A model of speech recognition for hearing-impaired listeners based on deep learning
Jana Roßbach, Birger Kollmeier and Bernd T. Meyer

Citation: The Journal of the Acoustical Society of America 151, 1417 (2022); doi: 10.1121/10.0009411
View online: https://doi.org/10.1121/10.0009411
View Table of Contents: https://asa.scitation.org/toc/jas/151/3
Published by the Acoustical Society of America

ARTICLES YOU MAY BE INTERESTED IN

Toward a better understanding of nonoccupational sound exposures and associated health impacts: Methods of the Apple Hearing Study
The Journal of the Acoustical Society of America 151, 1476 (2022); https://doi.org/10.1121/10.0009620

Measurement and modeling of the mechanical impedance of human mastoid and condyle
The Journal of the Acoustical Society of America 151, 1434 (2022); https://doi.org/10.1121/10.0009618

REVIEWS OF ACOUSTICAL PATENTS
The Journal of the Acoustical Society of America 151, 1428 (2022); https://doi.org/10.1121/10.0009653

The sound of one frog calling: The bullfrog's reactions to acoustic stimuli
The Journal of the Acoustical Society of America 151, R5 (2022); https://doi.org/10.1121/10.0009652

Vocal and semantic cues for the segregation of long concurrent speech stimuli in diotic and dichotic listening—The Long-SWoRD test
The Journal of the Acoustical Society of America 151, 1557 (2022); https://doi.org/10.1121/10.0007225

Wave propagation across the skull under bone conduction: Dependence on coupling methods
The Journal of the Acoustical Society of America 151, 1593 (2022); https://doi.org/10.1121/10.0009676
A model of speech recognition for hearing-impaired listeners based on deep learning

Jana Roßbach,1,a) Birger Kollmeier,2 and Bernd T. Meyer1

1Communication Acoustics and Cluster of Excellence Hearing4all, Carl von Ossietzky University, D-26111 Oldenburg, Germany
2Medical Physics and Cluster of Excellence Hearing4all, Carl von Ossietzky University, D-26111 Oldenburg, Germany

ABSTRACT:
Automatic speech recognition (ASR) has made major progress based on deep machine learning, which motivated the use of deep neural networks (DNNs) as perception models and specifically to predict human speech recognition (HSR). This study investigates if a modeling approach based on a DNN that serves as phoneme classifier [Spille, Ewert, Kollmeier, and Meyer (2018), Comput. Speech Lang. 48, 51–66] can predict HSR for subjects with different degrees of hearing loss when listening to speech embedded in different complex noises. The eight noise signals range from simple stationary noise to a single competing talker and are added to matrix sentences, which are presented to 20 hearing-impaired (HI) listeners (categorized into three groups with different types of age-related hearing loss) to measure their speech recognition threshold (SRT), i.e., the signal-to-noise ratio with 50% word recognition rate. These are compared to responses obtained from the ASR-based model using degraded feature representations that take into account the individual hearing loss of the participants captured by a pure-tone audiogram. Additionally, SRTs obtained from eight normal-hearing (NH) listeners are analyzed. For NH subjects and three groups of HI listeners, the average SRT prediction error is below 2 dB, which is lower than the errors of the baseline models.

(Received 23 February 2021; revised 27 December 2021; accepted 18 January 2022; published online 1 March 2022) [Editor: Sarah Verhulst] Pages: 1417–1427

I. INTRODUCTION
Communication plays a crucial role in our daily lives, and both speech production and recognition are at the core of our social interaction. In real-life conditions, human speech recognition (HSR) is often negatively affected by the influence of background noise, such as traffic noise or interfering speakers in multi-talker scenarios, by reverberation, or by impaired hearing.

Audio signal processing strategies such as hearing aid algorithms can potentially improve HSR. The gold standard for the evaluation of such algorithms is to perform listening experiments, which often have the aim to determine the speech recognition threshold (SRT), i.e., the signal-to-noise ratio (SNR) at which a specific percentage (commonly 50%) of words are recognized. Such tests are, however, time- and cost-intensive and not applicable to optimize algorithms by covering a high-dimensional parameter space. Hence, several model approaches have been proposed for predicting HSR based on an acoustic input.

In this study, we explore an HSR model based on deep machine learning and first provide a short overview of several established models, some of which are later compared to our experimental results: The speech intelligibility index (SII) (ANSI, 1997) is based on the articulation index (AI) (ANSI, 1969) and predicts a value between 0 (low HSR) and 1 (maximum HSR). This value is obtained by calculating the weighted sum of SNRs in different frequency bands. To predict the SRT, the value has to be mapped to the results of reference measurements with human listeners. A limitation of this approach is that it is not applicable for predicting HSR on short time scales because the SII uses the long-term spectrum of the signals. To take these modulations in time and frequency domain into account, the extended SII (ESII) (Rhebergen and Versfeld, 2005; Rhebergen et al., 2006) was developed and compared to the SII (Meyer and Brand, 2013). Instead of using the long-term spectrum of the speech signal and the noise, the signals are analyzed on short time scales using a weighted sum of SNRs. Another model for HSR prediction is the short-time objective intelligibility measure (STOI) (Taal et al., 2011). Its prediction is based on the cross correlation between the envelope of the clean speech and the envelope of the speech mixed with noise, assuming that a high correlation results in high HSR. To account for the spectral dependencies of speech, the extended STOI (ESTOI) (Jensen and Taal, 2016) also includes a spectral correlation between noisy and clean speech. Another extension of STOI is the non-intrusive STOI (NI-STOI) (Andersen et al., 2017), which estimates the envelope of the clean speech from the noisy signal. The prediction accuracy of NI-STOI is lower than of the intrusive STOI, but higher than the accuracy of other non-intrusive models.

Several established models have been proposed that have been derived from the power spectrum model (PSM)
The PSM is based on the assumption that the human auditory system contains a bank of auditory filters (Fletcher, 1940) where the filter with the highest long-term SNR is used to model tone-in-noise detection. Jørgensen and Dau (2011) developed the speech-based EPSM (sEPSM) which integrates an alternative forced choice decision model to define the relationship between the envelope SNR ($SNR_{env}$) and the HSR. To account for temporal fluctuations, a short-term analysis was added to the sEPSM. The resulting model is referred to as multi-resolution sEPSM (mr-sEPSM) (Jørgensen et al., 2013). The mr-sEPSM produces an absolute estimate of the HSR, which contrasts to the previously mentioned models that predict a relative change of HSR in comparison to a reference condition. Biberger and Ewert (2016) combined two models to exploit complementary information from the original PSM and the mr-sEPSM by considering the local amplitude modulation depth, which is referred to as the multi-resolution generalized power spectrum model (mrGPSM). In this model, the $SNR_{env}$ of the mr-sEPSM is logarithmically weighted and normalized with the local SNR that is calculated with the PSM.

The hearing-aid speech perception index (HASPI) (Kates and Arehart, 2014) can predict the HSR for normal-hearing (NH) and hearing-impaired (HI) listeners. HASPI is based on an auditory periphery model that returns the envelope and the temporal fine structure of the input signal (Kates, 2013). The speech signal is mixed with the noise and used as input for the HI auditory model that accounts for the individual hearing loss. To calculate the HSR, the results are compared with the outputs of an NH reference model that receives the clean speech as input signal.

The aforementioned models require a priori knowledge for HSR prediction in terms of the access to the clean speech signal and to the masking noise (SII, ESII), to clean and noisy speech (STOI, ESTOI) or—in the case of mr-sEPSM and mrGPSM—the masking noise and the masked signal. For predictions with HASPI, clean speech and noisy speech must be available separately. In this sense, the models require a speech reference and are referred to as intrusive. For mr-EPsM as well as for mrGPSM, information about the type of speech material (e.g., meaningful or meaningless sentences, single words, or only syllables) are additionally required because they are using a task-specific decision state. Intrusive models are valuable for measuring the benefit of hearing aid algorithms, but the requirement of having separate speech and noise signals is an important limitation of such models since it restricts their use to well-controlled stimuli that are typically only available in lab conditions.

In the last two decades, several models have been developed that combine knowledge of the auditory system with methods of automatic speech recognition (ASR). Cooke (2006) proposed a glimpsing model that estimates intelligibility based on spectrogram patches which are at least 3 dB above the noise floor, that serve as input for a Gaussian mixture model (GMM) combined with a hidden Markov model (HMM). The HSR predictions were produced by comparing the recognized speech with the true speech label and the glimpsing detection component requires a separate input of speech and noise.

Jürgens and Brand (2009) extended a model originally proposed by Holube and Kollmeier (1996), which combines the output of an auditory perception model (Dau et al., 1997) with a classification stage based on dynamic time warping (DTW) (Sakoe and Chiba, 1978). This model was able to predict the speech recognition of NH listeners with a high accuracy for non-fluctuating noises by using the identical speech signals for training and testing (a so-called optimal detector). Schädler et al. (2016) introduced a GMM-HMM-based approach as speech intelligibility model, which is referred to as framework for auditory discrimination experiment (FADE). The model produces accurate predictions for a large set of psychoacoustic measurements and is also able to predict the benefit of hearing aids (Schädler et al., 2018). However, since the training data contains some of the same speech signals later seen in testing, this approach is also based on an optimal detector.

An ASR-based model approach using GMMs and HMMs with separate speech for training/testing was introduced by Fontan et al. (2020). The authors modeled speech intelligibility for older, hearing-impaired listeners in quiet and observed high correlations for the predictions of the individual listeners with different word material. However, the results also show large root mean squared errors (RMSE) between measurement and prediction. Karbasi et al. (2020) combined an ASR-based system with different signal-based speech intelligibility measures. These measures were mapped to the results of listening experiments to predict the HSR. The mapping was done with binary classification neural networks, trained with the respective intelligibility measure labeled with the results of the listening tests. Spille et al. (2018) proposed a model for HSR prediction based on ASR that includes a deep neural network (DNN) for calculating phoneme probabilities. The word error rate (WER) was used to determine the SRT of the ASR-based model using sentences with a simple structure embedded in different noise signals, which range from stationary noise to a single background talker. The approach produced rather low average RMSE for SRT predictions (between 1.7 and 1.9 dB). However, the model was only evaluated with data obtained from NH listeners and did not take hearing impairment into account. Although this approach used data from different speakers for training and testing, it used identical, short masking signals for the noisy training and test sets, which bears the risk of overfitting since the model could learn the noise signals, which constitutes an indirect noise reference.

To explore if the ASR-based model approach is suitable to predict HSR for HI listeners, this study investigates if individual hearing impairment can be integrated into the DNN-based model by exploiting information about the frequency-specific hearing threshold of the better ear at the level of ASR features. To this end, HSR predictions
obtained with individual feature modifications are compared to SRTs of listeners with a very mild, mild, or moderate hearing impairment. The speech-in-noise recognition experiments are conducted using the same diverse noise types for speech masking as in the study of Spille et al. (2018). In addition to the SRTs, the slopes of the psychometric functions are also predicted, and the correlation between measurements and model output is calculated across all noise types and compared to five baseline models (SII, ESII, STOI, mr-sEPSM, and HASPI). We refer to this model as deep ASR for intelligibility (DAILY) prediction.

Second, we explore if accurate model predictions can be obtained if similar (but not identical) noise signals are used for training and testing. This question is addressed since the use of disjunct training and testing data is desirable for models that might ultimately be used in real-world applications, while many successful ASR-based perception models require access to speech and/or noise signals (Cooke, 2006; Jürgens and Brand, 2009; Schädel et al., 2016; Spille et al., 2018). Specifically, we investigate if the model proposed by Spille et al. (2018) suffers from overtraining, or if the approach is suitable for predicting HSR when only the noise statistics are known to the model, but the specific noise signals remain unknown.

II. METHODS

A sketch of the methods employed and the respective subsections explaining the details is provided in Fig. 1.

A. Stimuli

The speech material used for the listening experiments and for the model predictions is the German Oldenburg Sentence Test (OLSA) (Wagener et al., 1999a,b,c). The OLSA is a matrix test with the structure ⟨name⟩ ⟨verb⟩ ⟨number⟩ ⟨adjective⟩ ⟨object⟩ (e.g., “Peter bekommt vier nasse Ringe”/“Peter got four wet rings”). Each of the five categories contains ten different words which allows one to create 10^5 different sentences. For the actual test implementation, 120 different sentences have been recorded by a German female speaker. If an ASR system is trained with speech data from one speaker only, this usually results in degraded performance for other speakers, while errors for the speaker in the training set are reduced, i.e., the system’s performance becomes speaker-dependent. In order to obtain a speaker- and gender-independent ASR-based model, we used an inhouse corpus for ASR training containing matrix sentences that follows the matrix structure and vocabulary of the original OLSA test. This corpus has a total duration of ten hours and contains speech recorded from 20 different speakers (10 female, 10 male) who produced random matrix sentences (Meyer et al., 2015).

The masking of the speech signals was carried out using eight different noise types that can be divided into two groups: (A) speech-like maskers that were produced by using original speech recordings (which are independent of the clean training/testing sets described above) and modifying them and (B) noise-like maskers that were produced by starting out with a stationary noise with the long-term spectrum of speech, which is modified by applying different types of modulation. To produce a database of these noise signals with longer duration than the original noise signals, the collection from Schubotz et al. (2016) was considerably extended using the same synthesis properties.

The speech-like maskers include an international speech signal (ISS), a noise-vocoded version of the ISS (NV-ISS), a...
single talker (ST) and a noise-vocoded version of the ST (NV-ST). The ISS contains seven different languages (Danish, Dutch, English, French, German, Norwegian and Swedish) and is based on the speech recordings of 30 different speakers per language. To create the ISS, single words of the recordings were randomly combined with speech pauses that reflect the pause distribution of normal speech. The noise-vocoded signals NV-ISS and NV-ST were calculated by applying a Gammatone filterbank analysis to the audio signals, extracting the Hilbert envelope of each channel, multiplying it with a white noise and recombining the signals.

To create the noise-like maskers, a speech-shaped noise (SSN) was generated by using the ISS as a basis and randomizing its phase. This SSN was then modified in three different ways to generate noises that became gradually more similar to speech. The result of the first modification is a sinusoidal amplitude modulated SSN (SAM-SSN) to take the effects of temporal modulation into account. The highest perceptual sensitivity is achieved at half the modulation frequency (Simpson et al., 2013). To realize the temporal modulation, an 8 Hz modulation was applied to the SSN to represent the typical syllable rate of 4 Hz (Greenberg et al., 1996). The second one is the multiplication of the SSN with the envelope of a broadband speech signal (BB-SSN) of the inhouse corpus containing matrix sentences, as previously mentioned in order to obtain a closer representation of the temporal modulation changes of speech. The last modification is the across-frequency shifted SSN (AFS-SSN). To obtain this noise type, the 16 frequency channels of a gammatone filterbank were divided into four blocks and modulated with a random section of the speech envelope.

The listening experiments and model predictions are conducted with matrix sentences produced by a male speaker, because these recordings have been optimized to obtain an equal intelligibility over all sentences and result in a test-retest reliability of 0.5 dB for normal-hearing listeners (Wagener et al., 1999b,c). Using data from only one speaker for testing would be unconventional for regular ASR. In this ASR-based modeling approach, we use data from one speaker since this is the data used in clinical tests of listeners, i.e., only utterances for which human responses exist are being used. The recordings are embedded in eight noises derived from speech produced by (different) male speakers (male noise).

To avoid gender-dependent training of the ASR component of the model, we additionally created eight female maskers. Each of the 16 resulting noise signals has a length of 11 h; 80% of this data were used for ASR training, and 20% of the male masking signals were later used for ASR testing and listening experiments. 20% of the female maskers remained unused; they were not added to the training data since it should be balanced with respect to female/male data.

In order to evaluate DAILY, the predictions are compared with the results of listening experiments with NH listeners (Schubotz et al., 2016), with the predictions of Spille et al. (2018), as well as with HSR data collected for this study. The noise signals used in the referenced two studies are similar to the current study, but are limited to approximately 60 s in duration and are referred to as SBKE16 noises here. To distinguish between the SBKE16 noises and the noise signals of the current study, our signals are referred to as RKM20. Figure 2 shows an overview of the different noises and their use for the predictions of the various models. The authors of the comparative studies used the international speech test signal (ISTS) (Holube et al., 2010) instead of the ISS as well as a different single talker. The ISTS differs from the ISS noise that we used, since it is based on combined syllables from speech streams of six different languages (while the ISS is based on isolated words, as explained above). Note that the original ISTS combines carefully selected syllables that were manually stitched together, which is the reason why the resulting noises are quite short. For producing the ISS used for the first time in this study, we opted for an automated word-level segmentation since a manual processing for noises with a duration of 11 h was not feasible. Both signals cover several languages, sound very similar to speech and do not contain recognizable sentences. Our ISS was created using an automated script that selected individual words from clean speech signals with the aim of creating very long noise signals. This is also the reason that a manual syllable-based structure was not chosen since a careful design was not feasible for a signal with a duration of 11 h (in contrast to 60 s duration for the ISTS). The differences between the noise types could have a minor effect on listening experiments, but we presume they should not affect ASR results since both syllable-based and regular words are not contained in the ASR dictionary.

B. ASR features, training, and testing

The ASR training and testing are carried out with mixed audio signals containing speech and noise by using the
For each noise signal, a function is fitted to the averaged strongly depend on the randomly selected segment. To computation. For modulated maskers, the masking strength can between –30 and 20 dB to sample the psychometric func-
mixed with each of the male noises at 400 random SNRs trained on all noise types simultaneously.

ASR tasks) are used as training data. Thus, one DNN was chosen SNRs in a range of –10 to 20 dB SNR. The resulting female version of each of the eight noise types) at randomly speech files are mixed with each of the 16 noises (male and DNN dictionary contains only the words of the OLSA. The training material is limited to the OLSA recordings, the which results in a speaker-independent ASR system. Since

find the transcript with the highest probability given the estimate of the DNN is processed by a Hidden Markov model to

put of the DNN is processed by a Hidden Markov model to
ace normalization.

bank to combine different DCT bins and a mean and variance normalization.

These features are used as input to the DNN, which maps the features to context-dependent triphones. The network is a fully connected feed-forward model and consists of seven hidden layers with 2048 neurons per layer. The output of the DNN is processed by a Hidden Markov model to find the transcript with the highest probability given the estimated triphones.

The ASR training is carried out using the OLSA sentences produced by 20 different speakers (Meyer et al., 2015), which results in a speaker-independent ASR system. Since the training material is limited to the OLSA recordings, the DNN dictionary contains only the words of the OLSA. The speech files are mixed with each of the 16 noises (male and female version of each of the eight noise types) at randomly chosen SNRs in a range of −10 to 20 dB SNR. The resulting 80 h of mixed signals (a typical amount of data for smaller ASR tasks) are used as training data. Thus, one DNN was trained on all noise types simultaneously.

To test the ASR system, the original recordings of the OLSA sentences (male speaker) are used. The sentences are mixed with each of the male noises at 400 random SNRs between –30 and 20 dB to sample the psychometric function. For modulated maskers, the masking strength can strongly depend on the randomly selected segment. To compensate for the resulting variability, eight sentences are mixed at the random SNR and the predictions are averaged. For each noise signal, a function is fitted to the averaged data points to obtain the psychometric functions of the model. When calculating the fit, the chance probability of 10% of the OLSA test is taken into account and set as the function minimum.

For investigating if the newly generated RKM20 noises are responsible for the differences in the predictions of the model of Spille et al. (2018) and DAILY, DAILY is additionally trained in the same way as the original model. Instead of using 80% of the 11 h of each noise for training and 20% for testing of DAILY, only 60 s of each noise are used for training and testing (DAILY60s), thereby matching the duration of noise signals of previous studies.

C. Simulation of hearing loss in the model

For predicting SRTs of hearing-impaired listeners, we integrate a noise that simulates the elevation of hearing thresholds in hearing-impaired listeners (Beutelmann and Brand, 2006; Holube and Kollmeier, 1996; Kates and Arehart, 2005) into the processing chain of the ASR fea-
tures. After applying the mel-filter bank, the resulting magnitudes are converted into sound pressure levels, assuming a 65 dB broadband sound pressure level for the noise component (which had also been used for the listening experiments). Spectral components below the individual hearing threshold are replaced by Gaussian noise (at the level of the same hearing threshold) and become inaccessible for further processing steps. Afterwards, the levels are converted back into magnitudes. The remaining processing and training steps are identical to the procedure for NH listeners. To take into account the individual hearing loss of each listener, an individual ASR system is trained to represent each subject. This implies that hearing loss is integrated in training and test data and should therefore reflect HI listeners who are accustomed to their hearing loss.

D. Listening experiments

To evaluate the model predictions, HSR measurements with NH and HI listeners are performed. The comparison with NH subjects was carried out using the data from listening experiments of Schubotz et al. (2016). The authors measured the SRT using the OLSA test with the adaptive procedure of Brand and Kollmeier (2002) and eight NH subjects for the same noise types as described in Sec. II A.

For the evaluation of the HI extension of DAILY, listening experiments with 20 HI subjects (12 female, 8 male) between 53 and 80 years (median 71) were performed. All of them had a very mild, mild or moderate hearing loss (Bisgaard et al., 2010). The hearing losses were measured with the single-interval adaptive procedure based on the presentation of pure tones, described by Lecluyse and Meddis (2009) to obtain accurate audiograms for the HI simulation in DAILY. The model was developed to predict monaural HSR, so only the better ear of all subjects was used for modelling and also for the SRT measurements. Since model performance is usually evaluated using a group of listeners to reduce the influence of inter-subject variability (Biberger and Ewert, 2016; Jørgensen et al., 2013; Taal et al., 2011),
we group the HI listeners according to their hearing loss. The \( k \)-means clustering (based on minimizing the squared Euclidean distance between data points of the audiograms) was chosen as objective criterion to determine groups of HI listeners with different types of hearing loss. \( k = 3 \) was chosen since the total number of HI listeners was limited (20) and using a larger number of clusters could have resulted in a very small number of listeners per group. The result of \( k \)-means clustering was a grouping that very well matches the definition by Bisgaard et al. (2010) for very mild, mild, and moderate hearing loss. The choice of \( k = 3 \) was not optimized and the specific assignment of listeners to each group was made before any modeling results were obtained. The mean audiogram of each group is shown in Fig. 3. As for the NH listeners, SRT measurements using the OLSA were performed by each HI subject for all eight different noises.

E. Baseline models

To quantify the predictive power of the model, the predictions of DAILY for NH listeners are compared with the results of HASPI (Kates and Arehart, 2014), SII (ANSI, 1997), ESII (Rhebergen and Versfeld, 2005), STOI (Taal et al., 2011), mr-sEPSM (Jørgensen and Dau, 2011), and mrGPSM (Biberger and Ewert, 2016) and the ASR-based model of Spille et al. (2018). The predictions of SII and ESII are calculated by using the SIP-toolbox provided by Fraunhofer IDMT (2014). The ESTOI predictions are obtained by using the MATLAB implementation from Jensen and Taal (2016). The mapping from model outputs to recognition scores (RS) was performed by using the SSN condition as reference to determine the coefficients \( a \) and \( b \) of the mapping function 1.

\[
RS(x) = 100 \times \frac{1}{1 + \exp \left( a \times \text{ESTOI}(x) + b \right)} .
\]

To investigate if the difference between the model of Spille et al. (2018) and DAILY are based on the newly created noises (i.e., to assess if overfitting plays a role in previous model approaches), the predictions are obtained for the SBKE16 noises as well as for the RKM20 noises. For the HI predictions, HASPI is used as additional baseline model.

III. RESULTS

The results of this study are broken down into four groups, i.e., NH listeners and HI subjects with different degrees of hearing loss (cf. Fig. 3).

A. Model predictions for normal-hearing listeners

1. Psychometric functions

The upper column of Fig. 4 shows the psychometric functions of the NH subject mean (Schubotz et al., 2016), of the original model (Spille et al., 2018) and of the proposed DAILY model that uses different noise signals for training and testing.

The results are shown for each of the eight noise types. Overall, the NH model predictions seem to provide a good match for the SRT in comparison to the HSR data. In several conditions, the slopes of DAILY deviate noticeably from the slopes of the subjects’ psychometric functions (noise types SAM-SSN, ISS and NV-ISS). The RMSE of the SRTs and of the slopes at the 50\% HSR point of the psychometric functions are additionally listed in Table I (second-last column). The RMSE of the slopes (RMSE\(_{\text{slope}}\)) of DAILY (0.04) is identical to the RMSE\(_{\text{slope}}\) of DAILY\(_{60s}\) and only slightly higher in comparison to the RMSE\(_{\text{slope}}\) of the model provided by Spille et al. (2018) (0.03).

2. Effect of separate training and testing noises

In order to disentangle the effects of different training/testing noises and short vs long training noise signals [two important differences between Spille et al. (2018) and the model introduced in this study], we contrasted the results of our current DAILY model with a model for which the duration of training noises was limited to only 60 s (DAILY\(_{60s}\)), shown in Fig. 5(A). The figure shows that the predicted SRTs of DAILY\(_{60s}\) are very similar compared to DAILY, which uses the full-length noise signals. This is also evident from the comparison of RMSE, \( r_{\text{SRT}} \) and RMSE\(_{\text{slope}}\) of the two models in Table I. Note that the psychometric functions are not displayed in Fig. 4 because they largely coincide with the results obtained with DAILY.

3. Comparison with baseline models

To compare the results of the ASR-based approach (including the influence of the newly created RKM20 noises) and the baseline models, the RMSE\(_{\text{SRT}}\) and the correlation of the model predictions are listed in Table I. Both
ASR-based models provide consistently high and significant correlations between the predictions and the measurements. The baseline models HASPI, ESII, and mrGPSM also produce significant correlations, but the RMSE\textsubscript{SRT} are higher (between 4.6 dB and 11.4 dB) compared to the two ASR-based models (1.6 dB and 1.9 dB, respectively). In order to be able to calculate the HSR of the models for other SNRs, the slope at 50% recognition is also necessary. To take this into account, the RMSE\textsubscript{slope} of all baseline models are additionally listed in Table I. Compared to the ASR-based models, the RMSE of slope estimates is higher with absolute differences ranging from 0.02 to 0.14.

B. Model predictions for hearing-impaired listeners

The predictions as well as the measurements of the HI listeners are averaged within each of the three audiogram-based groups described in Sec. II.D. The three lower columns of Fig. 4 show the psychometric functions of the mean subject data of each HI group for the eight different noise types. Additionally, the psychometric functions of the predictions with DAILY and HASPI are shown. Figure 5(B) provides a different view on this data by directly comparing the model and HSR data. The groups of HI listeners are marked with different colors, the measured data and the two models are identified by different symbols. The accuracy of the DAILY predictions for HI listener groups is in the same range as the corresponding NH predictions. For two thirds of the data points, DAILY predicts slightly lower SRTs in comparison to the true SRT, i.e., the model overestimates the listeners’ performance for the majority of averaged measurements. All SRT predictions obtained with the HASPI model are higher (worse) than the measured recognition; indicating that HASPI underestimates listeners’ ability to recognize noisy speech in the conditions employed here. In Table II, the RMSE and the correlations between the measured results of the HI listener groups and the predictions of DAILY and HASPI are listed for each group as well as for

![Diagram of psychometric functions for eight different noise types and four groups of listeners.](Image)

**FIG. 4.** Predicted and empirically obtained psychometric for eight different noise types and four groups of listeners. Top row: Comparison of the mean subject data of NH listeners (Schubotz et al., 2016) with the model predictions from Spille et al. (2018) and DAILY. The lower three rows show the mean subject data and the predictions of DAILY and HASPI for three groups of HI listeners with different degrees of hearing loss. The SRTs are additionally shown in Fig. 5 for an easier comparison.

**TABLE I.** Overview for NH listeners of the root mean squared error (RMSE) and correlation of the SRT predictions and slope RMSE over all noise types.

| Noise: RKM20 | HASPI | SII | ESII | ESTOI | mr-sEPSM | mrGPSM | Spille | DAILY | DAILY\textsubscript{est} |
|-------------|-------|-----|------|-------|----------|--------|--------|-------|-------------------|
| RMSE\textsubscript{SRT} | 9.0  | 7.6 | 5.7  | 3.1  | 8.0      | 6.2    | /      | 1.6   | 2.2               |
| \textit{r}\textsubscript{SRT} | 0.79\textsuperscript{a} | -0.18 | 0.73\textsuperscript{a} | 0.87\textsuperscript{b} | 0.56 | 0.71\textsuperscript{a} | / | 0.90\textsuperscript{b} | 0.92\textsuperscript{b} |
| RMSE\textsubscript{slope} | 0.05 | 0.12 | 0.07 | 0.06 | 0.06 | 0.05 | /      | 0.04 | 0.04             |

| Noise: SBKE16 | HASPI | SII | ESII | ESTOI | mr-sEPSM | mrGPSM | Spille | DAILY | DAILY\textsubscript{est} |
|---------------|-------|-----|------|-------|----------|--------|--------|-------|-------------------|
| RMSE\textsubscript{SRT} | 11.4 | 7.4 | 7.7  | 4.1  | 7.3      | 4.6    | 1.9    | /      | /                 |
| \textit{r}\textsubscript{SRT} | 0.90\textsuperscript{b} | 0.44 | 0.81\textsuperscript{b} | 0.90\textsuperscript{b} | 0.62 | 0.84\textsuperscript{b} | 0.95\textsuperscript{b} | / | /                  |
| RMSE\textsubscript{slope} | 0.04 | 0.12 | 0.06 | 0.05 | 0.05 | 0.05 | 0.03 | / | /                  |

\textsuperscript{a}p < 0.05.  
\textsuperscript{b}p < 0.01.  
\textsuperscript{c}p < 0.001.  

---

J. Acoust. Soc. Am. 151 (3), March 2022
the average of these groups. The RMSE$_{SRT}$ of the DAILY predictions for HI listener groups (between 1.1 and 1.7 dB) are similar to the DAILY predictions for NH listeners (1.6 dB) and much lower than the RMSE$_{SRT}$ of HASPI (between 8.0 and 8.6 dB). The correlation coefficient of DAILY is for two HI groups highly significant (0.96 and 0.93). The significance level of the HASPI predictions is lower for these two groups as well as their correlation values (0.92 and 0.74, respectively). For the third HI group, both models produce no significant correlations.

IV. DISCUSSION

A. Model predictions for normal-hearing listeners

1. Psychometric functions

The psychometric functions in Fig. 4 show a good match between the measured subject data and the predictions. The SRTs are very similar and the slopes for the predictions of DAILY are a bit lower than for the subject data. These lower slopes are caused by the higher recognition rate of DAILY in low SNR conditions, which could be caused by different signal processing strategies in humans and machines in modulated noise: For human listeners, temporal masking effects such as forward and backward masking limits their ability to listen into the dips (Bronkhorst, 2000). The ASR-based modelling employed by DAILY only partially reflects this property of the auditory system: Even though the AMFB features represent some sluggishness in temporal envelope processing in the low-modulation-frequency-centered bands that yield some limitations for listening into the dips, there are still higher modulation frequencies accessible in the complete feature set vector space which does not prevent the ASR system from listening into the dips. Hence, it appears that the ASR system is able to exploit information from noise valleys even if their duration is only a few milliseconds (which is connected to the block length of 10 ms).

2. Effect of separate training and testing noises

The predictions of DAILY are compared to DAILY$_{60s}$, which included only the identical 60 s of each noise for training and testing. The results shown in Fig. 5(A) indicate that both model versions produce similar SRT predictions, which is also reflected in the RMSE between DAILY and DAILY$_{60s}$ (not shown in the tables), which amounts to 1.2 dB. The largest deviations are observed for the single-talker condition, for which DAILY$_{60s}$ produces SRTs closer to the HSR data. We assume this result reflects the difficulty of recognizing speech in the presence of a competing talker for an ASR system: This task should be much simpler if the competing signal is fairly well-known to the ASR system by using identical short segments for training and test data. NH listeners are fairly robust to unknown speakers, which is reflected in their very low SRTs (Fig. 5). However, even though the use of matched training/test noises often results in a decrease in the SRT, this does not imply improved predictions, since in some cases the ASR-based model is too robust with respect to noises derived from speech signals. Hence, using matched noises might be detrimental for HSR prediction since the ASR system might underestimate the SRT.

3. Comparison with baseline models

In our experiments, ASR-based models produced more precise predictions than the baseline models. There are

FIG. 5. Measured and predicted SRTs for (A) NH listeners, including data from a model trained and tested with identical noises (Spille et al., 2018) or trained with very short noise signals (DAILY 60 s) and (B) three groups of HI listeners. Whiskers denote the standard deviation.

![Graph showing SRT predictions for NH and HI listeners](image)

TABLE II. Overview for HI listener groups of the root mean squared errors (RMSEs) and correlations of the SRT predictions and slope RMSE over all noise types.

| HI    | HASPI RMSE$_{SRT}$ | DAILY RMSE$_{SRT}$ | $r_{SRT}^a$ | $r_{SRT}^b$ | RMSE$_{slope}$ | RMSE$_{slope}$ |
|-------|-------------------|--------------------|-------------|-------------|----------------|----------------|
| HI1   | 8.0 dB            | 1.3 dB             | 0.92$^a$    | 0.96$^b$    | 0.07           | 0.06           |
| HI2   | 8.4 dB            | 1.1 dB             | 0.74$^a$    | 0.93$^b$    | 0.10           | 0.10           |
| HI3   | 8.6 dB            | 1.7 dB             | 0.43        | 0.70        | 0.11           | 0.11           |
| HI total | 8.3 dB            | 1.4 dB             | 0.89$^b$    | 0.94$^b$    | 0.08           | 0.07           |

$^a p < 0.01.$

$^b p < 0.001.$

$^c p < 0.05.$
several different possible explanations for this difference. The baseline models need to be mapped to a reference condition or contain a noise-dependent decision stage, so they can be optimised for predictions of single noise types, but deteriorate for predictions of such a diverse noise set as used in this study. This is the reason for the large RMSEs of the baseline models. Moreover, they typically generalize across different speakers and are insensitive to the information contained in the speech material and its relevance for speech recognition. On the contrary, the ASR-based models use neither a mapping nor a decision stage and are optimized during the training to predict the HSR for the whole noise set. The limitations of the model are discussed in Sec. IV C. The predictions of all baseline models are based on a comparison of two signals (depending on the model there are different combinations of clean speech, clean noise, and noisy speech). Only the ASR-based models get merely the mixed signal and try to recognize speech instead of comparing signals. ASR systems have become relatively robust due to the developments in deep learning (in comparison to ASR systems that have been used a decade ago) so that they achieve a similar classification performance as human listeners, at least for a small-vocabulary speech task such as the Oldenburg sentence test. These systems are similarly affected by noise as human listeners, so speech masked with noise is no longer an issue for ASR with deep learning. This could partially be attributed to the use of features that are inspired by our auditory processing, in our case, the AMFB features. The diversity within one noise type could additionally affect the predictions. Therefore, they are carried out for the SBKE16 noises as well as for the RK2M20 noises. The newly generated noises have only a minor influence on the model predictions. The mean RMSE$_{SRT}$ over all baseline models is 7.1 dB (SBKE16 noises) and 6.6 dB (RK2M20 noises). This influence is also evident in the ASR-based models. The model of Spille et al. (2018) achieves an RMSE$_{SRT}$ similar to DAILY (1.9 and 1.6 dB, respectively). In addition, the up to significantly higher RMSE$_{slope}$ of the baseline models compared to the ASR-based models lead to much higher prediction errors in the upper or lower SNR regions of the psychometric function.

B. Model predictions for hearing-impaired listeners

Figures 4 and 5(B) as well as Table II show that the predictions of DAILY are close to the measured HSR data and are highly correlated to each other. Note that the predictions are only based on the individual audiograms, i.e., no supra-threshold processing deficits often seen in HI listeners is accounted for. This may be one reason why for the majority of conditions with HI listeners, DAILY overestimates the subjects’ performance (Fig. 5); this is true for 16 out of a total of 24 conditions, where 24 is the product of 8 (noise types) and 3 (groups of HI listeners). Effects of hearing loss not captured by the audiogram have been considered for the simulation of hearing loss in other ASR-based speech recognition models in the literature: Schädler et al. (2020, 2015) and Kollmeier and Kiessling (2018) modelled the suprathreshold distortion component (Plomp, 1978) by inserting a level uncertainty (equivalent to a multiplicative noise applied to the speech features prior to the recognition stage). They could improve their prediction performance by fitting the amount of this distortion to data from a different experiment. However, their approach was not applied to such a variety of (modulated) noise maskers as in the current study and, in addition, their speech recognizer was much more restricted to the Matrix sentence test material as the one employed here. Hence, the high prediction accuracy achieved here using only audibility information might be due to large dynamic range of the masker fluctuations which may yield a higher influence of the audibility of soft parts of the combined input signal than any distortions that gain more importance if all parts of the input signals are audible (which especially applies for stationary-noise conditions). Fontan et al. (2020), on the other hand, used a generic model of hearing impairment (with different combinations of hearing threshold, frequency selectivity loss and loudness recruitment) in combination with a generic ASR system to predict speech recognition in quiet. However, they experienced a considerable human-machine gap between actual and predicted performance, thus limiting the relevance of their finding for the problem under consideration here. Taken together, it is remarkable how small a prediction error is found here for the audiogram- and ASR-based prediction with DAILY for the variety of conditions employed here — even without taking additional factors of hearing impairment such as suprathreshold distortion or cognitive processing into account. Their appropriate incorporation into the ASR-based HSR prediction is a matter of future research.

C. Limitations of DAILY

Compared to all baseline models investigated in this paper, our modelling approach requires the training of an ASR system, which entails the requirement for a speech database with at least tens of hours of speech. Since training of deep learning algorithms on regular processors is time consuming, DAILY also profits from dedicated hardware in the form of graphical processing units (GPUs), which is especially important when individual models are trained to predict the SRTs of HI listeners. Although the feature extraction scheme of the DAILY models implements strategies such as frequency grouping based on frequency perception, the model also features a deep learning component for which the interpretability of the signal processing strategies is limited in comparison to models that implement individual processing steps of the auditory system. However, relevant input-output relations can still be identified with models that exploit deep learning [as has previously been done for simple time-frequency features (Spille et al., 2018)]. To calculate the WER during test time, DAILY also compares the true word labels to the transcript, while baseline models produce a prediction without the true word labels. On the other hand, since the model is based on a
regular ASR system, it does not need the clean speech reference (again, in contrast to all baseline models), which is one requirement for models in the loop that can be applied in real world scenarios, as briefly discussed in the following. A potential scenario of ASR-based models of perception is their use as a model in the loop, i.e., for continuous optimization of speech recognition in real-world situations; this could be achieved by monitoring different speech processing strategies (e.g., hearing aid settings) from which the best strategy is selected. The modeling approach presented here is a small step towards this direction, since it is blind with respect to the speech signals that are required for intrusive models of perception. Further, it drops the requirement for identical noise signals in training and test data compared to its predecessor. However, several further limitations of the model need to be addressed before it could be useful as model in the loop: First, knowing the true word labels before speech is produced is highly unrealistic in real-world situations. A remedy could be to estimate the WER of the ASR system, for instance by measuring the degradation of phoneme probabilities obtained from the DNN. This approach was shown to produce accurate predictions for subjective perceived speech quality (Huber et al., 2018b) and listening effort (Huber et al., 2018a) and should be explored in the context of HSR prediction as well. The combination of an ASR-based system with different signal-based recognition estimates (dispersion, entropy, likelihood ratio, time alignment difference, and normalized likelihood difference) has also been successfully implemented by Karbasi et al. (2020) and represents a promising approach to achieve a non-intrusive HSR model. Second, different segments of the same noise signals were used for training DAILY; this kind of a priori information would limit the model to very narrow application scenarios. In combination with the aforementioned approach for estimating the WER, larger training sets could be used, taking into account a large variety of training noises with the aim of generalizing to unseen noise types. Huber et al. (2018c) used a training set with 8000 h of noisy speech which generalized to new noises. It remains to be seen if this finding is transferable from modeling perceived listening effort to HSR prediction. Other factors that need to be addressed in the future are the exploration of reverberated speech, binaural hearing, and using arbitrary, continuous speech. Finally, real-time constraints also need to be considered: Although the training procedure is computationally costly, the forward run of the DNN is not and therefore compatible with co-processors for hearing aids, at least for smaller networks Castro Martínez et al. (2019). Our modeling approach should therefore be tested taking into account such smaller networks that are limited to few hidden layers with 500 neurons per layer.

V. CONCLUSION

This study extended and evaluated a modeling approach based on an ASR system for predicting HSR of 8 NH and 20 HI listeners in the presence of complex noise maskers. The model combines amplitude modulation features with a deep neural network and is referred to as DAILY; the resulting error rates are used to obtain a psychometric function and calculate its slope and the SRT, which are compared to corresponding HSR measurements in NH and HI listeners. While previous work used identical, short noise signals for training and testing (Spille et al., 2018), the current model uses similar but not identical noise signals in multi-condition training to obtain an SRT prediction error (in terms of the RMSE) of 1.6 dB for NH listeners, which is considerably lower than the RMSE produced by six competing baseline models. The current model allows to accurately predict the HSR performance of HI listeners (group RMSE below 2 dB) with different degrees of hearing loss for a variety of maskers with increasing complexity in temporal modulation and similarity to real speech. The remarkable precision is achieved even without considering the effect of suprathreshold distortions or other (e.g.) cognitive aspects by just considering the recognition process using auditory-perception motivated AMFB features. This indicates a higher weighting of audibility than suprathreshold distortion representation at least for the fluctuating masking conditions employed here. The correlation coefficients between predicted and observed SRTs are also consistently high ($r \geq 0.9$) for all groups (NH and HI) with the exception of the group with the strongest hearing loss ($r = 0.7$).

ACKNOWLEDGMENTS

This research was supported by the DFG (Cluster of Excellence 1077/1 Hearing4all). The authors want to thank Jasper Ooster for his support during the measurements and Jan Rennies-Hochmuth for providing the SIP-Toolbox. We also thank Constantin Spille, Thomas Biberger, Wiebke Schubotz, and Stephan Ewert for sharing their data and scripts.

Andersen, A. H., de Haan, J. M., Tan, Z-h., and Jensen, J. (2017). “A non-intrusive Short-Time Objective Intelligibility measure,” in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE, New York), pp. 5085–5089.

ANSI (1969). ANSI S3.5-1969, Methods for the Calculation of the Articulation Index (American National Standards Institute, New York).

ANSI (1997). ANSI S3.5–1997, American National Standard Methods for Calculation of the Speech Intelligibility Index (American National Standards Institute, New York).

Beutelmann, R., and Brand, T. (2006). “Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners,” J. Acoust. Soc. Am. 120(1), 331–342.

Biberger, T., and Ewert, S. D. (2016). “Envelope and intensity based prediction of psychoacoustic masking and speech intelligibility,” J. Acoust. Soc. Am. 140(2), 1023–1038.

Bisgaard, N., Vlaming, M. S., and Dahlquist, M. (2010). “Standard audigrams for the IEC 60118-15 measurement procedure,” Trends Amplif. 14(2), 113–120.

Brand, T., and Kollmeier, B. (2002). “Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests,” J. Acoust. Soc. Am. 111(6), 2801–2810.

Bronkhorst, A. W. (2000). “The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions,” Acustica 86(1), 117–128.

Castro Martínez, A. M., Gerlach, L., Payá-Vayá, G., Hersmansky, H., Ooster, J., and Meyer, B. T. (2019). “DNN-based performance measures
for predicting error rates in automatic speech recognition and optimizing hearing aid parameters.” Speech Commun. 106, 44–56.

Cooke, M. (2006). “A glimpsing model of speech perception in noise,” J. Acoust. Soc. Am. 119(3), 1562–1573.

Dau, T., Kollmeier, B., and Kohlrausch, A. (1997). “Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers,” J. Acoust. Soc. Am. 102(5), 2892–2905.

Ewert, S. D., and Dau, T. (2000). “Characterizing frequency selectivity for envelope fluctuations,” J. Acoust. Soc. Am. 108(3), 1181–1196.

Fletcher, H. (1940). “Auditory patterns,” Rev. Mod. Phys. 12, 47–65.

Fontan, L., Cretin-Maïnzenz, T., and Füllgrabe, C. (2020). “Predicting speech perception in older listeners with sensorineural hearing loss using automatic speech recognition,” Trends Hear. 24, 233121652091476.

Fraunhofer IDMT (2014). “SIP Toolbox,” https://www.idmt.fraunhofer.de/de/institute/projects-products/sip-toolbox.html (Last viewed January 11, 2021).

Greenberg, S., Hollenback, J., and Ellis, D. (1996). “Insights into spoken language gleaned from phonetic transcription of the switchboard corpus,” in International Conference on Spoken Language Processing 24–27.

Holube, I., Fredelake, S., Vlaming, M., and Kollmeier, B. (2010). “Development and analysis of an international speech test signal (ISTS),” Int. J. Audiol. 49(2), 891–903.

Holube, I., and Kollmeier, B. (1996). “Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model,” J. Acoust. Soc. Am. 100(3), 1703–1716.

Huber, R., Krüger, M., and Meyer, B. T. (2018a). “Single-ended prediction of listening effort using deep neural networks,” Hear. Res. 359, 40–49.

Huber, R., Ooster, J., and Meyer, B. T. (2018b). “Single-ended speech quality prediction based on automatic speech recognition,” AES J. Audio Eng. Soc. 66(10), 759–769.

Huber, R., Pusch, A., Moritz, N., Rennies, J., Schepker, H., and Meyer, B. T. (2018c). “Objective assessment of a speech enhancement scheme with an automatic speech recognition-based system,” in Proceedings of the ITG Conference on Speech Communication, pp. 86–90.

Jensen, J., and Taal, C. H. (2016). “An algorithm for predicting the intelligibility of speech masked by modulated noise maskers,” IEEE/ACM Trans. Audio Speech Lang. Process. 24(4), 2009–2022.

Jørgensen, S., and Dau, T. (2011). “Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing,” J. Acoust. Soc. Am. 130(3), 1475–1487.

Jørgensen, S., Ewert, S. D., and Dau, T. (2013). “A multi-resolution envelope-power based model for speech intelligibility,” J. Acoust. Soc. Am. 134(1), 436–446.

Jürgens, T., and Brand, T. (2009). “Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model,” J. Acoust. Soc. Am. 126(5), 2635–2648.

Karbasi, M., Bleec, S., and Kolossa, D. (2020). “Non-intrusive speech intelligibility prediction using automatic speech recognition derived measures,” http://arxiv.org/abs/2010.08574 (Last viewed May 14, 2021).

Kates, J. (2013). “An auditory model for intelligibility and quality predictions,” Proc. Migs. Acoust. 19, 050134.

Kates, J. M., and Arehart, K. H. (2005). “Coherence and the speech intelligibility index,” J. Acoust. Soc. Am. 117(4), 2224–2237.

Kates, J. M., and Arehart, K. H. (2014). “The hearing-aid speech perception index (HASPI),” Speech Commun. 65, 75–93.

Kollmeier, B., and Kiessling, J. (2018). “Functionality of hearing aids: State-of-the-art and future model-based solutions,” Int. J. Audiol. 57(5), S3–S28.

Lecluyse, W., and Meddis, R. (2009). “A simple single-interval adaptive procedure for estimating thresholds in normal and impaired listeners,” J. Acoust. Soc. Am. 126(5), 2570–2579.

Meyer, B. T., Kollmeier, B., and Ooster, J. (2015). “Autonomous measurement of speech intelligibility utilizing automatic speech recognition,” in Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 2982–2986.

Meyer, R. M., and Brand, T. (2013). “Comparison of different short-term speech intelligibility index procedures for fluctuating noise for listeners with normal and impaired hearing,” Acta Acust. Acust. 99(3), 442–456.

Moritz, N., Anemüller, J., and Kollmeier, B. (2015a). “An auditory inspired amplitude modulation filter bank for robust feature extraction in automatic speech recognition,” IEEE/ACM Trans. Audio Speech Lang. Process. 23(11), 1926–1937.

Moritz, N., Gerlach, S., Adilgolu, K., Anemüller, J., Kollmeier, B., and Goetze, S. (2015b). “A CHIME-3 challenge system: Long-term acoustic features for noise robust automatic speech recognition,” in 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) 468–474.

Povet, D., Ghoshal, A., Boulianne, G., Burgel, L., Glenbek, O., Goel, N., Herrmann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Steemer, G., and Vesely, K. (2011). “The Kaldi speech recognition toolkit,” in IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 1–4.

Rhebergen, K. S., and Versfeld, N. J. (2005). “A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners,” J. Acoust. Soc. Am. 117(4), 2181–2192.

Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (2006). “Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise,” J. Acoust. Soc. Am. 120(6), 3988–3997.

Sakoe, H., and Chiba, S. (1978). “Dynamic programming algorithm optimization for spoken word recognition,” IEEE Trans. Acoust. Speech Sign. Process. 26(1), 43–49.

Schädler, M. R., Hülsmeyer, D., Warzybok, A., and Kollmeier, B. (2020). “Individual aided speech-recognition performance and predictions of benefit for listeners with impaired hearing employing FADE,” Trends Hear. 24, 2331216520938929.

Schädler, M. R., Warzybok, A., Ewert, S. D., and Kollmeier, B. (2016). “A simulation framework for auditory discrimination experiments: Revealing the importance of across-frequency processing in speech perception,” J. Acoust. Soc. Am. 139(5), 2708–2722.

Schädler, M. R., Warzybok, A., Hochmuth, S., and Kollmeier, B. (2015). “Matrix sentence intelligibility prediction using an automatic speech recognition system,” Int. J. Audiology 54, 100–107.

Schädler, M. R., Warzybok, A., and Kollmeier, B. (2018). “Objective prediction of hearing aid benefit across listener groups using machine learning: Speech recognition performance with binaural noise-reduction algorithms,” Trends Hear. 22, 233121651876895.

Schubotz, W., Brand, T., Kollmeier, B., and Ewert, S. D. (2016). “Monaural speech intelligibility and detection in maskers with varying amounts of spectro-temporal speech features,” J. Acoust. Soc. Am. 140(1), 524–540.

Simpson, A. J., Reiss, D. J., and McAlpine, D. (2013). “Tuning of human modulation filters is carrier-frequency dependent,” PLoS One 8(8), e73590.

Spille, C., Ewert, S. D., Kollmeier, B., and Meyer, B. T. (2018). “Predicting speech intelligibility with deep neural networks,” Comput. Speech Lang. 48, 51–66.

Taal, C. H., Hendriks, R. C., Heudsen, R., and Jensen, J. (2011). “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” IEEE Trans. Audio Speech Lang. Process. 19, 2125–2136.

Vesely, K., Ghoshal, A., Burgel, L., and Povet, D. (2013). “Sequence-discriminative Training of Deep Neural Networks,” in Proceedings of INTERSPEECH. INTERSPEECH, Lyon, France, pp. 2345–2349.

Wagner, K. C., Brand, T., and Kollmeier, B. (1999a). “Development and evaluation of a German sentence test Part III: Evaluation of the Oldenburg sentence test,” Z. Audiol. 38(3), 86–95.

Wagner, K. C., Brand, T., and Kollmeier, B. (1999b). “Entwicklung und Evaluation eines Satztests für die deutsche Sprache Teil II: Optimierung des Oldenburger Satztests” (“Development and evaluation of a sentence test for the German language Part II: Optimization of the Oldenburg sentence test”), Z. Audiol. 38(3), 44–56.

Wagner, K. C., Köhnel, V., and Kollmeier, B. (1999c). “Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests” (“Development and evaluation of a sentence test for the German language I: Design of the Oldenburg sentence test”), Z. Audiol. 38(1), 4–15.