A simulation framework for auditory discrimination experiments: Revealing the importance of across-frequency processing in speech perception

Marc René Schädler,[a] Anna Warzybok, Stephan D. Ewert, and Birger Kollmeier
Medizinische Physik and Cluster of Excellence Hearing4all, Universität Oldenburg, D-26111 Oldenburg, Germany

(Received 24 July 2015; revised 30 January 2016; accepted 25 April 2016; published online 13 May 2016)

A framework for simulating auditory discrimination experiments, based on an approach from Schädler, Warzybok, Hochmuth, and Kollmeier [(2015). Int. J. Audiol. **54**, 100–107] which was originally designed to predict speech recognition thresholds, is extended to also predict psychoacoustic thresholds. The proposed framework is used to assess the suitability of different auditory-inspired feature sets for a range of auditory discrimination experiments that included psychoacoustic as well as speech recognition experiments in noise. The considered experiments were 2 kHz tone-in-broadband-noise simultaneous masking depending on the tone length, spectral masking with simultaneously presented tone signals and narrow-band noise maskers, and German Matrix sentence test reception threshold in stationary and modulated noise. The employed feature sets included spectro-temporal Gabor filter bank features, Mel-frequency cepstral coefficients, logarithmically scaled Mel-spectrograms, and the internal representation of the Perception Model from Dau, Kollmeier, and Kohlrausch [(1997). J. Acoust. Soc. Am. **102**(5), 2892–2905]. The proposed framework was successfully employed to simulate all experiments with a common parameter set and obtain objective thresholds with less assumptions compared to traditional modeling approaches. Depending on the feature set, the simulated reference-free thresholds were found to agree with—and hence to predict—empirical data from the literature. Across-frequency processing was found to be crucial to accurately model the lower speech reception threshold in modulated noise conditions than in stationary noise conditions.
[http://dx.doi.org/10.1121/1.4948772]

[MSS] Pages: 2708–2722

I. INTRODUCTION

Even though robust automatic speech recognition (ASR) systems have been shown to profit from knowledge about the human auditory system (Hermansky, 1990; Tchorz and Kollmeier, 1999; Kleinschmidt and Gelbart, 2002; Meyer and Kollmeier, 2011; Schädler *et al.*, 2012) and—in return—human auditory signal processing models may profit from the framework and rigid theory behind ASR systems (e.g., Holube and Kollmeier, 1996; Stadler *et al.*, 2007; Jürgens and Brand, 2009) both fields of research have traditionally evolved independently of each other. Typically, any exchange between the two is unidirectional in the sense that (modified) auditory signal processing models are considered as front-ends in ASR, but ASR front-ends are not considered as models of human auditory signal processing. Hence, the aim of this study is to revise the traditional idea of "fitting" auditory models "to the task" in favor of finding universally valid functional models which are able to perform as well as human listeners in a range of auditory recognition tasks. Such an approach should bridge the fields of ASR, human speech recognition, and psychoacoustics research. Compared to many models of human auditory signal processing, which are tailored to describe and model specific properties of the human auditory system, ASR features are subject to an extensive set of broad, sometimes even contradictory, demands, e.g., sufficient spectral/temporal detail but good generalization over acoustic conditions. These different objectives (descriptive model vs universally applicable model) are the reason for auditory models usually requiring considerable modification and engineering towards the appropriate ASR framework before they can be employed as front-ends for ASR purposes. From a modeling point of view, ASR features have desirable properties as a result of the selection process that they undergo in ASR experiments: They are the best known compromise between the diverse demands which are made on the signal representation by robust ASR tasks and, even beyond, audio classification tasks. Hence, auditory-inspired robust ASR features are often simpler than the models by which they were inspired because only the indispensable processing steps for solving the ASR task were actually used. In fact, many common ASR features incorporate only basic auditory signal processing principles such as a limited spectral resolution as well as a compressive intensity perception, e.g., all features which are based on logarithmically scaled Mel-spectrograms (LogMSs). It seems legitimate to ask for the "auditory fidelity" of auditory-inspired ASR features, or in other words, if they show those properties of the auditory system by which

[a]Electronic mail: marc.r.schaedler@uni-oldenburg.de

 0001-4966/2016/139(5)/2708/15/$30.00

they were originally inspired. Hence, one of the aims of this paper is to demonstrate how the "auditory fidelity" of signal representations, including traditional models of auditory signal processing, might be tested and established.

To evaluate ASR features and auditory models on a set of speech recognition and psychoacoustic discrimination tasks with varying complexity and to provide an unbiased, fair comparison between different features/models and empirical data, a common simulation framework which is able to obtain reference-free, i.e., without super-human prior knowledge, objective thresholds is highly desirable. Thus, in a first step, such a framework that allows the simulation of simple and complex auditory discrimination experiments (ADEs) using ASR features as well as the output of auditory models with a single universal parameter set is investigated.

Traditional modeling approaches employ predefined features of the change in the signal to be detected and are typically based on signal-to-noise ratios (SNRs) only, such as the power-spectrum model (Patterson and Moore, 1986), the Speech Intelligibility Index (ANSI, 1997), the envelope lowpass-filter model (Viemeister, 1979), or the envelope-power spectrum model (Ewert and Dau, 2000). The resulting detection threshold corresponds to a predefined feature value which may be formalized, e.g., by the Signal Detection Theory (Green and Swets, 1966). While these models only use long-term features and thus only require statistical representations of signal and noise, some more refined model versions such as the multi-resolution speech-based ESPM (Jørgensen *et al.*, 2013) require reproducible or so-called "frozen" noise to estimate SNRs in short time frames. More sophisticated modeling approaches (Holube and Kollmeier, 1996; Dau *et al.*, 1997; Jepsen *et al.*, 2008; Jürgens and Brand, 2009), perform a pattern match using an "optimal" detector to predict human performance, thus providing an automatic way of finding the appropriate feature(s) to be detected. However, the exact temporal alignment between template and pattern under consideration can only be secured by a "double-ended" approach, i.e., by deriving the template from a prior knowledge of the target signal alone or at a high SNR and a typical representation of the noise. Moreover, this approach is not able to predict plausible thresholds for the outcome of complex ADEs, such as speech intelligibility tests, without requiring an optimal detector that possesses super-human prior knowledge, such as, e.g., the exact temporal alignment of the target or masker signals.

As an alternative, the approach presented by Schädler *et al.* (2015) relieves the strong assumptions about the fixed temporal structure of the template, and hence about knowledge of the to-be-recognized target or noise signal *prior* to mixing, by assuming a training phase of a Hidden-Markov-model-based automatic speech recognizer (ASR) at a broad range of signal-to-noise ratios. During this training phase, the ASR system learns the task on noisy data, just like human listeners are assumed to do during an adaptation phase. Unlike other approaches and like human listeners, the ASR system then needs to infer the temporal alignment of the target signal from the noisy mixture. This can be denoted as a pseudo-single-ended approach which only relies on the knowledge of a probabilistically controlled succession of certain automatically learned features, which natively allows the use of processed signals (e.g., including the effect of noise reduction). Furthermore, this approach is reference-free, since the predicted thresholds are not dependent on any reference condition which is used by some traditional model approaches to fit detection parameters (such as, e.g., internal noise) to the average human performance.

Therefore, the modeling approach from Schädler *et al.* (2015), originally designed to predict the outcome of the German Matrix sentence speech recognition test, was extended to simulate generic ADEs and obtain reference-free objective thresholds. Schädler *et al.* (2015) successfully predicted the outcome of the German Matrix sentence test for different types of background noise by simulating the experiment using a standard ASR system. They trained and tested the ASR system with noisy matrix sentences on a broad range of SNRs and determined the speech reception threshold (SRT), i.e., the SNR at which the recognition rate is 50% correct. In the current study, this approach was extended to recognize tone-in-masker and only masker stimuli which allows one to simulate classical psychoacoustic detection and discrimination experiments. A set of general purpose back-end parameters was established with the aim of allowing the simulation of different experiments using different signal representations with the same parameter set. The extended framework with the general purpose parameters is referred to as the simulation framework for ADEs (FADE). The goal of FADE is to provide a general purpose framework to obtain thresholds which were constrained by the task and the signal representation.

FADE was used to simulate basic, psycho-acoustical experiments and more complex matrix sentence recognition tasks with a range of feature sets (front-ends). On the side of the psycho-acoustical experiments, simultaneous masking thresholds depending on tone duration were included as well as spectral masking thresholds depending on the tone center frequency. On the side of matrix sentence recognition tests, speech reception thresholds (SRTs) of the German Matrix sentence test were included in a stationary and a fluctuating noise condition. As signal representations, LogMSs, standard ASR features, auditory-inspired ASR features, and the output of a traditional "effective" auditory processing model were employed. Mel frequency cepstral coefficient (MFCC) features were used as standard ASR features. The recently proposed Gabor filter bank (GBFB) and separable Gabor filter bank (SGBFB) features, which were shown to improve the robustness of the standard MFCC-based ASR systems of Schädler *et al.* (2012) and Schädler and Kollmeier (2015), encode spectro-temporal modulation patterns of audio signals and were used as auditory-inspired ASR features. The LogMS was also considered as a signal representation because it represents the common basis for MFCC, GBFB, and SGBFB features. The signal representation of the perception model (PEMO) from Dau *et al.* (1997), referred to as PEMO features, represented the output of a traditional auditory signal processing model. ASR features are usually used with feature vector normalization, such as mean and variance normalization (MVN) (Viikki and Laurila, 1998), while signal representations in auditory models are not. To

assess the effect of MVN, LogMS, MFCC, and PEMO features were employed with and without MVN. All considered experiments were simulated using all feature sets and the obtained thresholds were compared to empirical and model data from the literature.

II. METHODS

A. Experiments

The stimuli, the empirical data, and the PEMO model data for the ADEs were taken from the literature (Moore *et al.*, 1998; Derleth and Dau, 2000; Wagener and Brand, 2005; Jepsen *et al.*, 2008). While the model and empirical data from the literature were measured using adaptive methods, the simulations using FADE were performed using a constant-stimulus method which is explained in detail in Sec. II C.

1. Simultaneous masking

The stimuli, the empirical data, and the PEMO model data for the tone-in-noise simultaneous masking experiment were taken from Jepsen *et al.* (2008). There, a 2-kHz tone signal needed to be detected in the presence of a broadband noise masker. The 500-ms Gaussian noise masker was limited to the frequency range from 20 Hz to 5 kHz and included 50-ms raised-cosine ramps. Detection thresholds corresponding to the 70.7%-correct point on the psychometric function were measured for signal duration from 5 to 200 ms which included 2.5-ms raised-cosine ramps.

2. Spectral masking

The stimuli and the empirical data for the tone-in-noise spectral masking experiment were taken from Moore *et al.* (1998). The signal was a tone and the masker a 80-Hz wide Gaussian noise centered at 1 kHz and presented at 45 dB sound pressure level (SPL). Detection thresholds corresponding to the 79.4%-correct point on the psychometric function were measured. The tone frequencies considered in this work were those at which the masking effect was expected to dominate the absolute hearing thresholds: 0.75, 0.90, 1.00, 1.10, 1.25, and 1.50 kHz. The original study considered more conditions including noise signals, tone maskers, additional masker levels, and additional center frequencies. The PEMO model data were taken from Derleth and Dau (2000), which used the same model parameters as Dau *et al.* (1997). In contrast to the original papers, the thresholds are presented in dB SPL rather than in dB masking. Therefore, the dB masking values were transformed to dB SPL using the absolute hearing thresholds defined in (ISO, 2003).

3. German matrix sentence test

The stimuli and the empirical data for the speech intelligibility experiment were taken from Wagener *et al.* (1999) and Wagener and Brand (2005). In the sentence test from Wagener *et al.* (1999), listeners needed to repeat sentences of five words with a fixed syntactical structure which were presented in noise. The SNR which corresponded to the 50%-correct point on the psychometric function, i.e., the SRT, was measured using an adaptive method. The speech material is phonetically balanced and represents the phonetic variety of the German language. In addition to the unmodulated test-specific noise condition, a condition with a single-speaker modulated speech noise from a male speaker at normal level, the IRCA5 noise from Dreschler *et al.* (2001), was considered. The corresponding empirical thresholds were taken from Wagener and Brand (2005).

B. Signal representations

The LogMS is the basis for all considered ASR features (MFCC, GBFB, SGBFB) in this study. Mel-spectrograms were extracted from an amplitude spectrogram of the input waveform with a window length of 25 ms and a window shift of 10 ms. Therefore, the linear frequency axis of the amplitude spectrogram was transformed into a Mel-frequency axis by combining the frequency bins from 64 Hz to 8 kHz with triangular filters into 31 equally spaced Mel-bands. Finally, the amplitude values are compressed with the decade logarithm. An example of a LogMS is depicted in the upper panel of Fig. 1. This 31-dimensional signal representation is referred to as LogMS features.

1. Mel frequency cepstral coefficients

MFCCs are widely used in ASR and acoustic detection tasks and are often used as a baseline. In this work, they were extracted from LogMSs by applying a discrete cosine transform (DCT) in the spectral dimension. Subsequently, the MFCCs corresponding to quefrencies above 0.58 cycles/Mel-band were removed and the remaining 18 MFCCs were concatenated with their first and second order discrete temporal derivative. The temporal derivatives are also called deltas and double deltas and were extracted by applying a slope filter with a total length of 5 frames once or twice,

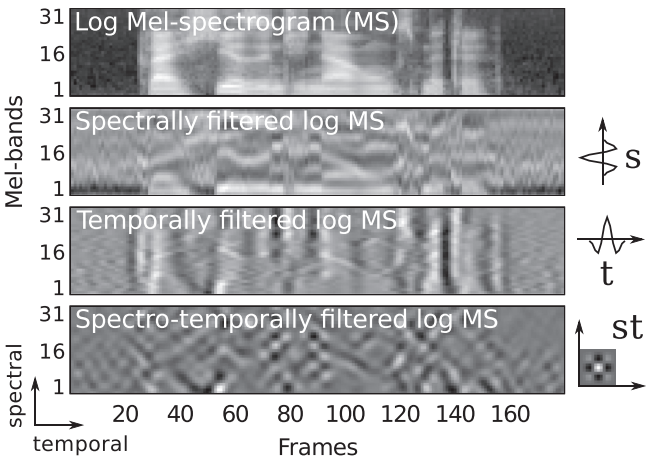

FIG. 1. Taken from Schädler and Kollmeier (2015). The LogMS of clean speech in the upper panel is 2D-convolved with a spectral 1D filter *s*, a temporal 1D filter *t*, and the corresponding spectro-temporal 2D filter *st*. The result of the filtering process is depicted to the left of the corresponding filter. The amplitude of the 2D filters and (filtered) spectrograms is encoded in gray scale, where white encodes high amplitude and black encodes low amplitude.

respectively. The 54-dimensional MFCC features were used with mean and variance normalization as explained in Sec. II B 5.

2. Gabor filter bank features

GBFB features were successfully employed as robust features for ASR by Moritz *et al.* (2013) as well as robust features for acoustic event detection by Schröder *et al.* (2013). They are auditory-inspired and extract spectro-temporal modulation patterns from LogMS using 2D Gabor filters. The shapes of the 2D Gabor filter that were used are depicted in Fig. 2 and were inspired by patterns found in neural correlates in the auditory cortex of cats by Qiu *et al.* (2003). To extract GBFB features, the LogMS was 2D convolved with each of the 2D GBFB filters. Each filtered version was subsequently (critically) down-sampled in spectral dimension by a quarter of the width in spectral dimension of the corresponding 2D filter. The filtered and down-sampled versions of the LogMSs were than concatenated and formed a 455-dimensional feature vector. Extensive descriptions of the GBFB feature extraction were given in Schädler *et al.* (2012) and Schädler and Kollmeier (2015). GBFB features were used with mean and variance normalization as explained in Sec. II B 5.

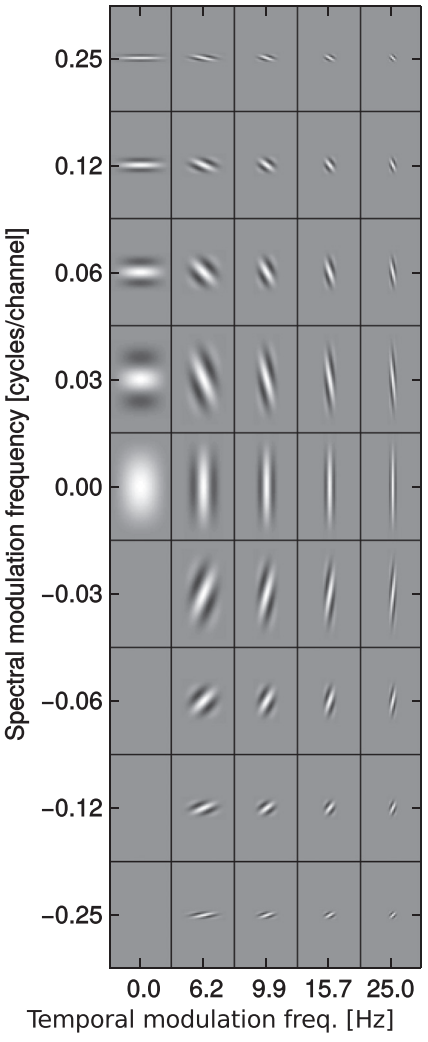

FIG. 2. Taken from Schädler *et al.* (2012). Filter functions of the 2D Gabor filter bank (GBFB) filters. Each tile represents the filter function of a spectro-temporal 2D Gabor filter, where the horizontal axis within each tile is the temporal one and the vertical axis is the spectral one. They are sorted by their spectral and temporal center modulation frequencies. The amplitude of the 2D filter functions is encoded in gray scale, where white encodes high amplitude and black encodes low amplitude.

3. Separable Gabor filter bank features

The difference between GBFB and SGBFB features is that SGBFB features are extracted with two separate modulation filter banks, a spectral and a temporal one, instead of using a filter bank of spectro-temporal filters. Nonetheless, they cover the same spectro-temporal modulation space. The SGBFB approach was shown to reduce the complexity of the features and even to improve the robustness of an ASR system (Schädler and Kollmeier, 2015). All SGBFB filter functions and the corresponding separable 2D filter functions of all combinations of spectral and temporal SGBFB filters are depicted in Fig. 3. In the current study 1020-dimensional SGBFB features were extracted using the full set, i.e., all nine spectral and all nine temporal filters, which are referred to as SGBFB features. An extensive description of the SGBFB feature extraction was given in Schädler and Kollmeier (2015). In addition to the 1020-dimensional SGBFB features, a reduced set of 255-dimensional SGBFB features which does not use the filters that are marked with I (for imaginary phase) in Fig. 3, were considered and are referred to as SGFB-RR features. Due to its design, the

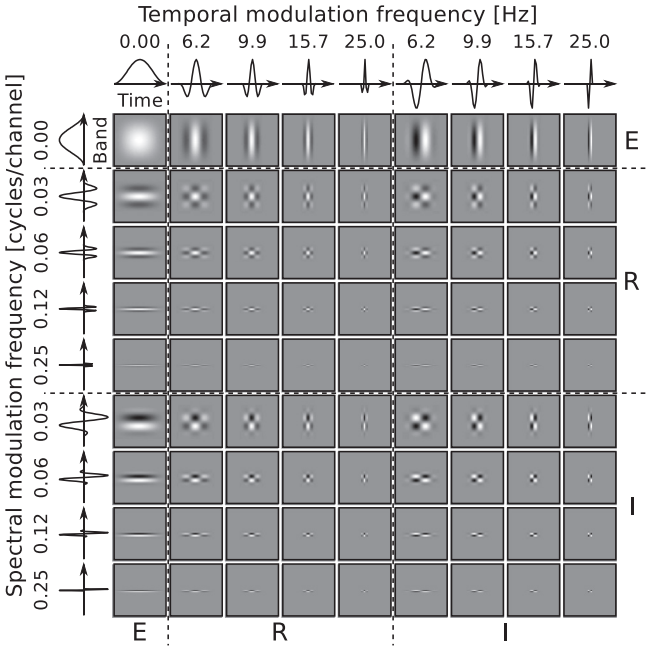

FIG. 3. Taken from Schädler and Kollmeier (2015). All possible combinations of spectral and temporal 1D Gabor filter bank filters, and their equivalent, separable spectro-temporal 2D filter functions. Each tile represents a separable spectro-temporal 2D filter function, with the horizontal axis within each tile being the temporal and the vertical axis being the spectral one. The 1D filters, depicted above and to the left of the 2D filters, are sorted by spectral and temporal center modulation frequencies, and are grouped according to the part of the complex 1D Gabor filter which is used: envelope (E), real (R), imaginary (I). The amplitude of the 2D filters is encoded in gray scale, where white encodes high amplitude and black encodes low amplitude.

SGBFB allows one to apply only the spectral or only the temporal modulation filtering. A set of features which was extracted using only temporal R (for real phase) and E (for envelope) filters is referred to as SGBFB-R-T, and another set of features which was extracted using only spectral R and E filters is referred to as SGBFB-R-S. All SGBFB based features were used with mean and variance normalization as explained in Sec. II B 5.

4. Perception model

The PEMO was successfully used to model various experiments in psychoacoustics (e.g., Dau *et al.*, 1997; Verhey *et al.*, 1999; Derleth and Dau, 2000). It was introduced by Dau *et al.* (1996a,b) and later extended with a temporal modulation filter bank by Dau *et al.* (1997). The PEMO includes a signal processing part (front-end) which effectively models several aspects of the human auditory system. In the current study the PEMO front-end from Dau *et al.* (1997) is used to extract features from input waveforms. Therefore, the freely available implementation from Søndergaard and Majdak (2013) at git commit *cc9c0d3c* was used, which considers auditory filters in the frequency range from 80 Hz to 8 kHz and temporal modulation frequencies in the range from 0 to 150 Hz. A block-diagram of the PEMO feature extraction is depicted in Fig. 4. A Gammatone filter bank was used to model the response of the basilar membrane to the input signal. The subsequently applied half-wave rectification and the 1 kHz low-pass filter model the hair cell deflection. The adaptation loops account for temporal properties of the nerve cell firing probability at different stages of the auditory pathway. The output of the modulation filter bank was low-pass filtered and down-sampled to 100 Hz. These 275-dimensional feature set is referred to as PEMO features.

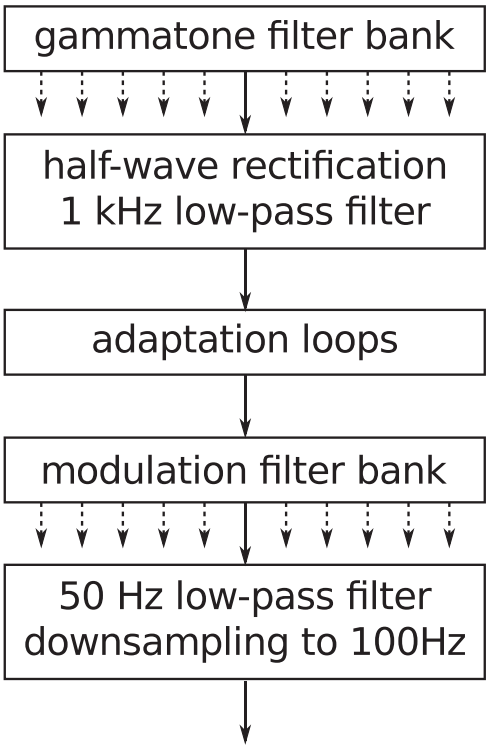

FIG. 4. Modified from Dau *et al.* (1997). Block diagram of the auditory signal processing which is used to calculate the internal representations with PEMO, also referred to as PEMO features. Essentially this is the model from Dau *et al.* (1997) up to the modulation filter bank. The low-pass filtering at 50 Hz and the down-sampling to 100 Hz is added to make the internal representation compatible for the use in the recognition system.

5. Feature normalization

Feature vector normalization such as mean and variance normalization (MVN) or histogram equalization were shown to improve the robustness of ASR systems (Viikki and Laurila, 1998; De La Torre *et al.*, 2005) and are usually employed in conjunction with robust ASR features. The auditory models used to explain psycho-acoustical experiments usually do not contain a similar processing step. This is why in the current study by default all ASR features (MFCC, GBFB, and SGBFB) are used with per-utterance/per-stimulus MVN, while the LogMS and the PEMO features are not.

In order to assess the effect of feature normalization, LogMS, PEMO, and MFCC features were tested with and without MVN. These feature sets are indicated by the suffix MVN and NOMVN, respectively.

C. Simulation framework for ADEs

The simulation FADE is based on the approach from Schädler *et al.* (2015), where an ASR recognition system was used to simulate—and hence predict the outcome of—the German Matrix sentence test with only few assumptions compared to traditional speech intelligibility prediction models. Here, this approach was extended to simulate tone-in-noise detection (i.e., tone-in-noise from only noise discrimination) experiments. A reference implementation of FADE is available online (FADE, 2016).

1. Front-end

In the original work by Schädler *et al.* (2015), only MFCCs were used as the front-end, while in this work all signal representations presented in Sec. II B were employed with FADE.

2. Back-end

The back-end used in FADE is the same as in Schädler *et al.* (2015). HTK was used to build left-to-right whole-word/stimulus Hidden Markov Models (HMMs) models with six states per word/stimulus and Gaussian Mixture Models (GMMs) with one component per state. For each training condition, which in the case of the German Matrix sentence test is determined by the SNR and for the psycho-acoustic experiments by the absolute tone level, the GMM/HMM parameters are estimated (learned) in a total of eight iterations. Since the material of the German Matrix test consists of 50 words, 50 whole-word models were learned during the training period. For the tone-in-noise detection experiments, two models were trained: A model for the stimuli in which the target is present (tone plus noise) and a reference one for the stimuli in which the target is absent

(noise only). In addition to the word/stimuli models, a START, a STOP, a PRE-SILENCE, and a POST-SILENCE model were trained for each training condition. The START/STOP model covers border artifacts which are common to all recordings of a training condition, while the PRE/POST-SILENCE models represent the indistinguishable signal parts before and after the speech/target. All four are shared between all sentences/stimuli of a training condition. The grammar, in HTK-terms, for a sentence/stimulus was (START PRE-SILENCE $sentence/stimulus POST-SILENCE STOP), where $sentence = ($word1 $word2 $word3 $word4 $word5) and $stimulus = (reference | target). The corresponding grammar was converted to a word network and used to limit the recognizer to search only for transcriptions with valid syntax for the corresponding experiment. This implements the knowledge of a trained listener, who knows about the grammatical structure as well as about the limited vocabulary of the matrix test. The effect of the number states per model and the number of states per special model (START, STOP, PRE-SILENCE, POST-SILENCE), and the number of training iterations was assessed in Sec. III D.

3. Simulation

The regions of interest of the values for the independent variables were defined as follows: For the simultaneous masking experiment, tone levels from +45 to +75 dB SPL in 5-dB steps were considered. For the spectral masking experiment, tone levels from −10 to +50 dB SPL in 5-dB steps were considered. For the German Matrix sentence experiments, SNRs from −24 to +6 dB in 3-dB steps were considered.

For each of these values, datasets for training and testing were generated in the same manner. For the tone-in-noise masking experiments, the two different types of stimuli (target and reference) were generated with random noise, such that a repetition of the same stimulus waveform is practically impossible. For the German Matrix sentence experiments, the 120 available sentences were mixed with the noise signal with random temporal offsets, such that a repetition of the same waveform is practically impossible even if the same sentences were mixed several times with the same noise signal. The 120 sentences contained each word of the 50-word vocabulary exactly twelve times, and mixing all sentences once with random portions of the noise signal resulted in twelve samples per word.

From these (statistical) *pools*, which directly reflect the difficulty of the corresponding recognition task at a given tone level or SNR, a number of samples was drawn and declared as the test data. Because the performance limiting factor, i.e., the difficulty of the task, is inherent to the test data under its projection into the feature space, an optimal training data set was desirable. Hence, the training data sets were drawn from the same *pool* as—but separate from—the test data sets. By this means, we aimed to minimize the influence of the training data set and at the same time to maximize the influence of the test data set on the recognition scores.

The recognition of all 120 available sentences of the German Matrix test produces 600 binary (correct or incorrect) decisions, which was chosen to be the size of the test data sets. It should be noted that each Matrix sentence results in five binary decisions, one for each word, while a presented psychoacoustic stimulus only results in one binary decision. The size of the training data sets of 96 samples for each word/stimulus was assessed in Sec. II C 4. For the matrix sentence test, these were achieved by mixing all sentences eight times with random portions of the noise signal. Features were then extracted from the generated training and test data sets.

For each condition (e.g., speech in fluctuating noise) separately, models were trained and tested for all considered values of the independent variable. For example, 11 models—one for each considered SNR—were trained on speech in fluctuating noise and each subsequently tested in the 11 considered SNR conditions, which resulted in $11 \times 11 = 121$ recognition scores. These were represented by a (square) matrix called "recognition result map" (RRM), where each row represents a psychometric function of which the value of the independent variable at a given target threshold could be derived. For the matrix sentence test, the SNR at 50%-correct, which is the standard procedure with human listeners, was determined. For the psychoacoustic experiments, instead of the 50%-correct point on the psychometric function (i.e., the SRT) the corresponding target %-correct point was considered. For each psychometric function, the value of the independent variable at the corresponding target %-correct point was interpolated together with its estimated uncertainty due to the size of test data set. Thus, for the tone-in-noise experiments, several levels at threshold depending on the training level, and for the German Matrix sentence test, several SRTs depending on the training SNR were available. As the result of the simulation, the lowest value at threshold was reported, where two standard deviations of margin were considered in order to report the outcome with the lowest 95th percentile (assuming normal distributions). This automatic determination of the optimal training data set, which may depend on the task itself, the amount of training data and the feature representation, is aimed to reduce its influence on the simulation results.

4. General purpose parameter set

At the core of FADE a set of general purpose parameters exists which was employed for all features and experiments, the simplest task being the detection of a tone, the most difficult the discrimination of words in modulated noise, the lowest-dimensional features being 31-dimensional LogMS features and the highest-dimensional being the 1020-dimensional SGBFB features. These parameters were

- HMM states START/STOP: 6.
- HMM states per model: 6.
- Training samples per model: 96.
- Training iterations: 8.

These parameters were considered to be especially important when differently complex features and tasks are

involved. To demonstrate that the chosen parameter values are optimal up to ±1 dB for different features in differently complex experiments, a set of features and experiments was performed when varying the parameter values. Optimal here means that the systems obtained the highest possible recognition rates which translates to the lowest possible thresholds, an optimization scheme commonly used in the field of ASR. Optimal here does not mean that the results were close to the empirical results, which is an optimization scheme commonly used in the field of psychoacoustic modelling. The considered values were

- HMM states START/STOP: 1, 2, 3, 4, **6**, 8, 12, 16, 24.
- HMM states per model: 1, 2, 3, 4, **6**, 8, 12, 16, 24.
- Training samples per model: 12, 24, 48, **96**, 192, 384.
- Training iterations: 1, 2, 3, 4, 6, **8**, 12, 16, 24.

Each of the parameters was varied while the others were left unchanged. Simulations of the simultaneous masking experiment and the German Matrix sentence test in the test-specific noise condition were performed with varied parameter values using MFCC and PEMO features.

5. Uncertainty calculation

The uncertainty of the simulated outcomes due to the limited test data, which was 600 binary decisions per condition, was estimated using bootstrapping. It turned out to be about 2.1 percentage points (pp) at 50% correct, about 1.8 pp at 75% correct, and about 1.2 pp at 90% correct. These estimated uncertainties were assumed to be normally distributed and propagated to derived values, such as SRTs or thresholds, where possible. The uncertainty due to the limited test data was not assessed as it would have required re-running the training stage several times with different data. In addition, the limited step size of training and test conditions could present another source of uncertainty, which was not assessed either. Hence, the uncertainties reported here only include those due to the limited test data and should be considered orientative. Nonetheless, the uncertainty can be assumed to be about 1 dB, which was justified in Sec. II C 4.

III. RESULTS

Apart from the parameter variation experiment, all simulations were performed with all features. The results are presented in tables and selected results are additionally plotted.

A. Simultaneous masking

Figure 5 depicts the simulated detection thresholds depending on the tone duration with PEMO, MFCC, and SGBFB features alongside the empirical results and PEMO model results from the literature (Jepsen *et al.*, 2008). Table I reports the corresponding results in numerical form for all considered feature sets, and in addition, the average detection threshold over all conditions.

FADE was able to predict detection thresholds for a simultaneous masking experiment with a variety of front-ends.

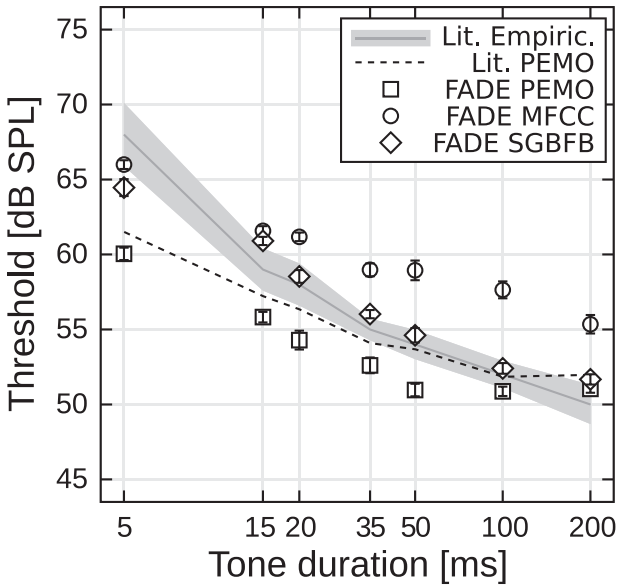

FIG. 5. Simulated detection thresholds for the simultaneous masking experiment depending on the tone duration with PEMO, MFCC, and SGBFB features alongside the empirical data and PEMO data from the literature Jepsen *et al.* (2008). The gray area indicates the 1-sigma uncertainty of the empirical data.

All simulated thresholds were consistent with the empirical thresholds within ±10 dB, i.e., in the correct order of magnitude. MFCC features resulted in the most pronounced over-estimation of the empirical thresholds, with an average detection threshold of 60.0 ± 0.2 dB SPL and PEMO features resulted in the most pronounced under-estimation of the empirical thresholds with an average detection threshold of 53.7 ± 0.2 dB SPL, while the empirical results showed an average detection level of 56.6 ± 0.5 dB SPL. Simulation results with all other features lay between simulation results with MFCC and PEMO features. The simulated thresholds with GBFB based features (GBFB, SGBFB, SGBFB-RR, SGBFB-R-S, SGBFB-R-T) were consistently found to be close to the empirical thresholds.

The simulated thresholds with PEMO features are generally about 2 dB lower than the PEMO data from the literature, over-estimating the empirical thresholds for tone-durations shorter than 100 ms. The simulated thresholds with MFCC features under-estimate the empirical thresholds for tone-durations longer than 15 ms. The simulated thresholds with SGBFB feature resemble the empirical thresholds remarkably well. Deviations of simulated thresholds from the empirical data are further analyzed in Sec. III E.

B. Spectral masking

Figure 6 depicts the simulated detection thresholds depending on the tone center frequency in Hz with PEMO, MFCC, and SGBFB features alongside the empirical results and PEMO model results from the literature (Moore *et al.*, 1998; Derleth and Dau, 2000). Table II reports the corresponding results in numerical form for all considered feature sets, and in addition, the calculated 20-dB-bandwidth.

FADE was able to predict detection thresholds for the spectral masking experiment with a variety of front-ends.

TABLE I. Simulated detection thresholds in dB SPL for the simultaneous masking experiment depending on the tone duration and the feature set. The empirical data were taken from Jepsen *et al*. (2008).

	Tone duration							
Features	5 ms	10 ms	15 ms	35 ms	50 ms	100 ms	200 ms	Average
Empirical	68.0 ± 2.1	59.0 ± 1.4	58.0 ± 1.4	55.0 ± 0.7	54.0 ± 1.0	52.0 ± 0.9	50.0 ± 1.3	56.6 ± 0.5
LogMS	63.0 ± 0.3	60.6 ± 0.7	58.4 ± 0.5	56.9 ± 0.2	54.0 ± 0.5	50.7 ± 0.3	48.2 ± 0.5	56.0 ± 0.2
LogMS-MVN	63.7 ± 0.4	61.6 ± 0.2	61.5 ± 0.2	56.8 ± 0.3	57.5 ± 0.3	52.5 ± 0.5	53.1 ± 0.5	58.1 ± 0.1
MFCC	66.0 ± 0.3	61.6 ± 0.3	61.2 ± 0.3	59.0 ± 0.4	58.9 ± 0.7	57.6 ± 0.6	55.4 ± 0.6	60.0 ± 0.2
MFCC-NOMVN	66.0 ± 0.3	61.3 ± 0.3	61.3 ± 0.3	58.5 ± 0.4	58.3 ± 0.5	56.2 ± 0.4	53.9 ± 0.5	59.4 ± 0.2
GBFB	64.8 ± 0.5	60.8 ± 0.3	59.0 ± 0.5	56.5 ± 0.3	56.1 ± 0.3	52.6 ± 0.4	51.9 ± 0.5	57.4 ± 0.2
SGBFB	64.5 ± 0.6	60.9 ± 0.3	58.5 ± 0.4	56.0 ± 0.3	54.6 ± 0.5	52.4 ± 0.4	51.7 ± 0.3	56.9 ± 0.2
SGBFB-RR	65.5 ± 0.3	61.2 ± 0.3	60.4 ± 0.4	56.8 ± 0.3	55.8 ± 0.4	52.1 ± 0.4	51.9 ± 0.4	57.7 ± 0.1
SGBFB-R-S	64.3 ± 0.5	61.8 ± 0.2	61.4 ± 0.2	57.5 ± 0.5	56.7 ± 0.3	52.7 ± 0.3	52.5 ± 0.6	58.1 ± 0.2
SGBFB-R-T	65.1 ± 0.5	60.5 ± 0.3	60.1 ± 0.4	56.0 ± 0.3	55.9 ± 0.4	54.0 ± 0.6	52.1 ± 0.6	57.7 ± 0.2
PEMO	60.0 ± 0.4	55.8 ± 0.3	54.3 ± 0.6	52.6 ± 0.5	51.0 ± 0.4	50.9 ± 0.3	51.0 ± 0.3	53.7 ± 0.2
PEMO-MVN	60.2 ± 0.4	56.0 ± 0.4	54.5 ± 0.6	52.8 ± 0.5	51.2 ± 0.4	51.2 ± 0.3	51.9 ± 0.3	54.0 ± 0.2

Almost all simulated thresholds are within ±10 dB of the empirical thresholds, i.e., in the correct order of magnitude. Only the PEMO and LogMS features with MVN resulted in thresholds outside that range. Generally, the simulations with all features show the highest thresholds at the noise center frequency of 1000 Hz and decrease as the tone frequency increases or decreases. Consistent with the results from the simultaneous masking experiment, the simulated thresholds with MFCC features exhibit the highest on-masker (1000 Hz) thresholds with 47.4 ± 0.2 dB SPL and the simulated thresholds with PEMO features, with 43.2 ± 0.3 dB SPL, one of the lower thresholds. These are—unlike in the simultaneous masking experiment—higher than the empirical threshold, which was 40.9 ± 6.4 dB SPL. The empirical 20-dB-bandwidth was calculated to be 229.5 ± 38.9 Hz.

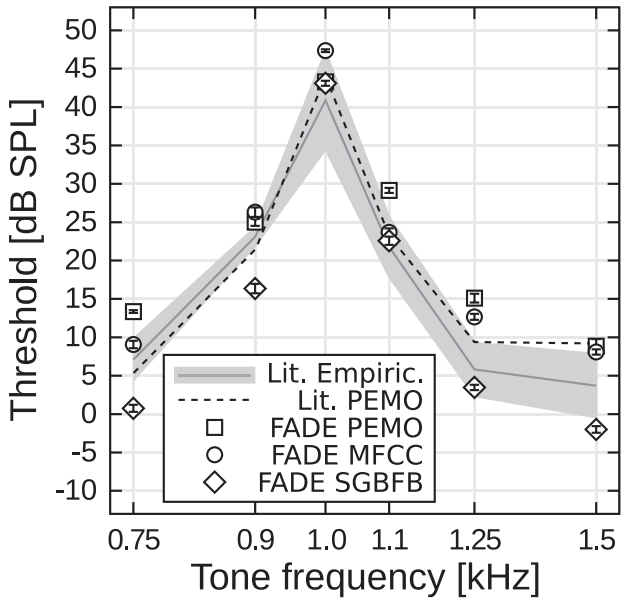

FIG. 6. Simulated detection thresholds for the spectral masking experiment depending on the tone center frequency in Hz with PEMO, MFCC, and SGBFB features alongside the empirical data and PEMO data from the literature (Moore *et al*., 1998; Derleth and Dau, 2000). The gray area indicates the 1-sigma uncertainty of the empirical data.

Almost all simulated results fell into the 2-sigma range (151.3 to 306.7 Hz) and hence did not differ significantly from the empirically derived bandwidth. Only the PEMO features with MVN exceeded this range with a bandwidth of 396.2 ± 5.8 Hz. All ASR features (MFCC, GBFB, and SGBFB) result in rather narrow bandwidths around 180 Hz, e.g., using SGBFB features, 172.3 ± 3.3 Hz.

The simulated thresholds with PEMO features were found to be similar or higher than the PEMO model data reported by Derleth and Dau (2000). With MFCC features, the simulated thresholds resembled the empirical data well while with SGBFB features the thresholds on the low frequency flank were over-estimated by about 6 dB. Deviations of simulated thresholds from the empirical data are further analyzed in Sec. III E.

C. German matrix sentence test

The recognition result map, which is the matrix that contains the recognition rates depending on the training and the test condition, and its evaluation is illustrated in Fig. 7 for the simulation results of the German Matrix sentence test with MFCC features. In panel (A), the RRM, i.e., the recognition performance depending on the training and test SNR, is depicted in gray-scale, where black corresponds to 0%-correct and white to 100%-correct. The iso-50%-correct contour is indicated by the dotted black-and-white line and the lowest achievable SRT, which at the same time is the simulation result, is indicated by a circle. The corresponding training condition (–3 dB SNR) is marked with a dash-dotted line and the corresponding psychometric function is depicted in panel (B). As expected, the recognition results are at chance level (10%-correct) at low SNRs and tend towards 100%-correct for high SNRs.

Figure 8 depicts the simulated SRTs depending on the noise condition and the employed feature set alongside the empirical results from the literature (Wagener *et al*., 1999; Wagener and Brand, 2005). Table III reports the corresponding results in numerical form for all considered feature sets and, in addition, the effect of modulation which is reported

TABLE II. Simulated detection thresholds in dB SPL for the spectral masking experiment depending on the tone center frequency in Hz and the feature set. The full widths were calculated at − 20 dB from the data. The empirical data were taken from Moore *et al*. (1998).

	Tone center frequency						
Features	750 Hz	900 Hz	1000 Hz	1100 Hz	1250 Hz	1500 Hz	Width [Hz]
Empirical	7.1 ± 2.7	23.1 ± 1.2	40.9 ± 6.4	21.8 ± 3.9	5.8 ± 3.5	3.7 ± 4.1	229.5 ± 38.9
LogMS	3.8 ± 0.2	15.6 ± 0.4	42.1 ± 0.3	24.6 ± 0.4	2.1 ± 0.5	−1.1 ± 0.4	192.3 ± 2.9
LogMS-MVN	5.1 ± 1.0	24.4 ± 0.8	47.3 ± 0.2	42.6 ± 0.4	4.2 ± 0.5	2.2 ± 0.6	246.6 ± 3.3
MFCC	9.1 ± 0.5	26.3 ± 0.6	47.4 ± 0.2	23.7 ± 0.5	12.7 ± 0.4	8.1 ± 0.3	179.6 ± 3.4
MFCC-NOMVN	1.3 ± 0.3	16.6 ± 0.3	43.0 ± 0.3	21.5 ± 0.4	2.8 ± 0.3	−2.7 ± 0.4	168.7 ± 2.1
GBFB	2.4 ± 0.4	18.6 ± 0.6	43.3 ± 0.3	23.2 ± 0.4	5.4 ± 0.6	−1.2 ± 0.4	180.4 ± 3.0
SGBFB	0.8 ± 0.5	16.4 ± 0.6	43.1 ± 0.3	22.6 ± 0.5	3.5 ± 0.3	−2.0 ± 0.4	172.3 ± 3.3
SGBFB-RR	1.7 ± 0.4	16.4 ± 0.5	43.0 ± 0.4	21.6 ± 0.6	3.4 ± 0.4	−1.9 ± 0.4	168.3 ± 3.3
SGBFB-R-S	5.1 ± 0.9	21.0 ± 0.8	43.2 ± 0.4	29.9 ± 0.5	4.5 ± 0.6	−1.3 ± 0.5	229.4 ± 4.7
SGBFB-R-T	4.2 ± 0.9	22.0 ± 0.6	47.2 ± 0.2	28.2 ± 0.7	7.2 ± 1.1	2.7 ± 0.7	186.0 ± 5.2
PEMO	13.3 ± 0.2	25.0 ± 0.5	43.2 ± 0.3	29.1 ± 0.3	15.1 ± 0.6	8.8 ± 0.9	284.8 ± 7.3
PEMO-MVN	16.5 ± 0.2	35.7 ± 0.3	44.6 ± 0.4	37.4 ± 0.5	19.9 ± 0.7	13.2 ± 0.2	396.2 ± 5.8

as the difference of the SRT in the modulated noise condition (ICRA5) and the test-specific noise condition (Olnoise).

For the stationary noise condition, the simulated SRTs were found to be in the range from −8.2 to −6.7 dB SNR, where the empirical value measured by Wagener *et al*. (1999) was −7.1±0.8 dB SNR. Hence, the stationary noise condition was well predicted by simulations with all features. Using GBFB features resulted in the lowest simulation results (−8.2±0.1 dB SNR). For the modulated noise, the picture changes considerably. Simulated SRTs ranged from −19 to 0 dB SNR depending on the employed feature set, where the empirical values measured by Wagener and Brand (2005) were on average −21.6±2.0 dB SNR. The lowest simulation results and hence, those closest to the empirical data, were obtained with GBFB and SGBFB features, with −16.2±0.5 dB SNR and −19.0±0.4 dB SNR, respectively, followed by MFCC features with −15.0±0.7 dB SNR. At the far end of the range, the use of LogMS and PEMO features resulted in simulated SRTs higher than in the respective stationary condition with −0.5±0.3 dB and −3.7±0.3 dB SNR, respectively.

The effect of modulation, which was defined as the difference in dB between the modulated (IRCA5) and the stationary (Olnoise) noise condition, was found to be −14.5±2.2 dB for the empirical data. This means that for listeners with normal hearing it was much easier to recognize speech in the modulated noise condition than in stationary noise condition. Comparing the modulation effect with the LogMS feature set (+6.3±0.3 dB), which performed no modulation processing, the SGBFB-R-T feature set (+6.2±0.3 dB), which only performed the temporal modulation filtering, the SGBFB-R-S feature set (−6.7±0.5 dB), which only performed the spectral modulation filtering, and the SGBFB-RR feature set (−10.8±0.4 dB), which performed both, shows that spectral modulation processing alone accounts for the major part of the modulation effect and that temporal filtering alone has no effect. Deviations of

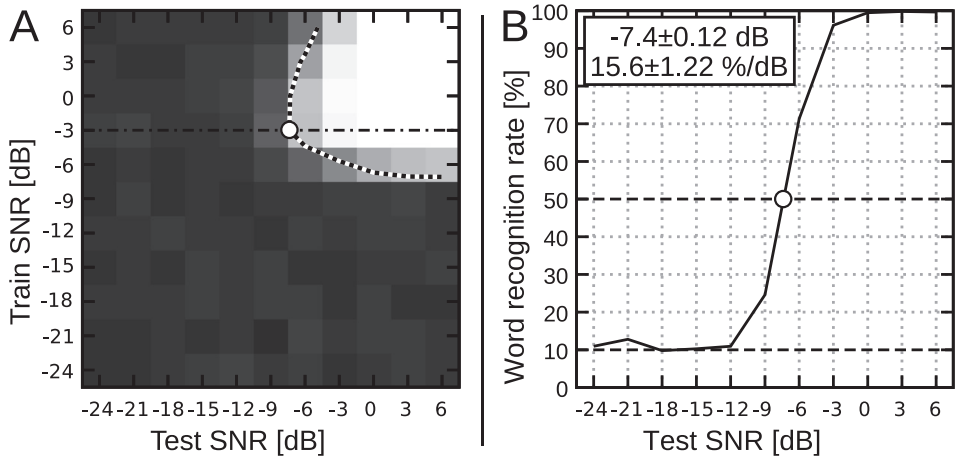

FIG. 7. (A) Recognition result map (RRM) for the test-specific noise condition with MFCC features. The obtained recognition performance is plotted depending on the training and testing SNR. The word recognition rates are encoded in gray-scale, with white representing 100% correct and black 0% correct. The dotted black-and-white line marks the iso-50%-correct contour. The dash-dotted line marks the training SNR which resulted in the lowest achievable test SNR at 50%-correct WRR (SRT). The white circle indicates the predicted SRT. (B) Word recognition rates depending on the test SNR for the system that achieves the lowest SRT [cf. the dash-dotted line in (A)]. The chance level (10%) and the 50%-threshold are marked with dashed lines. The white circle indicates the simulated SRT. The box shows the estimated SRT and slope of the psychometric function, respectively.

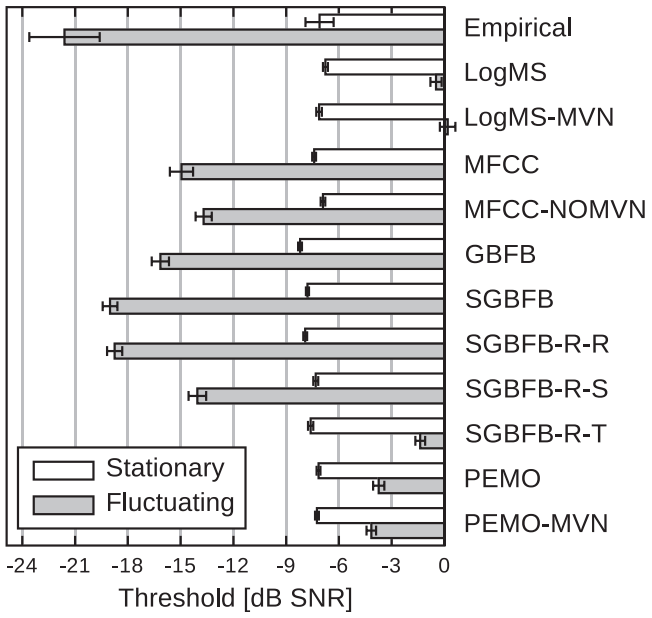

FIG. 8. Simulated SRTs in dB SNR for the German Matrix sentence test depending on the noise condition and the feature set. The empirical data were taken from Wagener *et al.* (1999) (stationary) and Wagener and Brand (2005) (fluctuating).

simulated thresholds from the empirical data are further analyzed in Sec. III E.

D. Effect of back-end parameter variations

The simulation results with varied back-end parameters are depicted in Fig. 9. For the simultaneous masking experiment the average simulated thresholds and for the German Matrix sentence test the simulated SRTs are plotted depending on the varied back-end parameters for MFCC and PEMO features.

Generally, the smallest parameter value from the range of considered values which resulted in the lowest thresholds ±1 dB was chosen if no reason existed not to do so. While for the matrix sentence test, the words were best modeled with HMMs with six emitting states, for the simultaneous masking experiment it was sufficient to use HMMs with a

TABLE III. Simulated SRTs for the German Matrix sentence test depending on the noise condition and the feature set in dB SNR. The empirical data were taken from Wagener *et al.* (1999) (Olnoise) and Wagener and Brand (2005) (ICRA5). The effect of modulation is reported as the difference of the SRT in the modulated noise condition (ICRA5) and the test-specific noise condition (Olnoise).

System	Olnoise SRT [dB]	ICRA5 SRT [dB]	Modulation effect [dB]
Empirical	−7.1 ± 0.8	−21.6 ± 2.0	−14.5 ± 2.2
LogMS	−6.8 ± 0.1	−0.5 ± 0.3	+6.3 ± 0.4
LogMS-MVN	−7.1 ± 0.2	+0.2 ± 0.4	+7.3 ± 0.5
MFCC	−7.4 ± 0.1	−15.0 ± 0.7	−7.5 ± 0.7
MFCC-NOMVN	−6.9 ± 0.1	−13.7 ± 0.5	−6.8 ± 0.5
GBFB	−8.2 ± 0.1	−16.2 ± 0.5	−7.9 ± 0.5
SGBFB	−7.8 ± 0.1	−19.0 ± 0.4	−11.2 ± 0.4
SGBFB-RR	−7.9 ± 0.1	−18.8 ± 0.4	−10.8 ± 0.4
SGBFB-R-S	−7.3 ± 0.1	−14.1 ± 0.5	−6.7 ± 0.5
SGBFB-R-T	−7.6 ± 0.2	−1.4 ± 0.3	+6.2 ± 0.3
PEMO	−7.2 ± 0.1	−3.7 ± 0.3	+3.4 ± 0.3
PEMO-MVN	−7.3 ± 0.1	−4.2 ± 0.3	+3.1 ± 0.3

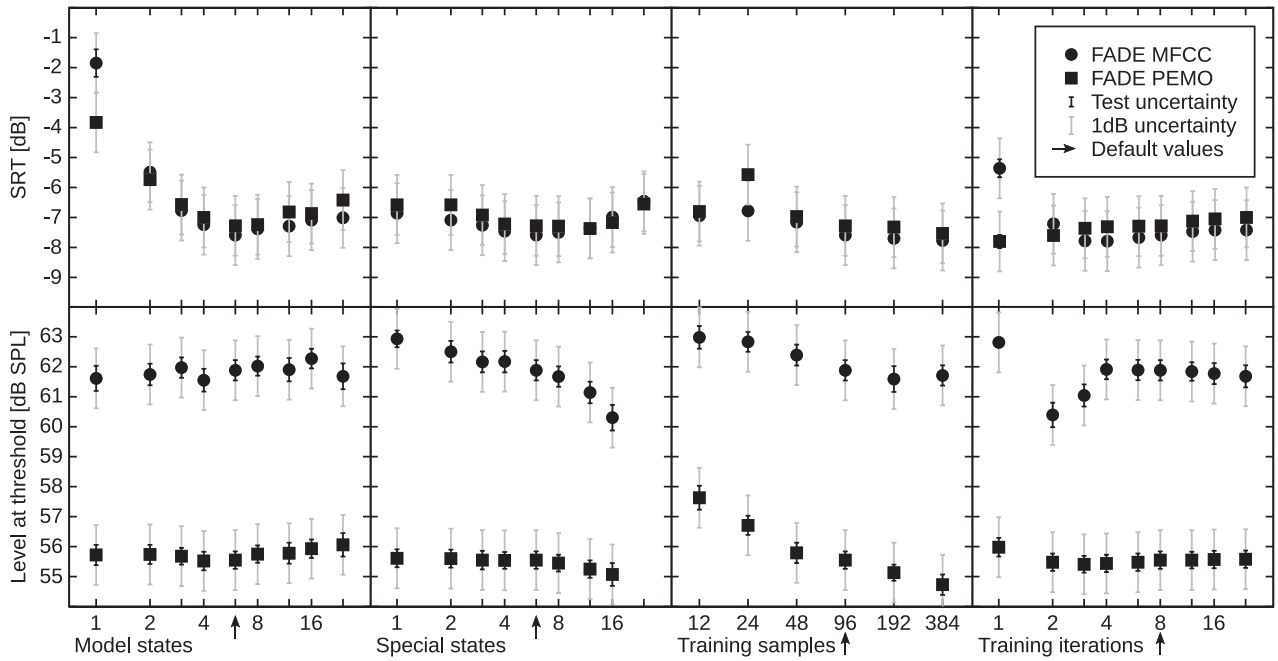

FIG. 9. Predicted SRTs (upper row) and tone detection thresholds in noise (lower row) from the back-end parameter variation experiment in dB SPL and dB SNR. The number of states per model, the number of states of the special models (START and STOP), the number of training samples per model, and the number of training iterations were varied over wide ranges of possible values. The predicted thresholds for the Matrix test in the test-specific noise condition and the average predicted thresholds for the tone-in-noise detection thresholds are plotted depending on the altered parameter values for both considered front-ends. The circles and squares indicate the results when using the MFCC and PEMO front-end, respectively. The arrows indicate the default parameter values. The small (partly hidden behind the markers) black error bars indicate the uncertainty due to finite number of testing samples, and the larger, gray error bars indicate the target precision of ±1 dB.

single emitting state. The special states (START and STOP) were chosen according to the simulation results from the German Matrix sentence test because for the simultaneous masking experiment long special states (>6 states) effectively narrowed the region to search for the target tone and hence improved the thresholds in an unwanted manner. Hence, the border effects were modeled best with HMMs with six emitting states. Reducing the amount of training data resulted in higher simulated thresholds, while increasing the amount of training data did not result in improvements of simulated thresholds exceeding 1 dB. It should be noted that the number of training samples per model guaranteed that each mean and each variance in the GMM was estimated from at least 96 samples, which was only the case if the corresponding HMM state occupied only one frame, i.e., the shortest possible duration of an HMM state. The number of training iterations was sufficient, with a security margin of factor 2, for all models to converge during the training procedure.

E. Man-machine gap

To get a comprehensive overview of the model fidelity depending on the employed feature set and experiment, the maximum and minimum differences from the empirical data are reported in Fig. 10. While negative values indicate an over-estimation of the empirical thresholds, positive values indicate an under-estimation. It should be noted that for human listeners no significant difference was found if the German Matrix test was presented in a closed-set or open response format (Warzybok *et al.*, 2015). The over all maximum can be interpreted as the remaining (unexplained) gap between human performance and machine performance. In this regard, the German Matrix sentence test in the modulated noise condition was the decisive condition, or very near (<1 dB) to the decisive condition, for all feature types. Over all considered experiments, only the feature sets without spectral modulation processing, or in more general terms *across-frequency processing* (LogMS, PEMO, SGBFB-R-T) resulted in under-estimated thresholds that were off by more than an order of magnitude (>10 dB). The ASR features (MFCC, GBFB, and SGBFB) provided simulation results which came closer to the human performance or even exceeded human performance in some tasks. The simulation results which least under-estimated the empirical thresholds were obtained using SGBFB features, by under-estimating the empirical performance by no more than 2.6 ± 2.0 dB, followed by GBFB features with 5.5 ± 2.1 dB, and MFCC features with 6.9 ± 3.5 dB.

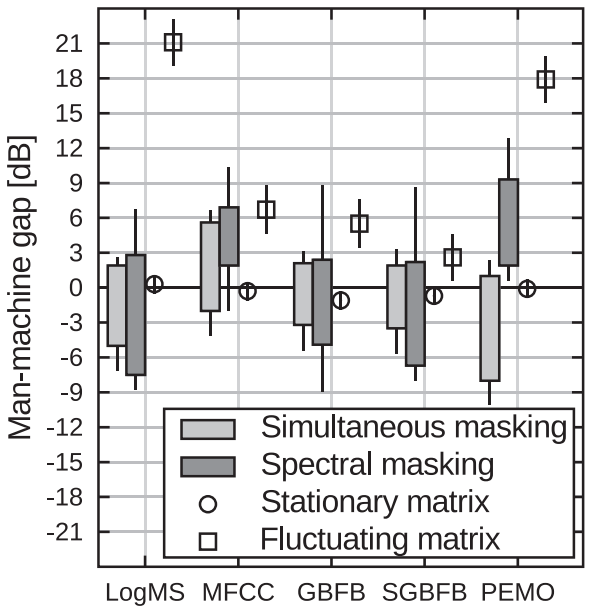

FIG. 10. Differences between simulated thresholds and empirical data depending on the feature set and experiment group. The difference is interpreted as the gap between human performance and machine performance; the lower the values, the smaller the gap, where positive and negative values indicate sub-human and super-human recognition performance, respectively. For the masking experiments, each of which has several conditions, the maximum and the minimum difference to the empirical data are depicted. The error bars indicate the uncertainty of the corresponding (minimum/maximum) value.

F. Effect of feature vector normalization

Considering the results in Tables I, II, and III, the MVN was found to have a minor effect on the simulation results except for the LogMS features in the spectral masking experiment simulation, where the deviation of a simulated threshold was exceptionally high. The use of MVN did neither qualitatively improve the overall simulation fidelity with LogMS-MVN or PEMO-MVN features nor did its omission when using MFCC-MVN features.

IV. DISCUSSION

It was shown that FADE enables the simulation of discrimination experiments of highly variable complexity using different feature vectors. The simplest experiment was a tone-in-noise detection task and the most complex the recognition of German Matrix sentences in a modulated noise condition. The feature vectors included traditional and robust ASR features as well as the output of a non-linear auditory model. The simulated thresholds were interpreted as predictions for the outcome of the corresponding experiment when performed by listeners with normal hearing.

A. Interpretation of simulated thresholds

All simulated thresholds are reference-free (i.e., neither the deviation from a reference-signal-based "optimal detector" nor an empirical reference threshold was employed) and were obtained with a recognition system that was primarily constrained by the input signals and the signal representation. In comparison to many models of psychoacoustic performance, the approach to construct or train an optimal detector with prior knowledge about the exact temporal stimulus alignment [such as, e.g., employed by Dau *et al.* (1996a); Dau *et al.* (1997)] is replaced by a training phase of the Hidden Markov Model and the selection of the respective training condition yielding the lowest predicted threshold. This selection requires *feedback* about the recognition performance and is the only information with that FADE in its current version is provided and human listeners usually not. However, human listeners

could probably guess the SNR at which they are listening. Hence, to better simulate the human recognition task, it seems worthwhile to investigate the possibility of taking the decision blindly in future work. It should be noted that the criterion for the decision on the optimal training data set is *recognition performance* and independent from any empirical data, as opposed to determining a fixed, e.g., training-test SNR offset, based on empirical data. The FADE approach, which decodes feature sequences instead of matching patterns, also models the uncertainty about the temporal alignment of the stimuli. Hence, it might be considered as more appropriate model of the human recognition process than an optimal detector, which requires *a priori* information that human listeners do not have access to. In comparison to state-of-the-art methods of robust ASR, the simulation of the German Matrix sentence test actually is an ASR experiment, but with most of the generic demands on a robust ASR system moved aside. That is to say, the ASR setup is not constructed to accommodate, e.g., generalization over speakers, noise conditions, reverberation, dialects, and other factors. Over a common ASR experiment, the approach to drastically reduce the number of those very broad demands has the advantage that it clearly shows when a feature set is not able to cope with a situation, like in the fluctuating noise condition of the German Matrix sentence test.

As a simulation with FADE is a very controlled A(S)R experiment, the same interpretation as in ASR is valid: the lower the threshold the "better" the system. In this context, thresholds below the corresponding empirical thresholds mean super-human recognition performance and thresholds above the corresponding empirical thresholds mean there is a gap in performance between the man and the machine, also referred to as the *man-machine-gap*. It should be noted that this interpretation is only possible because the thresholds with FADE are reference-free, objective thresholds and that this property translates naturally to the domain of psychoacoustic experiments.

While in the domain of ASR it is difficult to achieve (and hence predict) super-human performance because of its extensive demands, which result in a high variability of the signals to be recognized, in the domain of psychoacoustic experiments it is relatively easy to predict super-human performance because the trained detector stage (the HMMs in our case) can be highly specialized to the well-defined stimuli, which show less variability. This hypothesis is supported by the data in Fig. 10, where for the speech recognition tasks no significant super-human performance was predicted, while for the tone detection tasks, some simulated thresholds were below the corresponding empirical thresholds. For current optimal detector-based psychoacoustic models, the additional *a priori* information about the temporal alignment theoretically further facilitates achieving super-human performance predictions.

Even though the main prediction result of the current work concerns the threshold estimation discussed so far, more details of the FADE simulations might be considered to further validate the modeling of speech recognition and psychoacoustic tasks performed so far. For example, the slope of the psychometric function at the threshold could be derived from the recognition result map (RRM) and compared to empirical data. Likewise, the RRM could be evaluated for, e.g., each word group separately and word confusion matrices could be derived. Also, the selected training conditions could reveal differences between different feature sets.

B. Signal processing dependence of simulated thresholds

The simulated thresholds were found to depend on the employed feature set, where, in the speech recognition cases, the least variability was observed for the German Matrix sentence test in the test-specific noise condition and the most variability was observed for the German Matrix sentence test in the modulated noise condition. In the latter, the least fitting thresholds (–0.5 dB SNR) were obtained with LogMS features while the best predicting thresholds (–19.0 dB SNR) were obtained with SGBFB features, spanning a range of almost two orders of magnitude (20 dB). In the tone-in-noise detection experiments the dependence on the feature set was not as pronounced as in the modulated noise condition of the German Matrix sentence test. As, apart from the feature set, nothing in the setup was changed, this finding confirms the hypothesis that the signal processing employed in the feature extraction process plays an important role in modeling auditory experiments.

Interestingly, the simulated thresholds for the German Matrix sentence test in the test-specific noise condition were not found to depend on the very different feature sets, i.e., PEMO and MFCC features, while the simulated thresholds in the modulated noise condition exposed the decisive shortcomings of some of the considered feature sets (cf. Sec. IV E). Hence, the modulated noise condition was found to be the "critical" experiment to distinguish across the feature sets employed here.

Schädler and Kollmeier (2015) observed in a robust ASR experiment that an ASR system using GBFB features outperformed one using MFCC features, and one using SGBFB features outperformed one using GBFB features. Further, one can assume that the LogMS features will generally not outperform MFCC features in robust ASR tasks as well. The same pattern was observed in the simulated thresholds of the modulated noise condition. Obviously, the most complex experiment of the current study, the German Matrix sentence test in the modulated noise condition, poses very similar basic demands on the employed feature set as in realistic robust ASR tasks. In future work, it could be investigated if this correspondence holds for different features and robust ASR tasks.

C. Required assumptions for ADE simulations

In comparison to current psychoacoustical modeling approaches, FADE poses comparatively few assumptions about the tasks and stimuli, i.e., the following assumptions must be valid in order to simulate an experiment with FADE.

1. Psychometric function

The primary assumption is, that the goal of the experiment is to determine a point on a psychometric function. The

psychometric function needs to indicate the recognition rate on an auditory discrimination task depending on an independent variable which controls the difficulty of the task. The number of classes which have to be discriminated must be limited. In the current study the classes were either target and reference, or 50 different words of which 10 needed to be discriminated at a time, i.e., 1-out-of-2 and 1-out-of-10 discrimination tasks.

2. The same stimuli as in the original experiment

As the basic idea is to estimate the lowest obtainable threshold given a certain task, a set of stimuli, and a signal representation (features), the signals used to perform the simulation must be the same that were used in the original experiment. More technically, the method to generate signals of different classes (e.g., target and reference) for different values of the independent variable must be provided. The signal representations must exhibit a certain variability which may be due to the signal itself (such as, e.g., external noise or other sources of variations within the provided signals) or due to a stochastic process in the feature extraction (such as, e.g., internal noise or uncertainty about the signal and which feature is best suited). For the experiments in the current work, the noise and speech signals caused sufficient variations, and the feature extraction was deterministic. The shortest stimulus used in the current study was a tone which lasted 5 ms, the longest was a word (the German word "achtzehn") which lasted about 900 ms. Technically, no hard limitations with respect to the stimulus length exist.

3. Observable effects due to signal processing

The observable effect must originate from the interaction between the stimuli and the signal processing involved in the feature extraction, where the stimuli incorporate the task requirements and the signal processing the limitations of the human auditory system. This condition expresses the requirement that, differences in the stimulus which are not apparent in the signal representation cannot be detected by the recognition system and will hence not result in different thresholds.

D. Generalization of the FADE approach

One set of parameters was shown to suffice for a variety of experiments and features (cf. Fig. 9). The criterion to determine these parameters was the lowest obtainable thresholds and hence, they were independent of the empirical data of the considered tasks. These parameters also worked well in the simulation of the experiments which are not included in Fig. 9, i.e., the spectral masking experiment and the German Matrix sentence test in the modulated noise condition. Hence, the FADE approach generalized well over the considered experiments and features. The fact that a single set of parameters was sufficient for a variety of complex tasks and different types of features provides evidence that the underlying approach might be appropriate to simulate more experiments and that other features can be incorporated as well to model an even larger variety of experiments with the same set of parameters.

E. Across-frequency processing and relation to temporal processing

The data from Table III indicate that a correct direction of the modulation effect (i.e., a reduction in SRT by about 14.5 dB in humans due to modulations imposed on the noise) was only found for feature sets which incorporated some kind of across-frequency processing. For example, when extracting MFCCs, the DCT was calculated in the spectral dimension of the LogMS and hence MFCCs integrated over the whole spectral bandwidth. With GBFB and SGBFB features the LogMS was spectrally band-pass filtered. With these feature sets improved thresholds were found in the modulated noise condition. However, an opposite effect, i.e., the predicted thresholds increased in the modulated noise condition compared to the stationary noise condition, was observed for LogMS and PEMO features, of which the spectral bands are assumed to be independent. The SGBFB-RR features, a reduced set of SGBFB features, allowed one to perform either only the temporal modulation processing (SGBFB-R-T) or only the spectral modulation processing (SGBFB-R-S). The simulated thresholds with these tailored feature sets showed that the temporal processing alone (SGBFB-R-T) did not show an appropriate modulation effect, while the spectral processing alone (SGBFB-R-S) was sufficient to obtain an improved threshold in the modulated noise condition. With the set up implemented in this study it was not possible to explain the modulation effect without some kind of across-frequency processing.

A representation which allows the reliable detection of local spectral maxima based on, e.g., slope or curvature, instead of absolute values, which have to be relied on if no across-frequency comparison is performed, could probably help the back-end in decision-taking. Hence, it seems possible that at least some kind of across-frequency processing, in its most explicit form the spectral modulation processing performed by the SGFB-R-S feature set, is required to recognize speech in fluctuating noise. If true, this finding might have far-reaching consequences for any system (biological or technical) with the intention to recognize human speech, as it puts the common understanding that speech can be processed in independent frequency bands into question. For example, it might be desirable to preserve spectral modulation patterns rather than temporal modulation patterns in signal processing strategies of hearing devices if preserving speech intelligibility in non-stationary background noise is a declared intention.

Another yet unresolved question is if the spectral and temporal modulation processing in the human auditory system interact with each other or if they are separate processes. Schädler and Kollmeier (2015) observed that no spectro-temporal interaction in the modulation filtering, i.e., inseparable spectro-temporal filters, was needed to outperform MFCC and GBFB features in an ASR system employed in acoustically adverse conditions which included spectrally, temporally and spectro-temporally modulated noise. This

observation is supported by the thresholds of the modulated noise condition that were simulated in this study. In Fig. 10, the simulated thresholds obtained with SGBFB features were among the most suitable for explaining the empirical data. This could indicate that the SGBFB features might be a reasonable model of the auditory processing in the human auditory system and, if so, hint that spectral and temporal modulations in the human auditory system might be processed separately.

V. CONCLUSIONS

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

- FADE was successfully employed to simulate, and hence, predict the outcome of a broad range of auditory detection experiments with an increasing complexity while requiring fewer assumptions compared to traditional modeling approaches.
- A single set of general parameters was determined which was used to simulate all experiments from the basic tone-in-noise detection experiment to the complex speech-in-modulated-noise recognition task.
- Across-frequency processing was found to be crucial to predict the improved speech reception threshold in modulated noise conditions over stationary noise conditions.
- Of all considered signal representations, the Gabor filter bank based features with some across-frequency processing, most notably GBFB and SGBFB features, provide the most suitable model of human performance across the considered experiments.

ACKNOWLEDGMENTS

This work was supported by the Deutsche Forschungsgemeinschaft SFB/TRR 31 "The active auditory system" and the Cluster of Excellence Grant "Hearing4all."

ANSI (**1997**). S3.5-1997, *Methods for Calculation of the Speech Intelligibility Index* (Acoustical Society of America, New York).

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.

Dau, T., Püschel, D., and Kohlrausch, A. (**1996a**). "A quantitative model of the 'effective' signal processing in the auditory system. I. Model structure," J. Acoust. Soc. Am. **99**(6), 3615–3622.

Dau, T., Püschel, D., and Kohlrausch, A. (**1996b**). "A quantitative model of the 'effective' signal processing in the auditory system. II. Simulations and measurements," J. Acoust. Soc. Am. **99**(6), 3623–3631.

De La Torre, A., Peinado, A. M., Segura, J. C., Pérez-Córdoba, J. L., Benítez, M. C., and Rubio, A. J. (**2005**). "Histogram equalization of speech representation for robust speech recognition," IEEE Trans. Speech Audio Process. **13**(3), 355–366.

Derleth, R. P., and Dau, T. (**2000**). "On the role of envelope fluctuation processing in spectral masking," J. Acoust. Soc. Am. **108**(1), 285–296.

Dreschler, W. A., Verschuure, H., Ludvigsen, C., and Westermann, S. (**2001**). "ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment," Int. J. Audiol. **40**(3), 148–157.

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

FADE (**2016**). "Reference implementation of the simulation framework for auditory discrimination experiments" [computer program], http://medi.uni-oldenburg.de/FADE (Last viewed January 29, 2016).

Green, D. M., and Swets, J. A. (**1966**). *Signal Detection Theory and Psychophysics* (Wiley, New York).

Hermansky, H. (**1990**). "Perceptual linear predictive (PLP) analysis of speech," J. Acoust. Soc. Am. **87**(4), 1738–1752.

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.

ISO (**2003**). 226:2003, *Acoustics—Normal Equal-Loudness-Level Standard* (International Organization for Standardization, Geneva, Switzerland).

Jepsen, M. L., Ewert, S. D., and Dau, T. (**2008**). "A computational model of human auditory signal processing and perception," J. Acoust. Soc. Am. **124**(1), 422–438.

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.

Kleinschmidt, M., and Gelbart, D. (**2002**). "Improving word accuracy with Gabor feature extraction," in *Proceedings of Interspeech 2002*, ISCA, pp. 25–28.

Meyer, B. T., and Kollmeier, B. (**2011**). "Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition," Speech Commun. **53**(5), 753–767.

Moore, B. C. J., Alcántara, J. I., and Dau, T. (**1998**). "Masking patterns for sinusoidal and narrow-band noise maskers," J. Acoust. Soc. Am. **104**(2), 1023–1038.

Moritz, N., Schädler, M. R., Adiloglu, K., Meyer, B. T., Jürgens, T., Gerkmann, T., Kollmeier, B., Doclo, S., and Goetze, S. (**2013**). "Noise robust distant automatic speech recognition utilizing NMF based source separation and auditory feature extraction," in *Proceeding of CHiME Workshop 2013*, Vancouver, British Columbia, Canada, pp. 1–6.

Patterson, R. D., and Moore, B. C. J. (**1986**). "Auditory filters and excitation patterns as representations of frequency resolution," in *Frequency Selectivity in Hearing* (Academic Press, London), pp. 123–177.

Qiu, A., Schreiner, C. E., and Escabí, M. A. (**2003**). "Gabor analysis of auditory midbrain receptive fields: Spectro-temporal and binaural composition," J. Neurophysiol. **90**(1), 456–476.

Schädler, M. R., and Kollmeier, B. (**2015**). "Separable spectro-temporal Gabor filter bank features: Reducing the complexity of robust features for automatic speech recognition," J. Acoust. Soc. Am. **137**(4), 2047–2059.

Schädler, M. R., Meyer, B. T., and Kollmeier, B. (**2012**). "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," J. Acoust. Soc. Am. **131**(5), 4134–4151.

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

Schröder, J., Moritz, N., Schädler, M. R., Cauchi, B., Adiloglu, K., Anemüller, J., Doclo, S., Kollmeier, B., and Goetze, S. (**2013**). "On the use of spectro-temporal features for the IEEE AASP challenge 'Detection and classification of acoustic scenes and events,'" in *Proceeding of Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) 2013*, IEEE, pp. 1–4.

Søndergaard, P. L., and Majdak, P. (**2013**). "The auditory modeling toolbox," in *The Technology of Binaural Listening* (Springer, Berlin), pp. 33–56.

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

Tchorz, J., and Kollmeier, B. (**1999**). "A model of auditory perception as front end for automatic speech recognition," J. Acoust. Soc. Am. **106**(4), 2040–2050.

Verhey, J. L., Dau, T., and Kollmeier, B. (**1999**). "Within-channel cues in comodulation masking release (CMR): Experiments and model predictions using a modulation-filterbank model," J. Acoust. Soc. Am. **106**(5), 2733–2745.

Viemeister, N. F. (**1979**). "Temporal modulation transfer functions based upon modulation thresholds," J. Acoust. Soc. Am. **66**(5), 1364–1380.

Viikki, O., and Laurila, K. (**1998**). "Cepstral domain segmental feature vector normalization for noise robust speech recognition," Speech Commun. **25**(1), 133–147.

Wagener, K. C., and Brand, T. (**2005**). "Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: Influence of

measurement procedure and masking parameters," Int. J. Audiol. **44**(3), 144–156.
Wagener, K., Brand, T., and Kollmeier, B. (**1999**). "Entwicklung und Evaluation eines Satztests in deutscher Sprache Teil III: Evaluation des Oldenburger Satztests" ("Development and evaluation of a German sentence test—Part III: Evaluation of the Oldenburg sentence test"), Z. Audiol. **38**, 44–56.
Warzybok, A., Brand, T., Wagener, K. C., and Kollmeier, B. (**2015**). "How much does language proficiency by non-native listeners influence speech audiometric tests in noise?," Int. J. Audiol. **54**, 88–99.