condition, the results for the 20T and SIT condition are reported in numerical form. In Figure 5, the predicted individual and group benefits are plotted against the empirical individual and group benefits in the CAN condition for listeners with normal and aided impaired hearing. For listeners with normal hearing (left panel), the data points were closely aligned to the diagonal, which indicates highly accurate predictions. For the group of listeners with impaired hearing (right panel), the deviations from the diagonal were more pronounced than for listeners with normal hearing. This was an expected result due to less homogeneous nature of this group. Nonetheless, the group’s median values were found to be well aligned to the diagonal, which indicates still accurate predictions of the benefit for the group of listeners with impaired hearing in the CAN condition. The predicted group benefits with the different algorithms are directly compared with the empirically determined ones in Figure 6 for the group of listeners with normal hearing and in Figure 7 for the group of listeners with aided impaired hearing. For listeners with normal hearing, the predicted benefits were found to be in very good agreement with the median of the empirical data. The largest deviations were observed for the algorithms with the adaptive MVDR beamformer, which is the most complex algorithm and also resulted in a relatively large empirical interquartile range of about 2 dB. It is notable that the model correctly predicted that the SCNR did not improve speech intelligibility.

Table 3. Statistical Analysis of the Predicted SRTs for the Normal-Hearing Group (NH Median), the Hearing-Impaired Group (HI Median), and the Individual Outcomes of Listeners With Impaired Hearing (HI Individual) With Eight Different Binaural Preprocessing Strategies (Plus the No Processing Reference Condition) in Three Acoustic Conditions: 20-Talker Babble Noise (20T), Cafeteria Ambient Noise (CAN), and Single Interfering Talker (SIT).

| Cond. | Prediction | $R^2$ | Interval | $p$  | $B$  | RMS |
|-------|------------|-------|----------|------|------|-----|
| 20T   | NH median  | 0.87  | 0.51–0.97| <.01 | 1.8  | 1.9 |
| 20T   | HI median  | 0.63  | 0.08–0.91| <.05 | 1.6  | 2.1 |
| 20T   | HI indiv.  | 0.16  | 0.06–0.31| <.01 | 1.4  | 3.2 |
| CAN   | NH median  | 0.92  | 0.66–0.96| <.01 | 2.9  | 2.9 |
| CAN   | HI median  | 0.83  | 0.40–0.96| <.01 | 2.7  | 2.8 |
| CAN   | HI indiv.  | 0.10  | 0.02–0.23| <.01 | 2.4  | 3.9 |
| SIT   | NH median  | 0.04  | 0.00–0.59| >.05 | 13.0 | 13.3 |

Note. Pearson’s correlation coefficients ($R^2$) are reported including their 95% confidence intervals according to Fisher (1958) along with the root-mean-square (RMS) prediction error and the bias ($B$) for predicted SRTs in dB. SRT = speech reception threshold.
For listeners with impaired hearing, the median of the individually predicted benefits is in agreement with the empirical data, however, not to the same degree than it was found for listeners with normal hearing.

Interestingly, the interquartile ranges of the predictions were found to be consistently larger than those of the empirical data. This is probably due to the interaction of the hearing threshold, which is implemented as a hard

Figure 5. Predicted versus empirical benefits in SRT (difference to reference condition, higher values are better) for listeners with normal (NH, left panel) and aided impaired (HI, right panel) hearing in a cafeteria ambient noise condition with eight different binaural preprocessing algorithms. The individual predictions for listeners with impaired hearing are plotted with colored stars, where the color indicates the listener. The dash-dotted lines indicate perfect prediction. The error bars of the median predictions indicate the first and third quartile of the corresponding individual data, where for the predictions for normal hearing listeners, these were assumed to be the median ± 0.5 dB.

SRT = speech reception threshold.

Figure 6. Median predicted and empirical benefits in SRT (difference to reference condition) for the group of listeners with normal hearing in a cafeteria ambient noise condition with eight different binaural preprocessing algorithms: ADM) adaptive differential microphones, + coh) with coherence-based noise reduction, SCNR) single-channel noise reduction, MVDR) binaural minimum variance distortionless response beamformer, com PF) common postfilter based, and ind PF) individual postfilter based. The error bars of the empirical data indicate the first and third quartile of the corresponding individual data, and their length the interquartile range. The error bars of the predicted data indicate an assumed symmetric interquartile range of 1 dB.

SRT = speech reception threshold.
threshold as opposed to the probably smoother transition area from inaudible to audible in human listeners, with the compressive amplification scheme. Small deviations of the hard threshold will have stronger interactions with compressed signal levels than human listeners with smooth transition zones. This effect is reinforced by the low accuracy (±5dB) of threshold measurements in audiograms, which adds an additional noise source to the predictions not present in the empirical data. The benefit of the best and worst performing algorithms, fixed MVDR and SCNR, respectively, was overestimated by about 1 dB, while the deviations in the other conditions were smaller. Again, the predicted results did not suggest a benefit from using the SCNR algorithm. The group benefits in SRT for all conditions are presented in numerical form in Table 4, and a statistical analysis of the individual and group benefits in SRT is provided in Table 5. The proportion of explained variance of the empirical group data was similar in the 20T and the CAN condition. It was higher (about 90%) for the group of listeners with normal hearing than for the group of listeners with impaired hearing (about 75%). The bias and RMS error were particularly low in the CAN condition, with less than 0.5 dB and less than 1.0 dB, respectively. These were found to be slightly higher in the babble noise condition, with about 1 dB and less than 2 dB, respectively. In the SIT condition, the median predicted benefits were not significantly correlated with the empirical data. The predicted benefits for the individual listeners with aided impaired hearing were found to be significantly correlated with the empirical data. However, the proportion of explained variance was, with less than 25%, low, and the RMS prediction error, with more than 2.4 dB, comparatively high.

Correlation of Prediction Errors Across Test Conditions

To test if the pronounced individual prediction errors (difference between the predicted and the corresponding empirical SRT) for listeners with aided impaired hearing in the 20T and the CAN condition (cf. left panel in Figure 5) are systematic, that is, due to a common cause that the model did not consider, the individual prediction errors are plotted against each other in Figure 8. The prediction errors were found to be significantly correlated \( (R^2 = .69, p < .01) \), clearly indicating a systematic error. In other words, when the model over/underestimated an individual SRT in one condition, it also over/underestimated it in the other conditions.

Discussion

The main finding of this article is that the average benefit from several binaural noise-reduction strategies can be surprisingly well predicted using the comparatively simple, ASR-based FADE approach. FADE provides an objective estimate of the SRT that can be individualized to suit an individual listener with hearing impairment.
Völker et al. (2015) modeled the same empirical data with the BSIM (Beutelmann et al., 2010) and found that the difficult—but required for BSIM predictions—post hoc separation of the processed signal into a speech and noise part was a major problem for those conditions where nonlinear signal processing algorithms were used. The problems of Intrusive (1) classification concerns many speech intelligibility models (cf. Table 1). To overcome this problem in modeling, Völker et al. (2015) proposed to use FADE, which was successfully employed in this study. Another—often neglected—problem with many other prediction methods for human speech recognition performance is the lack of objectivity due to being based on empirical reference data, which also makes the

**Table 4.** Median Empirical and Predicted Benefits in SRT (Difference to Reference Condition) in dB for the Group of Listeners With Normal Hearing (NH) and Aided Impaired Hearing (HI) With Eight Different Binaural Preprocessing Algorithms in Three Conditions: 20-Talker Babble Noise (20T), Cafeteria Ambient Noise (CAN), and Single Interfering Talker (SIT).

| Cond. Data | ADM 1) | 1) + coh | SCNR fixed MVDR 3) | adapt MVDR 4) | 5) + com PF 6) | 6) + com PF 7) | 6) + ind PF 8) |
|---|---|---|---|---|---|---|---|
| 20T NH emp. | 4.5 | 4.1 | −0.6 | 4.1 | 3.7 | 3.7 | 3.2 | 3.4 |
| 20T NH pred. | 2.6 | 2.7 | −1.7 | 3.6 | 3.0 | 3.8 | 3.0 | 2.5 |
| 20T HI emp. | 3.5 | 3.4 | −0.7 | 4.1 | 3.8 | 3.4 | 3.1 | 3.2 |
| 20T HI pred. | 3.4 | 3.7 | 0.7 | 6.2 | 4.7 | 5.8 | 4.8 | 4.9 |
| CAN NH emp. | 3.5 | 4.0 | −0.5 | 4.1 | 2.7 | 3.1 | 2.3 | 2.2 |
| CAN NH pred. | 3.6 | 3.9 | −0.7 | 4.2 | 2.1 | 3.3 | 1.1 | 1.1 |
| CAN HI emp. | 2.9 | 3.2 | −1.0 | 3.8 | 3.2 | 3.3 | 1.7 | 2.6 |
| CAN HI pred. | 3.1 | 4.3 | 0.2 | 5.1 | 2.6 | 4.3 | 1.8 | 1.9 |
| SIT NH emp. | 2.4 | 3.0 | −1.4 | 3.2 | 4.1 | 1.3 | 3.2 | 3.2 |
| SIT NH pred. | 4.1 | 4.6 | −3.0 | 3.7 | −0.8 | 3.7 | −1.1 | −1.3 |

Note. Individual predictions for listeners with impaired hearing in the SIT condition were not performed. The numbers in the header identify the algorithms: (1) adaptive differential microphones (ADM), (2) ADM with coherence-based noise reduction (coh), (3) single-channel noise reduction (SCNR), (4) fixed binaural minimum variance distortionless response (MVDR) beamformer, (5) adaptive binaural MVDR beamformer, (6) common postfilter (com PF) based on fixed MVDR, (7) com PF based on adaptive MVDR, and (8) individual postfilter (ind PF) based on adaptive MVDR. Please note that the group predictions of the benefits in SRT do not have to coincide with the corresponding differences in the group predictions of SRTs provided in Table 2.

**Table 5.** Statistical Analysis of the Predicted Benefits (Difference to Reference Condition, Higher Values Are Better) in SRTs for the Normal-Hearing Group (NH Median), the Hearing-Impaired Group (HI Median), and the Individual Outcomes of Listeners With Impaired Hearing (HI Individual) in Three Acoustic Conditions: 20-Talker Babble Noise (20T), Cafeteria Ambient Noise (CAN), and Single Interfering Talker (SIT).

| Model | Cond. | Prediction | R² | Interval | p | B | RMS |
|---|---|---|---|---|---|---|---|
| iSNR⁴ | 20T | NH median | 0.62 | − | <.05 | −3.0 | 3.3 |
| FADE | 20T | NH median | 0.87 | 0.46–0.98 | <.01 | −0.8 | 1.0 |
| FADE | 20T | HI median | 0.75 | 0.17–0.95 | <.01 | 1.3 | 1.5 |
| FADE | 20T | HI indiv. | 0.23 | 0.10–0.39 | <.01 | 1.3 | 2.4 |
| iSNR⁴ | CAN | NH median | 0.35 | − | >.05 | −2.7 | 3.1 |
| FADE | CAN | NH median | 0.91 | 0.57–0.98 | <.01 | −0.4 | 0.6 |
| FADE | CAN | HI median | 0.77 | 0.20–0.96 | <.01 | 0.4 | 0.9 |
| FADE | CAN | HI indiv. | 0.13 | 0.03–0.27 | <.01 | 0.3 | 2.4 |
| FADE | SIT | NH median | 0.07 | 0.00–0.67 | >.05 | −1.1 | 3.1 |

Note. Pearson’s correlation coefficients (R²) are reported including their 95% confidence intervals according to Fisher (1958) along with the root-mean-square (RMS) prediction error and the bias (B) in dB for the predicted SRTs. SRT = speech reception threshold; RMS = root-mean-square; iSNR = intelligibility-weighted signal-to-noise ratio; FADE = framework for auditory discrimination experiment.

⁴The data of the iSNR-based predictions were taken from Baumgärtel et al. (2015).

**Intrusive Classification and Objectivity**

Völker et al. (2015) modeled the same empirical data with the BSIM (Beutelmann et al., 2010) and found that the difficult—but required for BSIM predictions—post hoc separation of the processed signal into a speech and noise part was a major problem for those conditions where nonlinear signal processing algorithms were used.
interpretation of prediction errors difficult. To the best of our knowledge, no other model with nonintrusive classification that allows the objective, that is, empirical reference-free, prediction of SRTs in aided conditions with fluctuating noise maskers exists at this time of writing. Nonintrusive classification. FADE requires only the same information as needed to perform the matrix sentence test with human listeners. This includes the ability to generate mixtures at defined SNRs that normally requires the separate availability of the clean matrix sentences and the noise signal. However, it does not give the recognizer (here, the ASR system) access to the separate signals. This closely resembles experiments with human listeners, where only noisy signals are presented to determine the outcome. Hence, FADE is a modeling approach with nonintrusive classification. In contrast to many other models (cf. Table 1), FADE works naturally on mixed signals, which allows to apply arbitrary signal processing to the mixed signals prior to presenting them to the recognition system (just like in the empirical measurements). The employed algorithm can be treated as a black box to which no access—apart from the signal input and the signal output—is needed.

Objective. FADE directly predicts the SRT as the primary outcome of the matrix sentence test employed. Index-based models predict an index value between 0 and 1 for each SNR. Assuming a strictly monotonic relationship between SNR and index, which is not always true, each index value can be mapped to an SNR. Very often, a reference value for the index that corresponds to the SRT, that is, the SNR at 50% correct, is determined by using the empirical data from one of the conditions that the model aims to explain. One implication of that procedure is that no single condition can be predicted. Another—more far-reaching—implication is that the predictions then depend on the reference condition and hence on (a) the complexity of the speech material, (b) the language, (c) the talker, (d) the noise, (e) the reverberation of the reference condition, and probably many more parameters whose influence is not predicted by the model. As a consequence, these SRT predictions are not strictly objective because they strongly depend on empirical measurements. In contrast, FADE can perform SRT predictions for single conditions solely based on fairly basic principles. For the predicted SRT to match the empirically determined one, it must correctly predict the joint effect of all relevant factors, for example, the complexity of the speech material, the language, the talker, the noise, possible binaural interactions, and of the signal processing. In doing so without using any empirical reference SRT data, it provides a truly objective measure.

Predicted Outcomes

Predicted outcomes of the German matrix sentence test were compared with empirical outcomes, whose differences are the prediction errors. An interpretation of the prediction error is a prerequisite for well-founded speculations on their causes, and hence on the causes of potential shortcomings of the model. Objective SRT predictions in combination with nonintrusive classification are presumably helpful for deriving a meaningful interpretation of the prediction errors. For example, if FADE cannot explain some aspect, the recognition performance gets worse and the predicted SRTs increase, hence underestimating the empirical recognition performance. This performance limitation is similar to most ASR systems that perform worse than listeners with normal hearing when confronted with situations for which they were not designed. Hence, a significant overestimation of the empirical performance (not observed here) indicates that the model was (a) provided with more information or (b) made better use of the information that is required to perform the corresponding recognition task. A significant underestimation of the empirical performance of listeners with normal hearing, on the other hand, indicates a suboptimal—or missing—implementation of at least one of the required principles to perform the corresponding recognition task. A small underestimation of the median empirical performance was observed in Section “Predicted SRTs” for the 20T and, slightly more pronounced, in the CAN...
condition. This could be due to the simplistic binaural better-ear strategy adopted by the current approach that might not be able to separate the target speaker from the noises as efficiently as listener with normal binaural hearing. The interaction of the spatial scene with binaural signal processing and binaural (impaired) hearing could also be the cause of the variability in the deviation of predicted median SRTs for listeners with impaired hearing in the 20T condition. A more sophisticated, true binaural processing stage with a direct across-ear processing of the input signal (e.g., Spille, Kollmeier, & Meyer, 2017) might improve the model behavior here.

A larger systematic underestimation of the recognition performance was observed in Section “Predicted SRTs” for the SIT condition, which resulted in large positive biases in Table 3 for the performance of listeners with normal hearing in all conditions. The considerable bias of 13 dB indicates that the model had a severe problem with voiced speech maskers. The system was neither able to learn nor provided with information about which speech portion is considered as target and which has to be considered as nonattended interferer. Human listeners achieve very low SRTs (<−20 dB) primarily by being able to listen in the dips of the masking speech. Probably, human listeners are better able to segregate the talkers based on their pitch and spatial cues than the model, an ability that is often referred to as the cocktail-party effect. The necessary signal processing to perform such a task by machines has been studied under the term computational acoustic scene analysis (CASA) and is largely missing in current ASR systems as well as in FADE. Emerging technical solutions to this problem could potentially improve ASR systems and lead to improved predictions in the SIT condition.

Because the predicted SRTs in the SIT condition were not found to be in agreement with the empirical data for normal-hearing listeners, no individualization, that is, prediction for listeners with impaired hearing was performed. In the 20T and CAN condition, the predictions for the group of listeners with impaired hearing showed the same bias (±0.2 dB) and hence higher SRTs than for the group of listeners with normal hearing (cf. Table 3) in agreement with the empirical data. According to the empirical data, some listeners with impaired hearing outperformed the average listener with normal hearing, by up to ≈2 dB. However, the individual predicted SRTs were not found to be more than 0.5 dB lower than the corresponding predictions without individualization, that is, for listeners with normal hearing. This can be partly observed in Figure 4 comparing the lowest individual predictions for listeners with impaired hearing (colored stars) to the lowest prediction for listeners with normal hearing (black squares) and also holds for the other algorithms (not shown). An explanation might be the inevitable uncertainty or statistical measurement error in the individual empirical data that could sum up to ≈ 2 dB depending in the noise condition. This expected error in the empirical data is due to small training effects, limited and varying attention of the listeners, and the small number of presented sentences (at most, 20) on which the SRT calculation is based. It should be noted that this statistical error is not—or to a much lesser extent—found in the predictions because SRTs are calculated based on results from all 120 sentences.

Apart from the already discussed bias, which also dominated the RMS prediction errors, the predicted group SRTs correlated significantly with the empirical data, explaining between 63% and 92% of the variance across algorithms in the 20T and CAN condition, respectively. The individual predictions for listeners with impaired hearing did not explain more that 16% of the variance across individuals and algorithms. These results already hint at good group predictions in general, but unsatisfactory individual predictions for listeners with impaired hearing. However, individual offsets originating from nonacoustical causes which are not modeled here (e.g., central neurosensory component of the HL or varying cognitive abilities) may affect all algorithms alike. They may cause a spread of the individual empirical SRTs and hence result in a lower correlation between the predicted and empirical data. To test this hypothesis and remove an individual bias, the benefits in SRT with respect to the reference condition (NoPre) were calculated individually for all listeners (to be discussed later).

Objective Instrumental Evaluation of Algorithms

The high prediction accuracy (RMS prediction error ≤1 dB) and the low absolute biases (<1 dB) of the benefits in SRT for the group of listeners with normal hearing (cf. Figures 5 and 6, and Table 5, respectively) indicate a solid basis for individualized predictions. Further, the predicted benefits for listeners with normal hearing do not depend on any empirical data, which is why FADE provided a near-optimal objective evaluation in this regard. Notably, in the realistic CAN condition, the model correctly predicted the fixed MVDR to result in the highest benefit and the SCNR in the lowest, where the benefit of the latter was correctly predicted to be negative, that is, it increased the SRT. While the ADM was predicted to perform worse than the fixed MVDR in the artificial 20T condition, which was not observed empirically, the negative benefit of the SCNR was still correctly predicted. The differentiation between improvement and deterioration is very valuable to algorithm developers and users alike.

In the instrumental evaluation of the algorithms from Baumgärtel et al. (2015), the iSNR, STOI, and
perceptual evaluation of speech quality (PESQ; ITU-T Recommendation, 2001), which can be considered representative for the methods that are used to objectively assess the performance of (speech) signal processing algorithms, were used to predict the benefit. None of them predicted the negative effect of the SCNR algorithm in the artificial 20T condition and the realistic CAN condition for listeners with normal hearing. There, benefits in SNR were only predicted with the iSNR because the output of the other measures were indexes that could not be mapped to SNRs or benefits in SNR due to the lack of empirical reference data in that study. This reflects a common, fundamental problem of reference-based models. For objective predictions of the benefit with the iSNR, a working point, that is, an input SNR at which the benefit was predicted, had to be chosen independently from the empirical data, which was 0 dB. If the empirical working point is unknown, which usually is the case when objective measures are used, a guess of 0 dB seems still reasonable, particularly if the predictions should also represent the behavior of listeners with impaired hearing or cochlear implants. These limitations, that is, the requirement of guessing or knowing the outcome prior to any prediction, clearly shows one advantage of FADE that can predict an SRT without any empirical reference. The statistical analysis of the predicted outcomes with the iSNR from Baumgärtel et al. (2015) for the group of normal-hearing listeners in the 20T and CAN condition is shown in Table 5. For the artificial 20T condition, the iSNR predictions explained about 60% of the variance in empirical data and the RMS prediction error exceeded 3 dB. For the realistic CAN condition, the ISNR predictions were not significantly correlated with the empirical data and the RMS prediction error also exceeded 3 dB. Compared with these results, the FADE predictions of benefits in SRT for normal-hearing listeners provided a much more reliable objective instrumental measure with RMS prediction errors of 1 dB and less in the respective 20T and CAN conditions.

The mediocre performance of the iSNR, STOI, and PESQ could be partly due to the method (phase inversion according to Hagerman & Olofsson, 2004) that was used to separate the speech from the noise signal after processing and that might not achieve a perfect separation. This seems probable because no true separate processed signals exist whereby the goodness of the estimation cannot be assessed. The phase inversion could be possibly replaced by a more sophisticated separation method, such as shadow filtering. However, even if this would lead to better predictions, it would not overcome the fundamental limitation that this kind of separation of the mixed signals can lead to uncontrolled biases, which are typically concealed by a calibration to empirical reference data. Because the classification with FADE is nonintrusive and works on mixed signals, predictions can be performed with virtually any signal preprocessing without additional modifications to the model. It seems worthwhile to further validate the suitability of FADE for the objective evaluation of algorithms on a broader basis of applications.

**Individual Aided Performance Predictions**

Individualization was performed using the audiograms, that is, the individual hearing thresholds of both ears, where listeners with normal hearing were assumed to have no HL. This is also why there were individual measurements from listeners with normal hearing but no individual predictions for them. The individualization for listeners with impaired hearing was reflected in the personalized HA setting: Here, it had consequences for the fitting of the multiband compressor of the MHA, where the HL was (partially) compensated (cf. Section “Binaural Preprocessing Algorithms”) and the feature extraction of FADE, where the individual hearing threshold was simulated (cf. Section “Model of Impaired Hearing”). Therefore, assuming an appropriate compensation for the main component of the moderate HIs encountered in our rather homogeneous group of listeners with impaired hearing, no strong individual deviation from the compensated main component of the HL should be expected.

However, despite similar audiograms and individually compensated HIs, the empirical results show (a) that the variance of the individual benefits in SRT were in the same order of magnitude as the effect size of the algorithms (cf. Figure 5) and (b) that the median SRTs for the group of listeners with impaired hearing were about 1 to 2 dB higher than for the group of listeners with normal hearing. The individual variance of the listeners with impaired hearing could not be explained by the model considering individual absolute hearing thresholds, where at most 23% of the variance could be explained with a considerable RMS prediction error of 2.4 dB (cf. “HI indiv.” predictions in the CAN condition in Table 5). However, the general trend, that the SRTs for the group of listeners with impaired hearing were higher than those for listeners with normal hearing, was predicted in accordance with the empirical data (cf. Table 2). Also, the predictions of benefits in SRT for the group of listeners with impaired hearing (median of the corresponding benefits in SNR) explained about 75% of the variance in the empirical data with RMS prediction error of 1.5 dB and 0.9 dB in the 20T and CAN condition, respectively.

A hypothesis is that the listeners with impaired hearing are not sufficiently characterized by the audiogram, as already suggested by Kollmeier et al. (2015). According to Kollmeier et al. (2015), a (monaural)
suprathreshold distortion component was important to accurately model the individual effect of HL on speech recognition tasks, and according to Plomp (1978), such a distortion component would affect the individual SRTs similarly in different conditions. This behavior could be observed in Figure 8, where the prediction errors in the 20T and CAN condition were plotted against each other and showed a significant and strong correlation \( R^2 = .69, p < .01 \). Any effect that was not explained by the model is accumulated in the prediction error, that is, deviations due to not considering a suprathreshold distortion component, missing binaural interaction, or neglected cognitive effects, as well as any measuring inaccuracies (apart from the audiograms). Errors in the measurement of the audiograms would affect the predictions in all conditions similarly, while errors in the measurement of SRTs can be assumed to be independent and hence would not lead to systematic errors. One possible explanation for the strong correlation of prediction error across conditions could be systematic biases due to errors in the measurement of the audiograms. Another possibility is that there could be another individual factor—apart from the audiogram—missing to accurately predict individual benefits in SRT with FADE. Such an uncontrolled individual distortion component could be responsible for up to 4 dB (cf. point spread in Figure 8), which would translate to an average group effect of up to approximately 2 dB. This is in agreement with the average differences in SRT between the group of listeners with impaired hearing and the group of listeners with normal hearing, which were empirically found to be 1.8 dB in the 20T and 1.3 dB in the CAN condition, and predicted to be 1.7 dB and 1.2 dB, respectively.

In its current form, FADE was able to accurately predict the benefit of algorithms for the groups of listeners with normal hearing and impaired hearing. Individual (aided) performance, however, could not be predicted based on an individualization with the audiogram alone. In future work, a possible individual distortion component, for example, as suggested by Kollmeier et al. (2015) in the form of a level uncertainty, should be considered to possibly enable accurate individual predictions of the aided performance of listeners with impaired hearing.

Another (related) explanation for the low correlation of the individual predictions could be an insufficient consideration of the (possibly even individual impaired) binaural interaction. The current approach considers better ear listening that does not include true binaural interactions across the signals presented to both ears. An individual component regarding the efficiency of the binaural processing would have the same effect as the (monaural) suprathreshold distortion component discussed before. In the end, a combination of both seems likely.

**Limitations and Opportunities of the Approach**

The FADE modeling approach has been extensively evaluated with the matrix sentence test—mainly because of its very good test–retest reliability—but still not with other speech tests. Because ASR systems are trained for each prediction on the corresponding speech material, there are no fundamental reasons why the same approach should not work with other speech material or vocabulary. In fact, it was already shown that the very same model predicts the outcome of the matrix sentence test in other languages (Schädler et al. 2016a) in different noise conditions without additional assumptions. However, larger vocabularies and more complex grammatical structures will probably increase the already high computational demands; currently, about 5 min per prediction can be achieved on high-end desktop PCs. It should be investigated in the future if the approach generalizes in this regard, that is, if it is able to predict the effect of the complexity of a speech test.

Also, these properties allow the interpretation of the prediction error. This might help to better understand which cues are missing in the feature vector for different kinds of tasks in comparison with full-blown complex ASR experiments where single factors cannot be isolated as easily, let alone the required resources to perform these experiments. A natural evolution of FADE could be the stepwise integration of methods/models from research on ASR, CASA, psychoacoustics, and even human physiology. Such extensions might be valid as long as the predicted SRTs for listeners with normal hearing do not increase. This appears to be a simple criterion that does not depend on empirical data and preserves the objectivity of FADE while being compatible with research on robust ASR.

If objectivity and interpretability of prediction errors are of no concern, FADE can be turned into a reference-based model by calibrating the predictions with empirical reference data. For example, if SRTs are empirically found to be 4 dB higher for a more complex speech test than for the matrix sentence test in a reference condition, the reference-free predictions for the matrix sentence test can be increased by 4 dB to predict the outcome of the more complex speech test. This approach is very similar to the calibration of reference-based models. It would allow performing reference-based predictions with FADE for data sets which otherwise could only be predicted with other reference-based models. Evaluations of predictions with FADE for other speech tests should consider testing the validity of a constant-offset assumption.

For individual predictions, the SRTs are allowed to increase depending on (dys-)functional parameters such as the audiogram or putative suprathreshold processing deficiencies. It is open for further research, which modification of the standard signal preprocessing and feature...
The predictions with FADE showed a general trend to result in a lack of across-frequency integration. This was attributed to the assumption of independent frequency bands and the resulting lack of across-frequency integration.

Conclusions

The most important findings of this work can be summarized as follows:

- FADE explained about 90% of the variance of the empirical speech recognition threshold (SRT) data resulting from eight different binaural preprocessing strategies in a realistic cafeteria environment for listeners with normal hearing. It predicted the corresponding benefits in SRT with an RMS error of only 0.6 dB without requiring any a priori knowledge of empirical data to perform those predictions. This result makes FADE a promising candidate for the objective evaluation of complex (speech) signal processing algorithms.

- A voiced speech masker (competing talker) at very low SRTs (< −20 dB in SNR) posed a problem to the model in its current form as it probably would to any ASR system. This led to pronounced underestimation of the empirically determined recognition performance. Further improvements in robust ASR features and the combination with CASA might help to overcome this problem in modeling.

- Median predictions across the group of listeners with impaired hearing—where individual factors average out—explained about 75% of the variance in the empirical data. Basing individual SRT benefit predictions only on audiograms for listeners with impaired hearing explained only an unsatisfactory small fraction (< 25%) of the variance in the empirical data. The high correlation of prediction errors between different noise conditions across listeners indicate that at least one further individual factor, such as a suprathreshold processing deficiency, should be accounted for to improve the accuracy of individual predictions with impaired hearing.

- The predictions with FADE showed a general trend to underestimate the empirical recognition performance. This bias indicates that the feature extraction stage and the simple binaural better ear listening strategy as well as the ASR model employed may not be optimal. This contrasts with the inherent assumption of FADE to estimate the lowest achievable SRT and shows a limitation of the current approach. However, the existence of the bias is due to two desired properties of FADE: (a) Its classification is nonintrusive, which prevents trivial manipulations of the bias towards lower SRTs using superhuman knowledge, and (b) its objectivity allows direct predictions of the outcomes and avoids the bias-removing necessity of a calibration to a reference condition.

Acknowledgments

The authors would like to thank Giso Grimm and Tobias Herzke for their support with the Master Hearing Aid.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Cluster of Excellence EXC 1077 “Hearing4all.”

Notes

1. URL: http://medi.uni-oldenburg.de/FADE
2. URL: http://medi.uni-oldenburg.de/SGBFB
3. Current x86-64 8-core CPU, 16+Gb of RAM, fast solid-state disk.

References

ANSI 1997. (1997). Methods for calculation of the speech intelligibility index. Washington, DC: Author.

Baumgärtel, R. M., Krawczyk-Becker, M., Marquardt, D., Völker, C., Hu, H., Herzke, T., ... Dietz, M. (2015). Comparing Binaural Pre-processing Strategies I: Instrumental Evaluation. Trends in Hearing, 19. doi: 10.1177/2331216515617916.

Beutelmann, R., Brand, T., & Kollmeier, B. (2010). Revision, extension, and evaluation of a binaural speech intelligibility model. The Journal of the Acoustical Society of America, 127(4), 2479–2497. doi: 10.1121/1.3295575.

Bibiger, T., & Ewert, S. D. (2017). The role of short-time intensity and envelope power for speech intelligibility and psychoacoustic masking. The Journal of the Acoustical Society of America, 142(2), 1098–1111. doi: 10.1121/1.4999059.

Bisgaard, N., Vlaming, M. S., & Dahlquist, M. (2010). Standard audiograms for the IEC 60118-15 measurement procedure. Trends in Amplification, 14(2), 113–120. doi: 10.1177/1084713810379609.

Byrne, D., Dillon, H., Ching, T., Katsch, R., & Keidser, G. (2001). NAL-NL1 procedure for fitting nonlinear hearing aids: Characteristics and comparisons with other procedures. Journal-American Academy of Audiology, 12(1), 37–51.

Cooke, M. (2006). A glimpsing model of speech perception in noise. The Journal of the Acoustical Society of America, 119(3), 1553–1573. doi: 10.1121/1.2166600.
Dau, T., Püschel, D., & Kohlrausch, A. (1996). A quantitative model of the “effective” signal processing in the auditory system. II. Simulations and measurements. *The Journal of the Acoustical Society of America*, 99(6), 3623–3631. doi: 10.1121/1.414960.

Elko, G. W., & Pong, A.-T. N. (1995). A simple adaptive first-order differential microphone. In 1995 IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics (pp. 169–172). IEEE. doi: 10.1109/ASPAA.1995.482983.

Falk, T. H., Cosentino, S., Santos, J., Suelzle, D., & Parsa, V. (2013). Non-intrusive objective speech quality and intelligibility prediction for hearing instruments in complex listening environments. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7820–7824). IEEE. doi: 10.1109/ICASSP.2013.6639186.

Fisher, R. (1958). *Statistical methods for research workers* (3rd ed.). Edinburgh, Scotland: Oliver and Boyd. doi: 10.1007/978-1-4612-4380-9_6.

Gerkmann, T., Breithaupt, C., & Martin, R. (2008). Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(5), 910–919. doi: 10.1109/TASLP.2008.921764.

Greenberg, J., Peterson, P., & Zurek, P. (1993). Intelligibility-weighted measures of speech-to-interference ratio and speech system performance. *The Journal of the Acoustical Society of America*, 94(5), 3009–3010. doi: 10.1121/1.407334.

Griffiths, L., & Jim, C. (1982). An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*, 30(1), 27–34. doi: 10.1109/TAP.1982.1142739.

Grimm, G., Herzke, T., Berg, D., & Hohmann, V. (2006). The master hearing aid: A PC-based platform for algorithm development and evaluation. *Acta Acustica United with Acustica*, 92(4), 618–628.

Grimm, G., Hohmann, V., & Kollmeier, B. (2009). Increase and subjective evaluation of feedback stability in hearing aids by a binaural coherence-based noise reduction scheme. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(7), 1408–1419. doi: 10.1109/TASL.2009.2020531.

Hagerman, B., & Olofsson, Å. (2004). A method to measure the effect of noise reduction algorithms using simultaneous speech and noise. *Acta Acustica United with Acustica*, 90(2), 356–361.

Holube, I., & Kollmeier, B. (1996). Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. *The Journal of the Acoustical Society of America*, 100(3), 1703–1716. doi: 10.1121/1.417354.

Holube, I., Wesselkamp, M., Dreschler, W. A., & Kollmeier, B. (1997). Speech intelligibility prediction in hearing-impaired listeners for steady and fluctuating noise. In: *Modeling sensorineural hearing loss* (pp. 447–459). Mahwah, NJ: Lawrence Erlbaum Associates.

IEC. (1998). *Sound system equipment - Part 16: Objective rating of speech intelligibility by speech transmission index*. Geneva, Switzerland: IEC.

ITU-T Recommendation. (2001). *P. 862. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*. Geneva, Switzerland: ITU-T.

Jensen, J., & Taal, C. H. (2016). An algorithm for predicting the intelligibility of speech masked by modulated noise masks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11), 2009–2022. doi: 10.1109/TASLP.2016.2585878.

Jørgensen, S., Ewert, S. D., & Dau, T. (2013). A multi-resolution envelope-power based model for speech intelligibility. *The Journal of the Acoustical Society of America*, 134(1), 436–446. doi: 10.1121/1.4807563.

Jürgens, T., & Brand, T. (2009). Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model. *The Journal of the Acoustical Society of America*, 126(5), 2635–2648. doi: 10.1121/1.3224721.

Kates, J., & Arehart, K. (2005). Coherence and the speech intelligibility index. *The Journal of the Acoustical Society of America*, 117, 2224–2237. doi: 10.1121/1.1865275.

Kates, J., & Arehart, K. (2014). The hearing-aid speech perception index (HASPI). *Speech Communication*, 65, 75–93. doi: 10.1016/j.specom.2014.06.002.

Kayser, H., Ewert, S. D., Anemüller, J., Rohdenburg, T., Hohmann, V., & Kollmeier, B. (2009). Database of multi-channel in-ear and behind-the-ear head-related and binaural room impulse responses. *EURASIP Journal on Advances in Signal Processing*, 2009(1), 1–10. doi: 10.1155/2009/298605.

Kollmeier, B., & Kiessling, J. (2016). Functionality of hearing aids: State-of-the-art and future model-based solutions. *International Journal of Audiology*. Advance online publication. doi: 10.1080/14992027.2016.1256504.

Kollmeier, B., Schäddler, M. R., Warzybok, A., Meyer, B. T., & Brand, T. (2015). Individual speech recognition in noise, the audiogram and more: Using automatic speech recognition (ASR) as a modelling tool. *Proceedings of the International Symposium on Auditory and Audiological Research*, 5, 149–156.

Kollmeier, B., Schäddler, M. R., Warzybok, A., Meyer, B. T., & Brand, T. (2016). Sentence recognition prediction for hearing-impaired listeners in stationary and fluctuation noise with FADE. *Trends in Hearing*, 20, 2331216516655795.

Plomp, R. (1978). Auditory handicap of hearing impairment and the limited benefit of hearing aids. *The Journal of the Acoustical Society of America*, 63(2), 533–549. doi: 10.1121/1.381753.

Rhebergen, K. S., Versfeld, N. J., de Laat, J. A., & Dreschler, W. A. (2010). Modelling the speech reception threshold in non-stationary noise in hearing-impaired listeners as a function of level. *International Journal of Audiology*, 49(11), 856–865. doi: 10.3109/14992027.2010.498446.

Schäddler, M. R. (2016). Robust automatic speech recognition and modeling of auditory discrimination experiments with auditory spectro-temporal features. Oldenburg, Germany: Bis-Verlag.

Schäddler, M. R., Hülsmeier, D., Warzybok, A., Hochmuth, S., & Kollmeier, B. (2016a). Microscopic multilingual matrix test predictions using an ASR-based speech recognition model. *Proceedings of Interspeech 2016*, 610–614. doi: 10.21437/Interspeech.2016-1119.
Schärdler, M. R., & Kollmeier, B. (2015). 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*(4), 2047–2059. doi: 10.1121/1.4916618.

Schärdler, M. R., Warzybok, A., Ewert, S. D., & Kollmeier, B. (2016b). A simulation framework for auditory discrimination experiments: Revealing the importance of across-frequency processing in speech perception. *The Journal of the Acoustical Society of America, 139*(5), 2708–2722. doi: 10.1121/1.4948772.

Schärdler, M. R., Warzybok, A., Hochmuth, S., & Kollmeier, B. (2015). Matrix sentence intelligibility prediction using an automatic speech recognition system. *International Journal of Audiology, 54*(sup2), 100–107. doi: 10.3109/14992027.2015.1061708.

Simmer, K. U., Bitzer, J., & Marro, C. (2001). Post-filtering techniques. In M. Brandstein, & D. Ward (Eds.), *Microphone Arrays. Digital Signal Processing* (pp. 39–60). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-662-04619-7_3.

Spille, C., Kollmeier, B., & Meyer, B. T. (2017). Combining binaural and cortical features for robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25*(4), 756–767. doi: 10.1109/TASLP.2017.2661712.

Stadler, S., Leijon, A., & Hagerman, B. (2007). An information theoretic approach to predict speech intelligibility for listeners with normal and impaired hearing. In *Proceedings of Interspeech 2007* (pp. 1345–1348). ISCA.

Steeneken, H., & Houtgast, T. (1980). A physical method for measuring speech transmission quality. *Journal of the Acoustical Society of America (JASA), 67*, 318–326. doi: 10.1121/1.384464.

Steeneken, H., & Houtgast, T. (2002). Basics of the STI-measuring method. In van Wijngaarden, S. J (Ed.), *Past, Present, and Future of the Speech Transmission Index, International Symposium on STI, TNO Human Factors, ZG Soesterberg, The Netherlands* (pp. 13–44).

Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2011). An evaluation of objective measures for intelligibility prediction of time-frequency weighted noisy speech. *The Journal of the Acoustical Society of America, 130*(5), 3013–3027. doi:10.1121/1.3641373.

Van Veen, B. D., & Buckley, K. M. (1988). Beamforming: A versatile approach to spatial filtering. *IEEE ASSP Magazine, 5*(2), 4–24. doi: 10.1109/53.665.

Völker, C., Warzybok, A., & Ernst, S. M. (2015). Comparing binaural pre-processing strategies III: Speech intelligibility of normal-hearing and hearing-impaired listeners. *Trends in Hearing, 19*, 2331216515618609.