Prediction of speech intelligibility with DNN-based performance measures

ANGEL MARIO CASTRO MARTINEZ*, CONSTANTIN SPILLE, BIRGER KOLLMEIER AND BERND T. MEYER

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

In this paper, we present a speech intelligibility model based on automatic speech recognition (ASR) that combines phoneme probabilities obtained from a deep neural network and a performance measure that estimates the word error rate from these probabilities. In contrast to previous modeling approaches, this model does not require the clean speech reference or the exact word labels during test time, and therefore, less *a priori* information. The model is evaluated via the root mean squared error between the predicted and observed speech reception thresholds from eight normal-hearing listeners. The recognition task in both cases consists of identifying noisy words from a German matrix sentence test. The speech material was mixed with four noise maskers covering different types of modulation. The prediction performance is compared to four established models as well as to the ASR-model using word labels. The proposed model performs almost as well as the label-based model and produces more accurate predictions than the baseline models on average.

INTRODUCTION

The intelligibility of speech is crucial for our social interaction, and it is an important measure for a diagnosis of hearing deficits through speech audiometry and for the optimization of speech enhancement algorithms in hearing aids or cochlear implants. Accurate models that predict the speech intelligibility (SI) in the presence of different masking noises are desirable since they can quantify the outcome of such an optimization and could, therefore, reduce the requirement of SI measurements that are usually time-consuming and costly.

Several models for SI prediction have been proposed that take into account the signal-processing strategies of the auditory system such as the speech-intelligibility index (SII; ANSI, S3 22-1997, 1997); the extended SII (ESII; Rhebergen and Versfeld, 2005) which extends SII to account for temporal modulations; the short-time objective intelligibility (STOI; Taal *et al.*, 2011), which is based on correlations between original and degraded signal; and the multi-resolution speech envelope power spectrum model (mr-sEPSM; Ewert and Dau, 2000), which incorporates temporal modulation filters in different frequency bands.

*Corresponding author: angel.castro@uni-oldenburg.de

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

Schubotz *et al.* (2016) compared these models in a study to determine how well they can predict the speech reception threshold (SRT), which is the signal-to-noise ratio (SNR) at which 50% of words presented are correctly recognized.

An alternative modeling approach combines signal extraction based on auditory principles with pattern matching algorithms borrowed from automatic speech recognition (ASR). For example, Barker and Cooke (2006) introduced a glimpsing model in which the above-threshold time-frequency patches (glimpses) were used as features for a backend that combines a Gaussian mixture model (GMM) with a hidden Markov model (HMM) to produce a transcript from the input glimpses, which was compared to listener responses. A GMM-HMM approach dubbed Framework for Acoustic Discrimination Experiments (FADE) was proposed in (Schädler *et al.*, 2015). This model produces SRT estimates by retraining a GMM-HMM system at different SNRs, and by selecting the model that produces the lowest SRT when using the same training and test sentences.

All previously mentioned models either require separate clean and degraded speech, or separate speech and noise signals. Motivated by the success of deep learning in ASR, Spille *et al.* (2018) proposed an ASR model that combines a deep neural network (DNN) trained to estimate phoneme probabilities given the acoustic observation with an HMM. The predictive power of this model exceeded the four baseline models mentioned above on the dataset collected by Schubotz and colleagues. The root-mean-square error (RMSE) between measurement and prediction was 1.8 dB on average when using multi-condition training as well as modulation features, which can be compared to the RMSE of baseline models in the range of 5.6 to 9.5 dB. The model is blind with respect to speech because training and test sets are speaker-independent. Therefore, it marks a step towards reference-free SI models, which could serve as models-in-the-loop in assisted hearing. A use case for such a model is the constant monitoring of SI in the current acoustic scene and to identify the speech enhancement algorithm that is optimal for that scene.

However, it requires the correct labels of the words in the utterance used as model input. These labels are compared to the transcript produced by the ASR system from which the recognition accuracy is calculated. For online applications of SI models, this is an essential limitation of the models.

In this paper, we introduce a model of SI prediction that does not require either the speech reference *or* the actual labels of the tested utterances. The model is based on the DNN-based approach introduced in Spille *et al.* (2018), but instead of computing the word error rate (WER), we test a method for *estimating* the WER directly from the phoneme posterior probabilities emitted by the DNN. The method explored here was first proposed for estimating phone error rates (Hermansky *et al.*, 2013) by analyzing the mean temporal distance or M-measure of phoneme vectors obtained from a neural network.

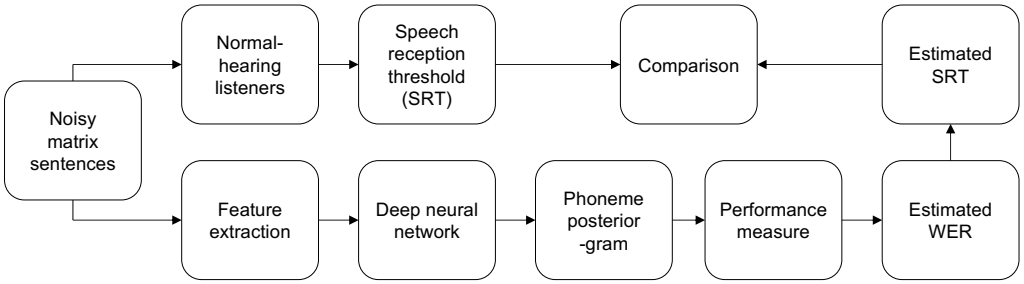

Fig. 1: Building blocks of the modeling approach: Speech intelligibility in noisy sentences is compared to the estimated SRT. To obtain this estimate, a DNN is trained as part of a standard ASR system and subsequently used to measure the degradation of phoneme representations in noise using a performance measure. From this, the WER of the ASR system is estimated, resulting in the predicted SRT.

In the current study, we quantify the relation of the M-measure with the WER and explore if accurate model predictions can be obtained with the DNN alone that operates on a relatively small temporal context window.

MODEL STRUCTURE

Figure 1 illustrates the structure of the proposed model. In previous work, Spille *et al.* (2018) estimated the SRTs from the target speech transcript, and the WER computed from the output of a regular hybrid ASR system. The components of this system are summarized below, together with the modification of using the M-measure to drop the requirement of word labels and estimate SRTs directly from the output of the acoustic model. The characterization of the acoustic model, together with the respective input features, closes this section.

ASR-based model of speech intelligibility

In the label-based ASR approach (Spille *et al.*, 2018), the acoustic model was trained on speech files mixed with different maskers at various SNRs using the Kaldi toolkit[†]. A fully-connected feed-forward DNN was used to map the acoustic features to posterior probabilities of context-dependent triphones. The time sequence of these probabilities was decoded using an HMM (three states for modeling phonemes and five for silence) to obtain a transcript of the utterance. This transcript was compared to the ground truth labels to obtain the word error rate (WER) from this sentence. By using utterances at various SNRs, a broad range of the corresponding WER estimations was obtained. These pairs of points were fitted to a psychometric function, as described by Wagener *et al.* (1999). SRT values served as a mean intelligibility predictor and were compared to SRTs obtained in listening experiments.

[†]The ASR was implemented using the Kaldi speech recognition toolkit (Povey *et al.*, 2011).

The model proposed in this paper differs from this approach by using the HMM during the training procedure only, and omitting the HMM (or any other language model) when predictions were obtained. Instead, a performance measure, as described in the next section, quantifies the degradation of the phoneme posteriorgrams. This measure could be informative about the WER and hence, about the SRT too. We hypothesize that the resulting measure should be sensitive to the SNR, but also show a similar sensitivity to masking noises similar to human listeners as long as the amount of training data is sufficient.

Estimating the word error rate

The WER estimation is based on a measure that quantifies the degradation of phoneme probabilities obtained from a DNN. We chose the mean temporal distance (MTD, also referred to as M-measure) (Hermansky *et al.*, 2013), which takes into account the distance of phoneme vectors and averages this distance. The underlying idea is that acoustically challenging conditions can have a temporal smearing effect on phoneme representations, i.e., the phoneme vectors become more similar. Acoustically optimal conditions produce very distinct phoneme activations, which become more distant in vector space. This distance is captured by the entropy-based divergence averaged over phoneme vectors in the interval from 50 to 800 ms.

The M-measure accumulates the average divergences of two phoneme posterior vectors $\mathbf{p}_{t-\Delta t}$ and $\mathbf{p}_t$ separated by a time interval of Δt and is defined as

$$\mathscr{M}(\Delta t) = \frac{1}{T-\Delta t} \sum_{t=\Delta t}^{T} \mathscr{D}(\mathbf{p}_{t-\Delta t}, \mathbf{p}_t) \qquad \text{(Eq. 1)}$$

where T is the duration of the analyzed representation, in this case, a portion of the posteriorgram. The symmetric Kullback-Leibler divergence is used as distance measure $\mathscr{D}$ between phoneme posterior vectors $\mathbf{p}_{t-\Delta t}$ and $\mathbf{p}_t$.

$$\mathscr{D}(\mathbf{p},\mathbf{q}) = \sum_{k=0}^{K} p^{(k)} \log \frac{p^{(k)}}{q^{(k)}} + \sum_{k=0}^{K} q^{(k)} \log \frac{q^{(k)}}{p^{(k)}} \qquad \text{(Eq. 2)}$$

As defined above, $p^{(k)}$ is the k-th element of the posterior vector $\mathbf{p} \in \mathbb{R}^k$.

We considered 16 values of Δt per utterance; from 50 to 800 ms in steps of 50 ms. For short Δt time spans, divergences are small, indicating neighboring frames often correspond to the same phoneme. The value increases with time up to a point at which both vectors $\mathbf{p}$, and $\mathbf{q}$ come from different coarticulation patterns, and the curve saturates.

As the acoustic model is trained to produce triphone posteriorgrams, to be decoded by the language model when performing ASR, an intermediate step of *grouping* the activations was performed to obtain monophone posteriorgrams. It is possible to

cluster triphones by mapping each transition as a branch of a decision tree, wherein the roots correspond to the central phoneme of the triphone. Monophone posteriorgrams of 42 dimensions were obtained by adding the corresponding activations, thus maintaining the distribution.

Monophone posteriorgrams yield M-measure values comparable to the ones obtained with the triphone equivalents at a lower computational cost without constraining the acoustic model of the temporal context if trained to produce monophones directly.

In our previous study (Castro Martinez *et al.*, 2019), we established the correlation between WER and the M-measure. In this work, we leverage this property, but the estimator ought to be in the same domain as the word recognition accuracy to estimate the SRTs and produce similar psychometric curves as human listeners. Given the non-linearity introduced by the M-measure, a mapping function is required to estimate WERs. The function used to map the M-measure to WER depends on the acoustic model and decays exponentially according to the following equation:

$$WER(\mathscr{M}) = A * e^{k*\mathscr{M}}. \qquad \text{(Eq. 3)}$$

The initial value A and the decay rate k were calculated on a cross-validation set comprised of utterances spoken by a speaker not included in the training set mixed with the same noise maskers described in the following section. Additionally, an upper boundary of 100 (the highest possible error rate) was imposed.

Features and deep neural network

The ASR system is trained with amplitude modulation filterbank (AMFB) features that are based on regular mel spectrograms with 40 frequency channels, which are processed with modulation filters in the range from 5 to 20 Hz (Moritz *et al.*, 2015). They were chosen since the explicit coding of temporal modulations increased model performance, especially for the across-frequency shifted speech-shaped noise (AFS-SSN) masker previously (Spille *et al.*, 2018). AMFB features were used as input to a DNN, which has the purpose of mapping the acoustic observations to phoneme probabilities. A fully-connected network (referred to as DNN) with six hidden layers and 2048 hidden (sigmoid) units was selected to compare this work and the previous SI prediction model from (Spille *et al.*, 2018). The network was trained to classify context-dependent triphones; every phone is modeled with three HMM states except for silence, which uses five states.

The training of the DNN described above was done in up to 20 epochs (stopping when the relative improvement was lower than 0.001). The starting learning rate was 0.008 (halving it every time the relative improvement was lower than 0.01). A soft-max layer of approximately 2000 units was attached to the output to produce the most likely posterior probabilities of each context-dependent HMM state.

An in-house corpus of 10 hours of speech from 20 speakers (10 male, 10 female) with the syntactical structure of Oldenburg Sentence Test (see next section) was selected as

a starting point to train the ASR system; sentences from the original speaker were not contained in the training set.

The training sets comprise of the clean data mixed with random parts of each of the eight different maskers (as described below) at random uniformly distributed SNRs ranging from -10 dB to 20 dB, resulting in 80 h of speech material. Two training sets were created which are based on noises created from a male or female voice. The test set to evaluate the ASR system was created by mixing eight random sentences from the speech material with parts of the respective masker for each of the 400 SNR values uniformly distributed between -30 dB and 20 dB to sample the whole psychometric function.

SPEECH MATERIAL, MASKERS AND SUBJECTIVE DATA

In this section, the speech material for both training and testing is described, the noise signals are introduced, and details about the listening tests from which the human SRTs were calculated by Schubotz *et al.* (2016) are provided.

Matrix test

Both the listening and ASR tests were performed using the *Oldenburger Satztest* (OLSA) (Wagener *et al.*, 1999), which is a matrix sentence test. It consists of 120 utterances produced by one speaker. Target sentences derived from a vocabulary of 50 words equally divided into five categories. For a review of matrix tests in several languages, please refer to (Kollmeier *et al.*, 2015). Each five-word sentence follows the fixed structure: <name><verb><number> <adjective><object>, e.g. *"Peter kauft sechs nasse Tassen"* ("Peter buys six wet cups"). Despite being grammatically correct, these sentences have no semantic context. Moreover, all combinations of words from each of the five categories can occur; therefore, predicting a sentence from previous ones is not possible.

Noise maskers

In the study carried out by Schubotz *et al.* (2016), a set of eight background maskers was created to evaluate the effect of energetic, amplitude modulation and informational masking on SI. We took this benchmark to evaluate our SI prediction model, focusing on four speech maskers described in the following.

First, a stationary SSN with the same long-term spectrum as the International Speech Test Signal (ISTS; Holube *et al.*, 2010) was used. Second, a sinusoidally amplitude-modulated SSN (SAM-SSN) was produced by adding an 8 Hz temporal modulation. The third masker was generated by multiplying the Hilbert envelope of a broadband speech signal with the SSN (BB-SSN). For the fourth, named across-frequency shifted SSN (AFS-SSN), the SSN was filtered in 32 frequency channels; subsequently, every four adjacent channels were multiplied with a different random time section of the

Hilbert envelope used for BB-SSN[‡].

To test the influence of same- or different-gender maskers, all maskers have a male and female version to match the long-term spectrum of the respective gender. Since the original ISTS contains female voices only, Schubotz *et al.* (2016) produced a male version of ISTS via the STRAIGHT algorithm (Kawahara *et al.*, 2008) to match its long-term spectrum.

Listening tests

To benchmark the performance of the proposed SI predictor, we compare the results from the listening tests performed in (Schubotz *et al.*, 2016). These experiments consisted of characterizing SI as a function of the SRT extracted from the adaptive procedure proposed by Brand and Kollmeier (2002). Eight normal-hearing participants (ages between 23-34) participated who were not previously exposed to the speech task; their hearing thresholds for pure tones did not exceed 20 dB at frequencies between 125 Hz and 8 kHz. During testing, the participants attended 20 OLSA sentences with an initial SNR of 0 dB; then, the SNR varied depending on the intelligibility measurement of the previous sentence. The procedure is set to determine the SNRs at which listeners correctly understand 50% (SRT) and 80% (SRT_{80}) of presented words. Each SRT resulted from a different set of sentences; in other words, each participant listened to 40 sentences per noise condition. Finally, the SRTs were averaged across the listeners to obtain the final SRT and SRT_{80} values, which are used to trace the psychometric function of the listeners (described entirely by the SRT and its slope). The slope of the psychometric function was estimated via a maximum-likelihood estimator (Brand and Kollmeier, 2002) with the 40 responses for each listener and masker.

RESULTS

Because the modeling approach presented in this paper is based on estimating the WER (cf. Figure 1), we first analyze if the error rate from the ASR is related to the predicted one based on the M-measure (Figure 2), where each data point corresponds to the error rate for eight matrix sentences.

While the WER with the SSN masker is overestimated and AFS-SSN data is underestimated, we observe a clear relationship between estimated and ASR WER for each masker. Additionally, the mapping is most sensitive at lower word error rates as the mapping function is a decaying exponential constrained to an upper boundary of 100.

To quantify the model performance, we compare the psychometric functions of the listeners to the approach using ASR generated transcripts and the proposed approach (Figure 3).

[‡]resulting in eight different adjacent modulation bands

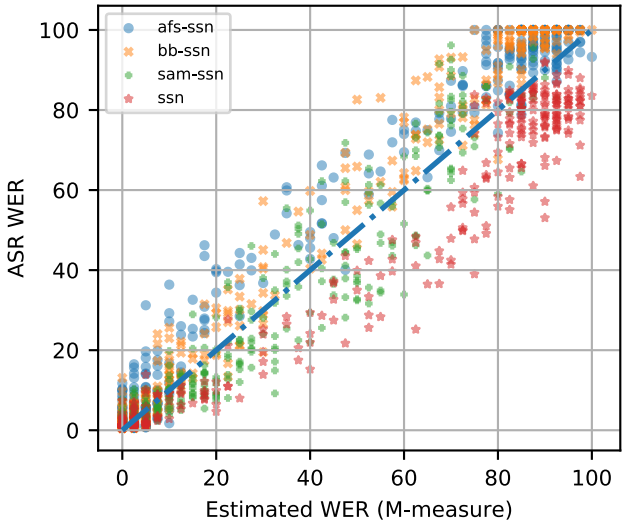

Fig. 2: Relation of estimated WER derived from phoneme probabilities and measured ASR WER for four different maskers.

When comparing both models, the overestimated WER values for the SSN result in a shift to a lower SNR for the new model (top left panel in Figure 3), while the shift to higher SNRs for the AFS-SSN and BB-SSN maskers is a reflection of an underestimation of the WER (noticeable in Figure 2).

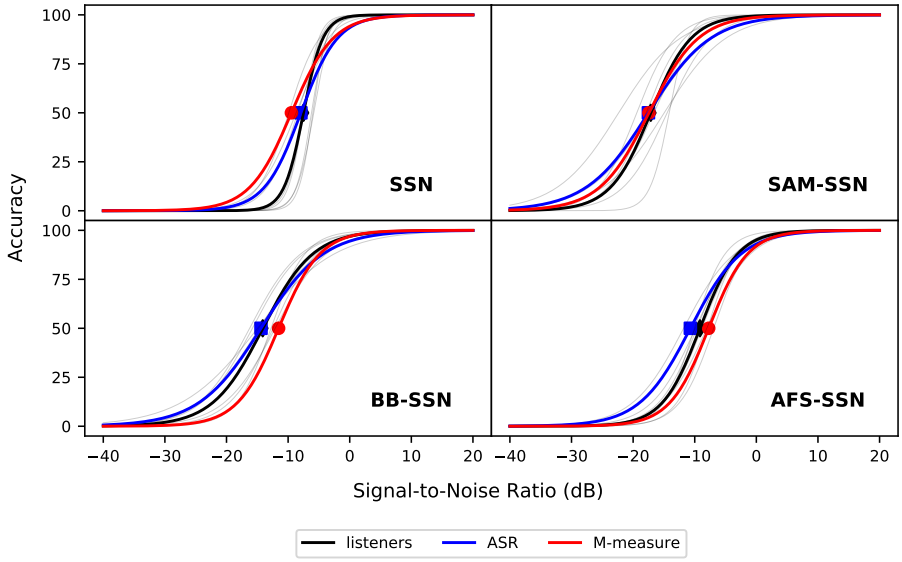

Fig. 3: Psychometric functions or NH listeners (mean in black, individual curves in gray), the ASR-based model that used *a priori* knowledge in the form of transcripts (blue) and the proposed model that estimates the SRT from phoneme probabilities without using the transcript (red).

(A) Average SRTs of listeners and proposed model in dB

	SSN	SAM-SSN	BB-SSN	AFS-SSN
female	-7.5	-17.1	-14.1	-9.2
AMFB-M-Measure	-9.5	-20.0	-14.3	-9.1
male	-8.2	-17.7	-14.9	-9.3
AMFB-M-measure	-10.0	-18.5	-13.0	-7.2

(B) Root-mean-squared error of observed and predicted SRTs

	baseline models				Spille et al	This work
	SII	ESII	STOI	mr-sEPSM	AMFB	AMFB-M-measure
male	6.0	2.2	8.7	4.0	1.6	1.7
female	6.0	2.8	8.5	8.5	0.9	1.7
avg.	6.0	2.5	8.6	6.2	1.3	1.7

Table 1: (A) SRTs of normal-hearing listeners and the corresponding SRT prediction for the proposed ASR-based models. (B) SRT prediction error for baseline models, the label-based previous model and the proposed approach. The rows *female* and *male* correspond to the maskers derived from speech from the female or male speaker, respectively (cf. section on noise types)

These effects, however, seem to be small and provide a good match with the human data (gray lines in the figure); this is also reflected by the RMSE between predicted and observed SRTs, calculated for four baseline models, as shown in Table 1. In Table 1(A), the average SRTs of listeners are compared with the ones yielded by the proposed model, referred to as AMFB-M-measure. The models trained on the female noise maskers are matched to the corresponding female normal-hearing SRTs; likewise, the male results were compared to a model trained on male noise maskers. In both setups, SSN and SAM-SSN, the predicted SNRs were lower than the observed ones, whereas the opposite behavior occurs with the BB-SSN and AFS-SSN noise maskers.

We compute the root-mean-square error (RMSE) between observed and predicted SRTs to measure the precision of the proposed model in all noise conditions shown in Table 1(B). Among the baseline models, ESII yields the lowest error with an average of 2.5. Both DNN-based models show lower RMSE than the previous models. The model from Spille *et al.* (2018), with an average RMSE of 1.3 dB, remains the closest to the human observed SRTs; the female version produces almost half the error as the male counterpart; the same pattern is observed in the mr-sEPSM model. For SII, STOI, and our proposed model, the error difference between the genders is very small.

Note that the DNN-based model was trained as a gender-independent speech with gender-dependent maskers; thus, it produces consistent predictions for both male and female maskers.

DISCUSSION

The proposed model produces accurate predictions while it does not require clean speech reference or the transcript of the utterance that is evaluated. Moreover, because the model was trained on speech data without semantic context, it could potentially generalize to other speech tests. However, in contrast to existing models, it requires a relatively large amount of training data in the range of 80 hours, and it is unclear if predictions can be obtained for SI across languages. It might, therefore, be challenging to apply this approach to low-resource languages without optimizing the training procedure. In related work targeting listening effort, the method of using phoneme probabilities was, however, successful for predicting the listening effort of the German matrix sentence test with fine-tuning using English data (Huber *et al.*, 2018), which indicates that across-language prediction could potentially work if the languages are phonetically not vastly different. An advantage of the proposed model is that it produces absolute predictions for the SRT, again in contrast to the baseline models that are normalized using the prediction for a reference condition, in this case, the stationary SSN.

In the future, the approach could be used as a model-in-the-loop (i.e., it could monitor and estimate the SI resulting from different processing strategies and settings in hearing aids and select the strategy that most likely maximizes SI). However, this would require the prediction of SI for hearing-impaired listeners, while the current model implementation has only been tested for normal-hearing listeners. A simple strategy to take into account the hearing loss that is reflected in a listener's audiogram would be to add frequency-dependent noise to mask the signal properties that are not accessible to the individual listener. Optimally, the corresponding calculations should be carried out on mobile hearing aid hardware. In previous research, we have shown that running at least one feed-forward neural network can be achieved on a hearing aid co-processor (Castro Martinez *et al.*, 2019). However, more efficient net topologies such as time-delay neural networks (Peddinti *et al.*, 2015) need to be considered in the future, as well as taking into account hearing loss, given that a comparison of different processing algorithms requires at least two networks can be used simultaneously.

SUMMARY

This paper explored a modeling approach for SI prediction based on ASR without the requirement of a transcript in the model. It was shown that the model is suitable to predict the SRT of normal-hearing listeners with very similar accuracy to the prediction performance of an ASR-based model that used *a priori* knowledge in the form of transcripts. This achievement was enabled by measuring the degradation of frame-level phoneme representations obtained from a DNN. Our model also

outperforms four established baseline models in four masker types with different types of modulation. Future research should focus on a wider range of maskers and take into account the computational complexity of the approach, which needs to be considered for real-time applications of SI prediction. As the approach was only tested for normal-hearing listeners, we also need to investigate if the model can be extended for predicting SI of (aided) hearing-impaired listeners, which would be a significant step towards using it as model-in-the-loop for real-time optimization in assistive hearing.

ACKNOWLEDGMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2177 - Project ID 390895286 and the SFB/TRR 31/3' The active auditory system,' Transfer Project T01).

REFERENCES

ANSI, S3 22-1997 (**1997**), "Methods for calculation of the speech intelligibility index," American National Standard Institute.

Barker, J. and Cooke, M. (**2006**), "Modelling speaker intelligibility in noise," Speech Commun.

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.

Castro Martinez, A. M., Gerlach, L., Payá-Vayá, G., Hermansky, 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.

Ewert, S. D. and Dau, T. (**2000**), "Characterizing frequency selectivity for envelope fluctuations." J. Acoust. Soc. Am., **108**(3), 1181–96.

Hermansky, H., Variani, E., and Peddinti, V. (**2013**), "Mean temporal distance: Predicting ASR error from temporal properties of speech signal," Proc. IEEE ICASSP, 7423–7426.

Holube, I., Fredelake, S., Vlaming, M., and Kollmeier, B. (**2010**), "Development and analysis of an International Speech Test Signal (ISTS)." Int. J. Audiol., **49**(12), 891–903.

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

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., and Banno, H. (**2008**), "Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation," Proc. IEEE ICASSP, 3933–3936.

Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M. A., Uslar, V., Brand, T., and Wagener, K. C. (**2015**), "The multilingual matrix test: Principles, applications,

and comparison across languages: A review," Int. J. Audiol., **54**(sup2), 3–16.

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

Peddinti, V., Povey, D., and Khudanpur, S. (**2015**), "A time delay neural network architecture for efficient modeling of long temporal contexts," Proc. International Speech Communication Association.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (**2011**), "The Kaldi speech recognition toolkit," Proc. IEEE ASRU.

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.

Schädler, M. R., Warzybok, A., Hochmuth, S., and Kollmeier, B. (**2015**), "Matrix sentence intelligibility prediction using an automatic speech recognition system." Int. J. Audiol., 1–8.

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.

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., Heusdens, R., and Jensen, J. (**2011**), "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. Audio Speech Lang. Process., **19**(7), 2125–2136.

Wagener, K., Brand, T., and Kollmeier, B. (**1999**), "Development and evaluation of a German sentence test part III: Evaluation of the Oldenburg sentence test," Zeitschrift Fur Audiologie, **38**, 86–95.