Acta Acustica 2022, **6**, 31

https://doi.org/10.1051/aacus/2022028

Available online at:
https://acta-acustica.edpsciences.org

Topical Issue - Auditory models: from binaural processing to multimodal cognition

TECHNICAL & APPLIED ARTICLE

OPEN ACCESS

Interactive spatial speech recognition maps based on simulated speech recognition experiments

Marc René Schädler*

Medizinische Physik and Cluster of Excellence Hearing4all, Universität Oldenburg, 26111 Oldenburg, Germany

Received 2 April 2021, Accepted 1 July 2022

Abstract – In their everyday life, the speech recognition performance of human listeners is influenced by diverse factors, such as the acoustic environment, the talker and listener positions, possibly impaired hearing, and optional hearing devices. Prediction models come closer to considering all required factors simultaneously to predict the individual speech recognition performance in complex, that is, e.g. multi-source dynamic, acoustic environments. While such predictions may still not be sufficiently accurate for serious applications, such as, e.g. individual hearing aid fitting, they can already be performed. This raises an interesting question: *What could we do if we had a perfect speech intelligibility model?* In a first step, means to explore and interpret the predicted outcomes of large numbers of speech recognition experiments would be helpful, and large amounts of data demand an accessible, that is, easily comprehensible, representation. In this contribution, an interactive, that is, user manipulable, representation of speech recognition performance is proposed and investigated by means of a concrete example, which focuses on the listener's head orientation and the spatial dimensions – in particular width and depth – of an acoustic scene. An exemplary modeling toolchain, that is, a combination of an acoustic model, a hearing device model, and a listener model, was used to generate a data set for demonstration purposes. Using the spatial speech recognition maps to explore this data set demonstrated the suitability of the approach to observe possibly relevant listener behavior. The proposed representation was found to be a suitable target to compare and validate modeling approaches in ecologically relevant contexts, and should help to explore possible applications of future speech recognition models. Ultimately, it may serve as a tool to use validated prediction models in the design of spaces and devices which take speech communication into account.

Keywords: Interactive speech recognition map, Speech perception model, Speech recognition performance, Complex acoustic scene, Speech masking, Speech in noise, Impaired hearing, Hearing loss compensation

1 Introduction

"What could one do with a speech perception model that accurately predicts the individual speech recognition performance of a listener in any listening situation?" While this still remains a philosophical question, with the advances in automatic speech recognition (ASR) and especially speech recognition modeling, it will become more tangible in the future. It also adds a problem-centric perspective to the often rather model-centric perspectives on speech intelligibility, because it asks for concrete applications. Natural applications could be: hearing aid fitting/evaluation, speech-perception aware acoustic room/space design, or virtual assessments of hearing-related technology with respect to speech perception. In any case, it will mean to predict *a lot* of data. But huge amounts of data are not necessarily useful by themselves, which raises the question which data would be useful to predict in the first place.

In this contribution, as a first step towards a scenario where predictions of human speech recognition performance are accurate and inexpensive, an accessible visual representation of extensive modeling data is presented and investigated. The focus lies on presenting a concrete example application with exemplary data using existing model components in the context of speech perception with impaired hearing. However, its aim is not to compare or discuss the prediction accuracy of existing models that predict speech intelligibility. Its aim is to demonstrate the potential of future speech perception models in a detailed example, and to motivate thinking about suitable evaluation criteria for such models.

Human speech recognition performance in realistic spatial listening conditions is affected by many non-linearly interacting factors, even when only the acoustic modality is considered. The situation gets even more complex (technically speaking, the simulation problem gets more degrees of freedom) if, in addition, a hearing loss limits the speech perception of a listener. And when this specific listener uses

*Corresponding author: marc.rene.schaedler@uni-oldenburg.de

an individually fitted hearing aid, an important question is whether the device will improve the individual speech recognition performance sufficiently to enable a conversation. Improving the speech recognition performance is the main goal when a listener with impaired hearing is provided with hearing aids [1]. Modern hearing aids are inherently non-linear complex adaptive signal processing devices [2]. In different communication situations or for different users they can have a very different effect on speech recognition performance [3]. Understanding the interactions between the acoustic environment, the hearing devices, and an individual listener in this context is the key to identify and address acoustic communication difficulties. A perfect oracle speech perception model would show the same behavior, that is, speech recognition performance, than human listeners and a suitable representation of its predictions would make this behavior observable.

To get a concrete idea of the complexity involved, picture a living room with a TV set, a door, and a listener on the couch in front of the TV set. Talking to this listener in that environment will result in a percept which will depend on the following incomplete list of factors:

- The uttered sentence,
- any competing signals,
- the room acoustics,
- the head orientation of the listener,
- the position of the talker,
- the individual hearing abilities,
- and the signal processing of any hearing aids.

Without going into further detail, the speech recognition performance of a listener with aided impaired hearing in a complex spatial acoustic environment depends on too many parameters to assess their interactions – and hence their relative importance – by means of listening experiments. Measurements with human listeners are too expensive to screen the many possible parameter combinations. However, recent developments in the modeling of human speech recognition performance bring the simulation of individual speech recognition experiments for listeners with aided impaired hearing in complex acoustic conditions within reach.

Spatial acoustic scenes can already be modeled/simulated with acceptable effort, either statically with measured impulse responses, or dynamically with acoustic rendering techniques [4], or even mixtures of both techniques, e.g. [5–7]. Hearing devices are signal processing devices, which can be modeled very accurately with corresponding reference implementations, e.g. [8]. And recently, the individual speech recognition performance of listeners with aided impaired hearing was accurately predicted using a modeling approach that employs a re-purposed and modified automatic speech recognition (ASR) setup to simulate speech recognition experiments [9]. These approaches to model complex acoustic environments, individual hearing aid processing, and individual speech recognition performance with processed signals can already be connected to build a toolchain for modeling the spatial aided speech recognition performance in complex acoustic scenes. Such connected models, or modeling toolchains, are now feasible and were already built, e.g. in the ongoing "Clarity challenge"[1] to approach hearing loss compensation with machine learning. An overview of the exemplary toolchain which was used in this contribution, and which is described later in more detail, is depicted in Figure 1. In essence, speech recognition experiments are performed using a re-purposed ASR system (which consists of a front-end and back-end), where all acoustic paths are simulated.

Eventually, one of these approaches could allow one to predict human speech recognition performance in more individual listening conditions – individual to a single listener, e.g. due to the individual hearing devices – than a single listener could ever measure. With increasing accuracy of the predictions, the simulation data will be useful for diverse applications. But the diverse nature of the model parameters also poses a challenge to present such simulation results in an accessible and meaningful way, and more so, to compare predictions of different models. Therefore, it seems to be an appealing approach to put a specific problem statement first and then develop and improve the required model components instead of putting a specific (speech intelligibility) model first and only consider problems that can be solved with it. This is closely related to the question asking which aspects of speech perception are relevant for human listeners in their communication environments, and hence relevant to consider in models.

To define such a problem statement and to explore, interpret, and present the corresponding predicted data sets, suitable tools are needed. Such a tool, an interactive spatial speech recognition map, is presented in the following and its suitability to expose relevant listener behavior is evaluated in an exemplary acoustic scene with the demonstration toolchain depicted in Figure 1 which consists of existing model components. *Relevant listener behavior* here means, e.g., that a listener with unaided impaired hearing shows low speech recognition performance of a distant talker when a nearby TV set interferes, while a listener with normal hearing experiences no problems, especially when TV and talker are spatially separated. Hence, the spatial dimension of the acoustic scene, especially distance and direction of sound sources relative to the listener, plays an important role.

In an earlier approach to visualize the spatial dimension of speech perception, it was proposed to render so called *complex "intelligibility maps"* [10] based on predictions of a binaural speech intelligibility model. There, in a top view of an acoustically rendered room, these maps showed the target and interferer positions, and the target-to-interferer ratio as a gray-scale overlay. The values could be interpreted as a proxy for the speech intelligibility that a listener facing the target would encounter at a position in the room. These maps were introduced *to illustrate the potential applications of the prediction method to support the design of social interaction spaces* [10]. The application the authors had in mind was a tool to aid room design. By letting the top view determine the primary dimensions of the

[1] http://claritychallenge.org/.

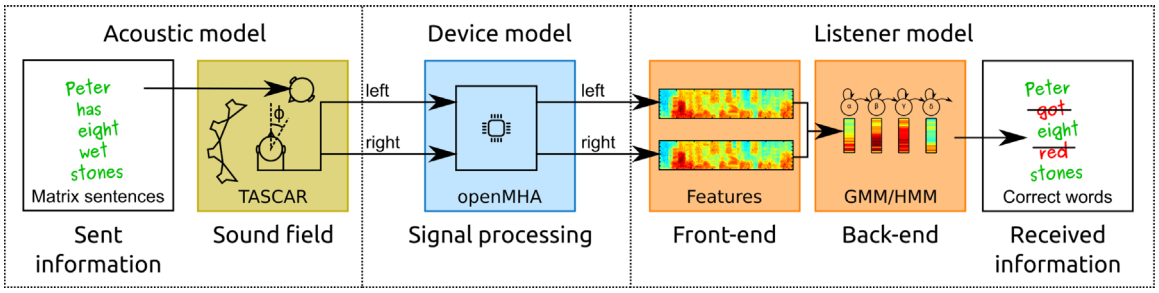

Figure 1. Overview of the employed combination of existing model components to create a toolchain for predicting the human speech recognition performance in complex (that is, challenging) acoustic listening conditions. Speech information is presented in a simulated sound field, the acoustic signals are rendered to the ear level and optionally processed with a hearing device model before they are recognized by a re-purposed automatic speech recognition system. In the first square, speech information is encoded by selecting a matrix sentence. The sentence is presented in a virtual acoustic scene created with the software TASCAR. The output of that step, a noisy binaural speech recording, is optionally processed with a hearing aid model, implemented with the software openMHA. The (optionally processed) noisy recording is recognized with an ASR system trained to recognize matrix sentences (orange squares), where a hearing loss can be configured in the feature extraction stage. In the last square, the percentage of correctly recognized words is evaluated. The approach is described in detail in the Section 2.

representation, an intuitive orientation was guaranteed. Also, the 2D[2] representations assign to each point in the relevant space one relevant quantity, and hence avoid adding distracting dimensions to the representation itself. To show the effect of parameter variations, Figures 7 and 8 in the corresponding publication [10] show many of these intelligibility maps side-by-side in a tiled view. While tiling cannot be avoided in a non-interactive medium, interactivity, or at least click-ability, is now widely exploited in media use.

The current contribution specifically aims to advance on this appealing concept and take it to a modern medium with a demonstration example using "toy" data generated with existing models, including an exemplary interpretation of that data. It is a demonstration of how predictions of speech recognition performance in complex acoustic environments can be obtained, accessibly presented, and potentially evaluated or compared in the future with interactive media. In order to avoid an unintended generalization, the specific implementation presented here is referred to as the *demonstration example* or, synonymously, *the demonstrator*.

2 Methods

2.1 Design of interactive speech recognition maps

Interactive (clickable) "speech recognition maps" are proposed, where the following aspects guided the design decisions:

1. The user interacts with the depicted scene by single clicks (no double-clicks), possibly via a touch screen.
2. The model parameter values can be intuitively changed or selected.
3. A quantity that can be measured in listening experiments is depicted as the primary data.
4. An aid to interpret the depicted quantity is always visible.
5. The user interface makes use of colors to facilitate the interpretation.
6. The primary interaction with the scene is to move the head of the virtual listener by clicking.

The interactive parameter selection via touch screen will allow one to change parameters in the scene, e.g. head position, hearing aid use, and background noise. A target-to-interferer ratio [10] is a technical quantity which cannot be determined from listening tests with human listeners. To enable a direct comparison of model predictions and human speech recognition performance, the outcome of a speech recognition test would be a suitable quantity, e.g. a speech reception threshold (SRT). Using models that predict SRTs has the advantage that the modeled absolute speech levels can be directly related to the speech levels which possible interlocutors would naturally use or physically able to produce [11]. It suggests a natural interpretation of the depicted quantity, e.g.: *Will the talker need to shout to be intelligible?* A colored encoding of the effort that is required to produce the modeled speech levels facilitates the identification of spatial zones in which speech communication is feasible. Allowing head movements, or at least head rotation, will allow one to highlight the effect of any spatial signal processing of a hearing device on speech recognition performance, e.g. the effect of adaptive beamformers.

The concept is too abstract to illustrate it in general. Hence, the focus of this contribution is on a demonstration example. It includes a technical description of the model parts, their combination which is referred to as the *modeling toolchain*, and a fully functional proof-of-concept of the graphical user interface (GUI). Figure 2 presents the proposed GUI in two states: In the left panel for a head orientation to the left, and in the right panel for a head orientation to the right. The color at each position encodes the simulated SRT for a target talker at that position.

[2] Because the underlying data structure is a two-dimensional array.

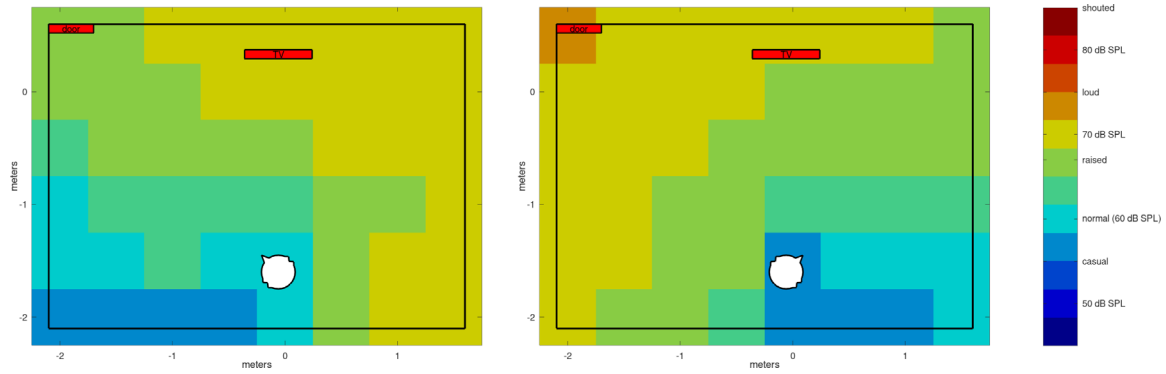

Figure 2. Proposed interactive graphical user interface to explore results of simulated speech recognition experiments; here, for a demonstration example. The quantized colors encode the simulated speech levels that are required such that the listener (white head symbol) understands 50% of the words uttered by a talker located at that position. The state of the representation can be changed by clicking or touching it; here, the states for two different head orientations are depicted.

For these simulations, talker positions were chosen on a grid with a mesh size of 50 cm. The colors are quantized, such that the representation resembles a contour plot and easily identifies zones of similar speech recognition performance. The SRT is presented in dB sound pressure level (SPL). The represented data includes the effect of the acoustic scene, including the listener and talker positions. Essentially, each value indicates the speech level that a talker at a given position has to produce to reduce the masking of his/her speech signal such that at least 50% of the words can be correctly recognized by the listener, where the listener is modeled by an ASR system. This acoustic path can be altered arbitrarily, e.g. by adding a model of a hearing device. The individual hearing abilities can be modeled by removing the information that is not available to listeners with impaired hearing, e.g. low-level signal variations, from the signal path.

The remainder of this contribution describes the implementation of the demonstrator and discusses the representation of the corresponding simulation results in spatial speech recognition maps in the context of their suitability to explore and interpret model predictions.

2.2 Demonstrator

The modeling toolchain of the demonstrator consists of three main parts (cf. Fig. 1): 1) An acoustic scene model, which was used to render environmental noise recordings (used as masker signals) and impulse responses (used to include the target speech signal), 2) a hearing device model, which was used to compensate for an assumed hearing loss of the listener, 3) a model of human speech recognition performance, which predicts the outcome of a speech recognition test, that is an SRT, and which can be configured to include a hearing loss.

The model was used to simulate speech recognition experiments in the demonstrator scene (cf. Fig. 2) with

5 head orientations from −90° to 90° in 45°-steps,
48 talker positions on a grid with 0.5 m mesh size,
2 TV set modes (on and off),
2 door states (open and closed),
2 unaided listeners (normal and impaired hearing),
1 aided listener (only impaired hearing).

This resulted in $5 \times 48 \times 2 \times 2 \times (2 + 1) = 2880$ simulated SRTs. The simulation results were loaded and presented by a proof-of-concept interactive GUI written in GNU/Octave.[3]

2.3 Acoustic scene model

The acoustic environment was a living room scene, as depicted in Figure 2, with three masking sources: 1) A TV set playing back a German weather forecast at a comfortable level in front of the listener position, 2) an informal conversation originating in the neighboring room which is connected by a door, 3) and a dishwasher operating in the neighboring room. The demonstration example scene was implemented and rendered with the Toolbox for Acoustic Scene Creation and Rendering (TASCAR) [7]. TASCAR itself is free and open-source software. Unfortunately, the employed scene description could not be published, because it was based on a pilot version which was still under development and included sounds with unresolved license issues. However, the most important information here is how to employ the acoustic scene in the context of the modeling toolchain, and not which scene exactly was used. The following procedure to obtain suitable data for simulations should be applicable to arbitrary virtual acoustic scenes, including future published versions of the living room scene.

Acoustic scenes in TASCAR are described with a markup language, which allows to define the geometry of the room, the location of the sound sources, and acoustic properties of the surfaces. The software propagates the signals of the sound sources through the room, including reflections at surfaces, and records the result at a desired position. Different receiver types can be chosen which allow binaural recordings.

To manipulate the scene parameters, e.g. the head orientation, placeholders were used in the scene description. These placeholders were substituted by the corresponding

[3] https://octave.org/.

parameter values before using scene description to render the scene to a waveform. For example, the description allows to mute sound sources with the property "mute", which was used to switch the TV source on and off and to open or close the door. The azimuth of the receiver (the head of the listener) was manipulated in the same way.

In one operation mode (called *environment*), the sound sources in the scene were rendered to ear canals of the listener, which results in a binaural recording. For this, in the demonstration example, the receiver type "ortf" was used, which models the ORTF stereo microphone system which can be used to record stereo sound. Recently, a receiver type which models a generic head-related transfer function (HRTF) was added to TASCAR, which should be preferred in the future to better consider head shadow effects. The result of the rendering process, which was performed with the tool `tascar_renderfile`, is a binaural recording of the environmental sounds in the scene, which was used as a noise masker in the simulation of the speech recognition experiments. The duration in the demonstration example was 10 s. Apart from the noise sources, the scene was static, that is the sources and the receiver did not move during the recording. For dynamic scenes with fast moving objects, shorter segments might be recommendable.

In a second operation mode of the scene (called *hrir*), all noise sources were switched off and the impulse responses from a specific talker position to the ear canals of the listener were rendered. For this, all noise sources were muted and a probe source, specifically added for this operation mode, was un-muted. The corresponding addition to the scene description was:

```
<source name="probe" mute="PROBEMUTE">
<position>0 PROBEXXX PROBEYYY PROBEZZZ
</position>
<sound>
<plugins>
<sndfile name="sounds/impulse.wav"
  levelmode="calib" level="65"
  position="PROBESTART" />
</plugins>
</sound>
</source>
```

The words in capital letters were replaced by the corresponding values each time before the scene was used. The result of the rendering process in this mode, which was also performed with the tool `tascar_renderfile`, is a binaural impulse response describing the characteristics of the acoustic path from the talker position to the listeners ear canals. Because of potentially stochastic processes in the scene description, such as, e.g., modules adding late reverberation, the recording of the impulse response was repeated several times; five in the demonstration example. The recorded impulse responses had a duration of 1 s and were employed to filter the speech material before it was used in the simulation of the speech recognition experiments. For the demonstration example, the listener and the target speaker were assumed to rest at a specific position. But generally, some acoustically important objects (sources which may interfere with the target speaker) might move. For dynamic scenes, multiple impulse responses, recorded at different fixed temporal positions could be used to reflect the short-term variability in the acoustic conditions. The approach to render and use impulse responses for embedding the speech signals in the acoustic scene does not allow for a time-variant simulation of the acoustic path of the speech signal. However, this compromise allows one to use the fast convolution to generate large corpora of noisy speech signals, and it reduces the time required for the corpus generation in the demonstrator from hours to seconds. The procedure also makes the scene description independent of the speech material, which allows to re-use the generated files with different speech test material, e.g. different languages or talkers.

Other acoustic rendering software packages should allow to implement a similar procedure to obtain a multichannel recording of the environmental sound and multichannel impulse responses for different (semi-)static states of the acoustic scene. The number of channels of the recordings depends on the receiver type and can either represent the signal at the entry of the ear canal, as in the demonstration example, or at other typical positions of hearing device microphones, e.g. to correctly reflect spatial cues with behind-the-ear devices, and to allow the use of multi-channel noise reduction techniques in the hearing aid processing. The rendering was abstracted using a BASH-script with the following application programming interface (API)

```
render.sh <TYPE> <OUTFILE>
<START> <DURATION> <X> <Y> <Z>
<RECEIVERTYPE> <RECEIVERAZIMUTH>
<TV> <CR> <REVERB>
```

where **TYPE** is the operation mode (`environment` or `hrir`), **OUTFILE** the file name for the recording, **START** the temporal position in the scene where the recording should start, **DURATION** the duration of the recording in seconds, *X*, *Y*, Z, the position of the probe (only relevant for hrir mode), **RECEIVERYTPE** the type of receiver (in the demonstrator `ortf`), **RECEIVERAZIMUTH** the azimuth of the receiver to control the head orientation, **TV** the switch to mute the TV source (1 means *on* in environment mode, 0 *off*), **CR** the switch to mute the connected room sources (1 means *on* in environment mode, 0 *off*), and **REVERB** the switch to control the reverberation module (1 means *on*, 0 *off*, was always enabled in the demonstration example). While this abstraction is specific for the scene, it is simple and allows one to replace the acoustic renderer as well as for an adaptation of the demonstrator to other acoustic scenes. Because the possibilities to combine parameters in a complex acoustic scene are infinite, the limitation by such an abstraction helps to focus the experimental design on a manageable set of manipulable parameters, which can then be controlled by interactive elements in a spatial speech recognition map.

2.4 Speech communication model

In order to measure the suitability of an acoustic channel (here, one-directional from the target speaker location

to the listener in a defined acoustic environment) with respect to speech communication, a speech recognition test with an adaptive speech level control is suitable. An adaptive test, which considers a variable speech presentation level, avoids the disadvantages of potential ceiling or flooring effects. Importantly, a speech presentation level as an outcome can be interpreted in the context of the speech production levels which interlocutors would naturally choose to communicate in the considered acoustic environment.

For the demonstrator, the male German matrix sentence test [12] was used as a model for speech communication. The squares "sent information" and "received information" in Figure 1 correspond to this model. *Model* here means, the speech material was used instead of recordings of real conversations. The matrix test, which exists in more than 20 languages, comprises 50 phonetically balanced common words of which sentences with a fixed syntax, such as "Peter got four large rings" or "Nina wants seven heavy tables", are built. This speech test is typically performed with human listeners to adaptively measure their SRTs in noise. The outcome is the speech level, often reported relative to the noise level as the signal-to-noise ratio (SNR), which is required for the listener to correctly recognize a given percentage of the presented words. Usually, a target recognition rate of 50%-word-correct is used, because the slope of the psychometric function is steepest in this region, which allows precise measurements. Any matrix sentence test can be performed with the data generated from the acoustic scene by using the rendered environmental noise recording as a masker signal and by convolving the speech signals with the rendered impulse responses. Hence, any condition considered in the following simulations could also be measured with human listeners.

The simple and fixed syntax of the test sentences allows an implementation of the recognition experiment with an ASR system with low complexity. Most importantly, it allows one to build a language model that only allows the recognition of valid matrix sentences and which is invariant to different weightings against the acoustic model, that is, an ideal model for the matrix sentence recognition task. Human listeners can also be assumed to dispose of such a close-to-ideal language model, after an initial training session. Because an ideal simple language model can be assumed, the matrix test is mostly sensitive to the contribution of the acoustic model to the speech recognition performance. In other words, the (probably individual) language model is eliminated as an unknown from the simulation of the speech recognition experiment. The limited number of words also favors an implementation with low complexity and comparatively low computational demands, which allows one to run simulations quickly.

In summary, the matrix test is an established measurement tool for the speech recognition performance of human listeners in noisy environments which is also very suitable for simulating the same experiments with correspondingly adapted ASR systems. In the future, more complex speech tests might be employed to better assess the properties of an acoustic channel with respect human speech communication.

2.5 Simulation of speech recognition experiments

For the demonstration example in this contribution, speech recognition experiments were simulated with the Simulation Framework for Auditory Discrimination Experiments (FADE[4]) to predict their outcome [13]. The squares "front-end" and "back-end" in Figure 1 correspond to the re-purposed ASR system employed to recognize the sentences of the speech test in FADE. While the demonstrator was constructed using FADE, the aim was that the setup can be adapted to other modeling approaches (toolchains) in the future, to allow a direct comparison of modeling results, and a comparison with empirical data. For a substitution, a speech intelligibility model should, at least, accept binaural noisy non-linearly processed speech signals and allow one to configure a hearing loss. However, at the time of writing, only few approaches existed which were compatible with these requirements, cf. [14] for a discussion on the compatibility of some existing models. Adding the requirement that the model should predict the outcome of a speech recognition test, only the FADE modeling approach was compatible at the time of writing. And while FADE is compatible in the sense that it is technically possible to predict the outcome of the considered speech recognition tests, it is important to mention that it was not evaluated with empirical data in these conditions.

The only change to the originally proposed prediction setup [13] was the manipulation of the feature extraction stage which is explained in Section 2.6 and the processing of the noisy speech signals with the hearing device model as explained in Section 2.8. Hence, the FADE simulation approach is only outlined here, and we refer the interested reader to the original descriptions [9, 13].

Predictions with FADE are performed completely independently for each condition. Hence, an independent simulation process was started for each of the 2880 considered parameter combinations of the demonstrator scene. Also, the predictions do neither depend on any empirically measured SRTs, nor on predictions in any reference condition. This allowed to predict the outcomes of speech recognition tests for which no empirical data existed. For the prediction of the SRT-50, that is the speech level required to achieve a recognition score of 50%-words-correct, the following standard procedure [13] was used.

An ASR system was trained on a broad range of SNRs with noisy speech material from the considered condition. For this, a corpus of noisy speech material at different SNRs was generated as follows: The clean matrix test sentences ("sent information" in Fig. 1) were convolved with the recorded impulse responses ("sound field" in Fig. 1), and then, randomly chosen fragments from the recorded environmental sounds were added to the them. The noisy signals were optionally processed with the hearing device model ("signal processing" in Fig. 1) when an aided listening condition was considered. From the noisy (and optionally processed) speech signals, feature vectors were extracted ("front-end" in Fig. 1), where this step included

[4] https://doi.org/10.5281/zenodo.4003779.

the implementation of a hearing loss, as described in Section 2.6. Subsequently, an ASR system using whole-word models implemented with Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM) ("back-end" in Fig. 1), was trained on the feature vectors. This resulted in 50 whole-word models for each training SNR. These GMM/HMM models were then used with a language model that considers only valid matrix sentences (of which 10^5 exist) to recognize test sentences on a broad range of SNRs with noisy speech material from the same considered condition. For each combination of a training SNR and a test SNR, the transcriptions of the test sentences were evaluated in terms of the percentage of correctly recognized words ("Received information" in Fig. 1). The resulting recognition result map (cf. Fig. 7 in [13]), which contained the speech recognition performance of the ASR system depending on the training and testing SNRs in 3 dB steps, was queried for the SRT. For a given target recognition rate, here 50% words correct, the lowest SNR at which this performance was achieved was interpolated from the data in the recognition result map and reported as the predicted SRT for the considered condition. The whole simulation process, including the creation of noisy speech material, an optional processing of this noisy speech material with the hearing device model, the feature extraction (which depends on the listener profile), the training of the ASR system, the recognition of the test sentences, and the evaluation of the recognition result map, was independently repeated for each considered condition. In summary, with this approach, the speech level that is at least required for a matched-trained ASR system to achieve a word recognition rate of 50% in the matrix sentence test is measured and reported as the predicted outcome.

2.6 Listener profiles

For the demonstration example, three listener profiles were considered: (1) A listener with normal hearing, (2) a listener with unaided symmetric impaired hearing, (3) the same listener with aided symmetric impaired hearing. The accurate prediction of the individual effect of hearing loss on speech recognition performance requires an efficient model of hearing loss. Such a model was recently proposed for FADE [9, 15], where in addition to the absolute hearing threshold, a supra-threshold factor, the level uncertainty, was introduced to model the effect of supra-threshold hearing deficits on speech recognition performance. The individual increased hearing thresholds and level uncertainty were implemented in the feature extraction stage of the ASR system, with the aim to remove the information that was not available to an individual listener. An important property of this model of hearing loss is that the effect of increased hearing thresholds can be compensated by simple linear amplification, but the effect of the level uncertainty cannot. This is in line with observations that linear amplification can only partially compensate a hearing loss in noise [3].

As already mentioned for other components of the modeling toolchain, the main point here is to illustrate how impaired hearing and aided hearing can be included as factors in the interactive representation of the speech recognition map. For natural listeners with impaired hearing, the effect of their hearing loss on speech recognition performance can only be alleviated, but often not fully compensated by means of a hearing device [3]. Using a hearing loss model which shows this behavior is instructive for the demonstration example, but technically it is not necessary.

The hearing loss was implemented in the feature extraction stage of the ASR system that was employed as the listener model, such as it was proposed in [9]. The hearing threshold was implemented as a hard lower threshold in the log Mel-spectrogram domain, which is the spectro-temporal signal representation on which the feature extraction in ASR systems is usually based. For the listener profile with normal hearing (Profile 1), normal hearing thresholds were considered. For the listener profiles with impaired hearing (Profiles 2 and 3), the hearing thresholds for a standard moderate hearing loss profile were taken from the standard profile N3 [16]. The level uncertainty was implemented by adding random values drawn from a normal distribution to the signal representation in the log Mel-spectrogram domain. For the listener profile with normal hearing (Profile 1), a level uncertainty of 1 dB was configured, this means the standard deviation of the normal distribution was 1 dB. For the listener profile with impaired hearing (Profiles 2 and 3), a level uncertainty of 10 dB was configured, which means the random values were drawn from a normal distribution with a standard deviation of 10 dB. The expected effect of the such modeled hearing loss on the speech recognition performance as measured with the matrix sentence test in noise is that, compared to the normal-hearing configuration, SRTs in quiet environments will be much elevated (due to the increased hearing thresholds) and SRTs in noisy environments will be somewhat elevated (due to the level uncertainty). The chosen model parameter values could be replaced by values inferred from measurements with human listeners [9]. However, such profiles were only measured and validated for monaural listening conditions.

2.7 Binaural hearing model

Binaural cues are highly relevant for spatial hearing, and they are also suspected to play an important role in spatial speech perception in complex acoustic scenes (often referred to as the "cocktail party effect" when the interfering signals are speech signals [17, 18]). A part of the spatial release from masking can often be explained by so called "better-ear listening", where it is assumed that the SNR on one ear is better due to the head-shadow effect. In the presented example scene, the ortf receiver can be seen as a very coarse model of the head shadow effect; the hrtf receiver would provide a better model. With an omni-directional microphone as a receiver, one would expect weaker spatial effects. Better-ear listening can be seen as an extension to monaural listening where the signal representation with the better SNR is used. It is still subject to research which mechanisms human listeners employ for

binaural listening in complex acoustic scenes beyond better-ear listening, and more so, how impaired hearing affects this ability [19].

As already mentioned several times, the main point here is to illustrate how the effect of binaural hearing can be included in the interactive representation of the spatial speech recognition map. However, especially the properties of the binaural hearing model will be responsible for patterns in the spatial representation. Hence, it is highly recommendable to employ a binaural model of speech perception which can take binaural hearing beyond better-ear listening into account.

In order to model binaural hearing beyond better-ear listening, a recently proposed expansion of FADE for predicting SRTs in binaural listening conditions, called KAIN, was employed [20]. Instead of just concatenating the feature vectors of the left ear channel and the right ear channel, the difference of the signal representations of the left and right ear channel in the log Mel-spectrogram domain was calculated to extract binaural features. This simple-to-implement approach achieved a good performance in predicting the spatial release from masking in an anechoic listening condition, where the effect of binaural unmasking is typically most pronounced [20]. The main advantage of the approach is that it integrates seamlessly with the hearing loss model and that it works *blindly*, which means that there are no parameters that need to be adapted to the stimuli, such as it usually is the case for equalization-cancellation based models, e.g. [21]. Also, it is compatible with noisy processed signals, a property sometimes referred to as non-intrusiveness, which means that the clean speech and noise signals are not required to be known in the decision stage, such as it is the case for other binaural models, e.g. [22]. However, it is not be noted that the KAIN approach was not evaluated in other than anechoic listening conditions.

The proposed spatial representation of the speech recognition performance would be highly suitable to intuitively compare and judge predictions of different models of binaural speech perception, which, however, is out of scope for this contribution. Because even slight modifications of the binaural signals might have an effect on the speech recognition performance, it can be expected that the binaural hearing model will interact with the hearing device model.

2.8 Hearing device model

For the application in the demonstrator, the hearing device model can be a signal processing black box. Again, the most important information here is how to employ a hearing device model in context of the proposed toolchain, and not which device, exactly, was used. In the following, the procedure that was used in the demonstrator is described, but it should be applicable in a similar way for other hearing device models.

The open Master Hearing Aid (MHA) [8], which is an open-source software platform for real-time audio signal processing, was used to simulate an individually aided listening condition for the listener with impaired hearing. With openMHA, the signal processing is described in a human-readable configuration file which allows to load, configure, and connect signal processing plugins. The same configuration can be used in real-time applications, e.g., mobile hearing aid prototypes,[5] and in off-line batch processing. The latter mode was used to optionally process the noisy speech signals prior to using the data in the feature extraction stage of the listener model, that is, before the hearing loss is applied.

In the demonstration example, it was assumed that the device model would represent two fully occluding in-the-canal hearing aids, each with one microphone. Hence, the direct sound was assumed to be completely blocked by the device and replaced by the processed signal. This (over-) simplification allowed to use the same recordings from the acoustic scene for the unaided and for the aided conditions. For more complex hearing device configurations which include multiple microphones per device, the signals from the acoustic scene would need to be rendered to the corresponding microphone positions. For future applications, this will definitely be a very interesting option, because it would allow to study the interaction between multi-channel noise reductions algorithms, the head orientation, and a number of spatially distributed noise sources. Also, the direct sound path could be added to simulate the interference of the processed signal with the leaking signal.

The signal processing was configured to implement a multi-band dynamic compressor with band center frequencies of 250, 500, 1000, 2000, 4000, and 6000 Hz. The attack time constant was 50 ms, the decay time constant 500 ms. The gains were prescribed based on the standard audiogram profile N4 [16] with the NAL-NL2 prescription rule [1]. The decision to use N4 instead of N3 (as it was configured for the listener profile) was due to the additional supra-threshold hearing loss, which does not only increase SRTs in noise but also the hearing thresholds in quiet in the model. Of course, this configuration is only a first guess, or rather a "first fit", and it could be replaced and compared to other fitting rules.

The calibration was assumed to be such that a signal with a root-mean-square (RMS) amplitude of 0 dB relative to full scale represented a sound with 130 dB SPL. The correct digital level calibration is especially critical, because of the non-linear level-dependent signal processing performed by the hearing device model. The signal processing of the left and right channel was synchronized, that is, the start end end points of the processing windows were identical. The signal processing was abstracted with a BASH-script with the following API:

```
batch_process <SOURCELIST> <TARGETLIST>
              <INCREMENT>  <OFFSET>
```

where **SOURCELIST** is a text file which contains one input file path per line, **TARGETLIST** is a textfile which contains the corresponding output file paths, and **INCREMENT** and **OFFSET** are integer numbers which

[5] For example: https://github.com/m-r-s/hearingaid-prototype or https://batandcat.com/portable-hearing-laboratory-phl.html.

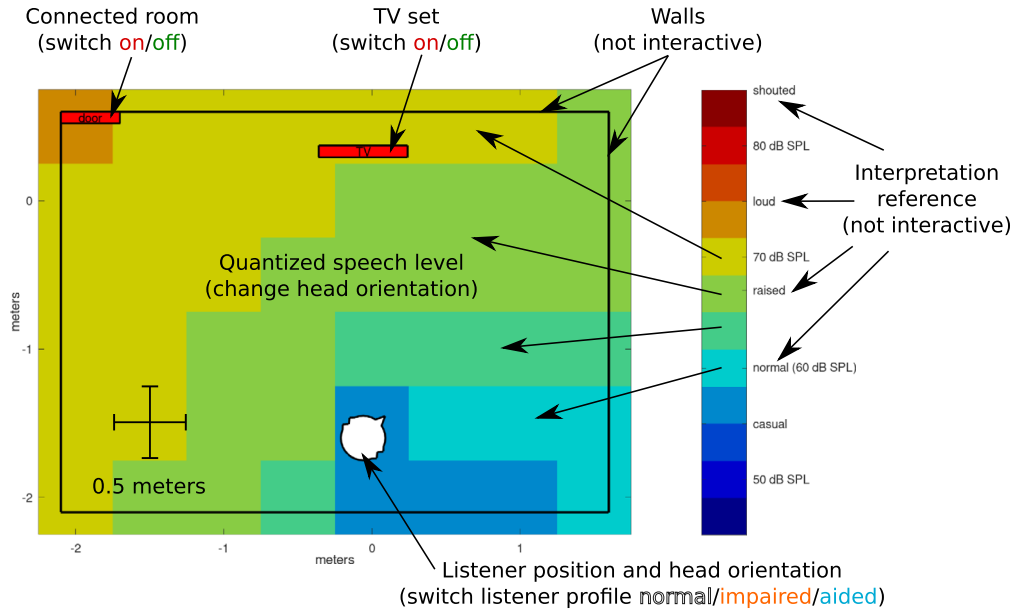

Figure 3. Explanation of the elements of the proposed graphical user interface. The interactivity of each element is explained in parenthesis. Interaction is performed by single-clicking or touching an element.

indicate the initial item and the step size for the iteration over the file list. FADE can use this API to execute the signal processing, where the increment and offset values are used to implement Poor-man's parallelization. It would also be possible to pass additional arguments, e.g. parameters which control noise suppression or other hearing device settings, to the hearing device model. However, for the demonstration example, only the described hearing aid model configuration was used.

2.9 Interactive spatial representation

A fully functional proof-of-concept of the interactive GUI was implemented in GNU/Octave[6] according to the considerations in Section 2.1. Two states of the proposed GUI with the data for the normal-hearing listener profile from the demonstration example were already presented in Figure 2. Figure 3 explains the elements of the GUI and their function.

The SRTs in dB SPL for the selected condition are depicted as a nearest-neighbor interpolated color-encoded image, where the number of color grades was limited to 12 between 45 dB SPL and 85 dB SPL in the demonstrator, which resulted in approximately 3.3 dB per color. With the aim to not distract the user from the speech recognition performance data, the illustration of the room dimensions, the interactive objects, and any possibly important furniture were kept minimalistic. Interaction is only performed by single-clicking or touching the representation. For example, the noise sources can be switched on and off by clicking the corresponding surfaces, which are red if the sources are active, and green if inactive. The head orientation can be changed by clicking anywhere on the quantized color-encoded representation of the speech levels. The listener status can be cycled (normal → white, unaided impaired → orange, aided impaired → light blue) by clicking on the head drawing. Once a desired change is expressed by clicking any interactive element, the figure is updated.

The GUI application, on startup, loads the simulation results from text files generated by the simulation script. After loading the simulation results, the proof-of-concept code first draws all objects, stores their handles into a struct, which in turn is stored in the `UserData` variable of all objects. In addition, for each interactive object, a callback function is defined and registered to the `ButtonDownFcn` field of that object. This allows to access and change the properties of any object from any `callback`. A special function, `refresh_data`, is called after any change of an object property, where the struct with the object handles is passed as the single argument. It implements the update of the figure according to the state of all objects, e.g., the correct display of the head orientation and the corresponding result image. The code of the GUI and the result files of the demonstration modeling toolchain are available [23].

2.10 Availability of resources

Most of the software that was used to develop the demonstrator is free and open-source software. The employed modeling toolchain is based on a previously published software package.[7] However, some parts could not be published to due to restrictive licenses or unresolved license issues. Alternative sources or possible replacements are proposed where suitable. An updated version of the

[6] Tested with GNU/Octave versions 5.2.0, 6.1.1, and 6.2.0 under Ubuntu Linux and Windows 10.

[7] https://doi.org/10.5281/zenodo.3734152.

modeling toolchain (without components than cannot be published) including the GNU/Octave script which implements the interactive spatial speech recognition map for the demonstration example is available [23]. The components that cannot be published by the author of this contribution are:

- The TASCAR scene description due to missing rights. The authors of TASCAR provide extensive examples and tutorials for building acoustic scenes on their website.[8]
- The Hidden Markov Toolkit (HTK) which was used to build the speech recognizer. HTK can be downloaded from the authors website.[9]
- The German Matrix test speech material due to missing rights. An alternative is the freely available *Synthetic German matrix speech test material created with a text-to-speech system*[10] which was evaluated and compared to the female German matrix sentence test [24].

3 Results and discussion

All simulation results discussed in the following were created for the purpose of demonstration. They were not validated against the performance of human listeners.

3.1 Simulation performance

Each state of the interactive spatial speech recognition map, that is, each image, shows simulated data which would require about 48 (number of simulations per image) × 3 (minutes per simulation) = 144 min of very repetitive measurements, if it would be measured with human listeners. While it is feasible to measure single conditions with human listeners, measuring the whole spatial array is at least impracticable. In total, the demonstration example simulations represent an equivalent of about 2880 × 3 = 8640 min (6 days) of uninterrupted measurements. On a modern high-end desktop system,[11] this was also approximately the time needed to complete the 2880 simulations, that is the real time factor of the simulation process is approximately 1. Because all 2880 simulations can be performed independently, they could run in parallel on different compute nodes, and be expected to scale up to a factor equal to the number of simulations. With sufficient hardware, the simulations required for the demonstration example could be calculated in less than 5 min with the current implementation. Given sufficient compute nodes and further optimization and parallelization of the simulation code, even lower simulation delays of less than 1 min might be achievable. While this would not be sufficient to realize a real-time update of the simulation data in the GUI without losing responsiveness, it would be sufficient for model-aided design and development applications. Basing design decisions on simulation results raises questions about their accuracy.

3.2 Simulation accuracy

The prediction accuracy of the demonstrator modeling toolchain depends on the fidelity of all involved model components (acoustic model, device model, and listener model) for that purpose. The focus of this contribution is not on discussing the suitability of the model components, which would also depend on the specific problem to be solved (e.g. individual fitting of a hearing device). In more general, the suitability of approaches in hearing sciences is increasingly often discussed in the context of ecological validity:

> *In hearing science, ecological validity refers to the degree to which research findings reflect real-life hearing-related function, activity, or participation.* [25]

A discussion in the context of ecological validity would require, apart from questioning the prediction accuracy, questioning the suitability of the speech test, as well as of the employed acoustic scene; which is out of scope for the current contribution.

In single listening conditions, the predicted outcomes could be compared against measured outcomes with human listeners to assess the prediction accuracy. However, such an evaluation could only validate the modeling toolchain for that very specific condition. But even if the model predictions were perfect, the degree to which this result reflects real-life hearing-related function, activity, or participation, needs to be shown separately. For example, if some practical conclusions can be drawn from the simulated data which later prove to be correct and relevant for a listener in real life, it would hint at the ecological validity of the simulations. The proposed representation is thought as a tool to explore the – as necessary, also verifiable – simulation data, which can be performed with the objective to identify model behavior that could be ecologically relevant.

Even if not in the focus here, there are many aspects to consider with respect to the fidelity of the employed model components, of which only a few are compiled in the following. The employed speech material was recorded in a quiet environment, but human talkers change their pronunciation when putting effort into producing high speech levels, which is known as the Lombard effect and reported to set in already at background noise levels as low as 43 dB (A) [26]. The employed acoustic renderer (TASCAR) was created for the real-time rendering of acoustic scenes and puts a focus on interactivity, e.g. within virtual reality. It is still subject to research which acoustic properties of a scene are relevant for speech recognition [25] and if TASCAR is capable to adequately model these [7, 27]. The employed scene description was a pilot version of a (still) non-existent room, which is why the employed acoustic simulation could not be verified against a real room. The talker was considered to be an omni-directional source; a very rough assumption. The employed receiver model, ortf, provides only a coarse model of a head-related transfer function, which, however, is fundamental to convey the binaural cues that

[8] http://www.tascar.org/.
[9] http://htk.eng.cam.ac.uk/.
[10] https://doi.org/10.5281/zenodo.4501212.
[11] AMD Ryzen 9 3900X 12-Core CPU with 64Gb of DDR4-3200 RAM.

are important for spatial hearing. While the digital signal processing part of a hearing device could be modeled very accurately in theory, real hearing devices do not use openMHA but proprietary, and usually unavailable, implementations. The coupling of the hearing device to the acoustic sound field and to the ear canal of the listener was over-idealized; real hearing devices do not reduce the leakage of the direct sound to the same extent. Also, any analog properties of the hearing device model, such as, e.g. the self-noise of the microphones, were not considered. The employed normal-hearing listener model, that is FADE, was only evaluated in a range of laboratory listening conditions [13, 14]. The employed model of impaired hearing was only evaluated in monaural conditions [9]. The employed model of binaural hearing, KAIN, was only evaluated in an anechoic condition.

This list of shortcomings of the modeling toolchain components is incomplete and should only demonstrate the challenges which predicting speech recognition performance in realistic and relevant listening conditions poses. For a model to be ecologically valid, all model components *together* will be relevant. Only considering their isolated simulation accuracy might be a good starting point, but will eventually not suffice. The joint consideration of the many parameters that affect speech communication in real-life opens a vast parameter space, whose exploration requires suitable tools.

3.3 Spatial speech recognition maps

The main feature of the spatial representation (cf. Fig. 3) is the quantized nearest-neighbor interpolated image of the SRTs (speech levels needed for 50% word recognition rate). It creates a contour-like partitioning of the image into zones of similar speech levels, without suggesting a higher resolution of the underlying data samples. Due to the nearest-neighbor interpolation, the spatial resolution and the location of the simulated samples are intuitively mediated. The simulated samples are located in the center of each square (pixel). The quantized color map with 12 color grades allows to classify the corresponding zone according to the interpretation reference, which is provided in the legend. The colors from dark blue over cyan, green, yellow, and orange to red, indicate the effort that a talker at the corresponding position would need to put into producing speech at the corresponding level [11]. The levels range from 45 (dark blue) to 85 dB SPL (dark red), because speech levels above 85 dB and below 45 dB SPL are difficult to produce for human talkers. Hence, the color encoding allows to judge from which positions in the scene a communication with the listener could be successful. For a given a speech production level, e.g. raised effort, it enables the user to establish an idea of the listeners range for successful speech communication in the scene. The representation integrates many factors which are relevant for speech recognition into a physical dimension that the affected listener is able to experience: the limitation of the "horizon for speech communication". The interpretation ranges from "no limitation" (all dark blue map) to "speech communication impossible" (all dark red map), with many intermediate states which may depend on the head orientation.

3.4 Head orientation

The head orientation is a particularly important parameter, because it can be controlled by the listener. It was shown to have a potentially pronounced effect on speech recognition performance [18, 28]. Figure 4 illustrates the effect of different head orientations with the normal-hearing listener profile in the demonstrator. Here, only five different head orientations in 45°-steps were simulated, which were chosen to limit the number of simulations. In the interactive GUI, the head is oriented such that it is directed towards any clicked position on the colored SRT image. Intuitively, the head follows the finger when touching the representation on a touch screen.

In the demonstrator, the position of the talker and the head orientation had a pronounced effect on the speech recognition performance with the normal-hearing lister profile, when all noise sources were active. The required speech levels to achieve 50% word recognition rate ranged from "casual" to "loud". As a strong tendency, the required speech levels were lowest (speech recognition performance was best) for a given target talker position when the head was oriented towards that position. The only, albeit little, exception can be observed for talker positions between the TV set (the dominant noise source) and the listener, where turned-away head directions resulted in lower SRTs. Possibly, the early reflections of the sources (target talker and TV) on the walls provide spatial cues which are not available when the sources and the listener are aligned. This is an hypothesis, which however could be investigated. Qualitatively, the observations here are in line with the notion that binaural listening can reduce masking in situations with spatially separated sources [18, 28].

The late (and diffuse) reverberation in the example scene had only little energy. With more late reverberation, the effect of the head orientation can be expected to be smaller. In the context of ecological validity, this would be, for example, an observation with practical relevance for designing rooms with comfortable room acoustics.

3.5 Noise sources

Another considered parameter which can also be potentially controlled by the listener are the noise sources. Figure 5 illustrates the effect of switching the noise sources on and off with the normal-hearing listener profile in the demonstrator. For the case that all noise sources are switched off (first panel), this is, the door is closed and the TV set is switched off, the simulated SRTs are below 45 dB SPL. This means that any speech produced at typically realizable speech levels should be intelligible in this otherwise completely quiet room. By un-muting the noise sources from the connected room (second panel), this is, the door is open and a conversation and the dishwasher are audible, the required speech levels increase, particularly in the region close to the door. This, again, is in line with

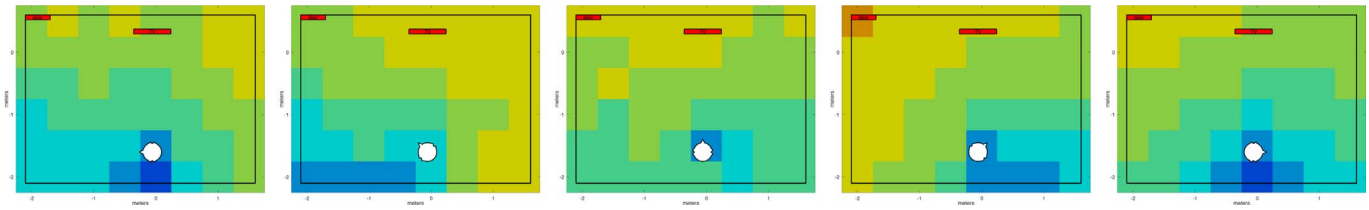

Figure 4. Illustration of the effect of head orientation on the spatial speech recognition map. In the demonstration example, five different directions in 45°-steps were simulated. The legend of the color encoded speech levels is identical to the one in Figure 3.

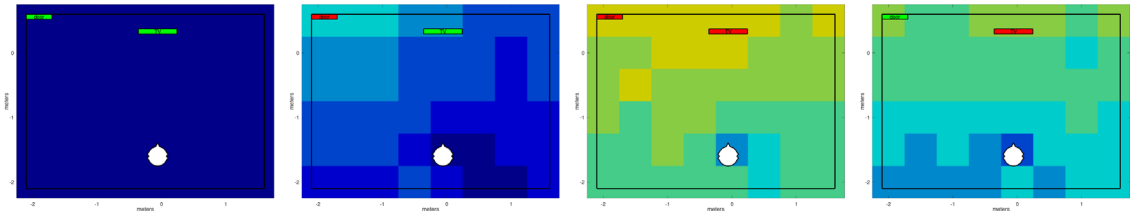

Figure 5. Illustration of the effect of switching on and off noise sources on the spatial speech recognition map with the normal-hearing listener profile. Green boxes with text indicate a muted noise source, red boxes an active noise source. The legend of the color encoded speech levels is identical to the one in Figure 3.

the notion that spatially separated noise maskers can have a reduced masking effect [18, 28]. The levels increase up to an equivalent of normal speech production effort indicated with cyan color, that is about 60 dB SPL.

By only switching on the TV set (last panel), the required speech levels are increased in the whole room. The increase is most pronounced close to the wall where the TV set is located with levels up to 70 dB, and least pronounced close to the wall where the listener is located with levels up to 60 dB. The whole map is relatively symmetric with respect to vertical axis, which is due to the symmetry of the scene.

The speech levels due to the TV set alone are further increased by also opening the door to the neighboring room (third panel). If both noise sources would be co-located, stationary and additionally had similar long-term spectra, one would expect that dominant noise source would determine the required speech levels. But the noise sources are spatially separated and fluctuating, although the dishwasher sound does not fluctuate as much as the speech maskers. The expected interaction, that both interfering noise sources increase the SRT, can be observed in the represented simulation results. In some regions, e.g. between the two noise sources, an increase by two color grades can be observed due to opening the door when the TV is already switched on (compare fourth and third panel). This corresponds to an increase of speech levels by approximately 6.7 dB. In other regions, e.g. close to the right wall, the increase was less with only one color grade. This illustrates how interactions between two typically idealized effects, namely spatial release from masking and modulation masking release, can be observed in a realistic, and possibly relevant, acoustic setting. The proposed speech recognition map can be used to find realistic acoustic scene configurations which are suitable to investigate the interaction between fundamental concepts of speech information masking and, e.g., impaired hearing.

3.6 Impaired hearing

With sufficiently increased hearing thresholds, one would expect that the interfering sources would not be audible anymore and hence show no effect on the speech recognition performance. Figure 6 illustrates the effect of switching the noise sources on and off with the hearing-impaired listener profile in the demonstrator. As expected, the simulation results for the different combinations of noise maskers are virtually identical. A random deviation of up to one color grade can be expected because of the stochastic simulation process. The map colors range from orange to dark red, which indicates a relatively high effort would be needed to produce these speech levels. In a free-field experiment, one would expect the required speech levels to increase with the distance from the listener. The simulation results suggest that the acoustic conditions in the simulated room were far from similar to free-field. The required speech levels are surprisingly homogeneous and rather depend on the side of the room where the talker was located than on the distance from the talker position; they tended to be lower on the same side and higher on the opposite side. The only way of communicating with the listener in the scene without talking *loudly* is to get very close, independently of whether the room was quiet or noisy. This is in line with the idea that noises with levels below the hearing threshold don't have an effect on speech recognition performance. It also suggests that noise sources at moderate levels pose no (additional) problem for listeners with sufficiently elevated hearing thresholds. On the contrary, if noise

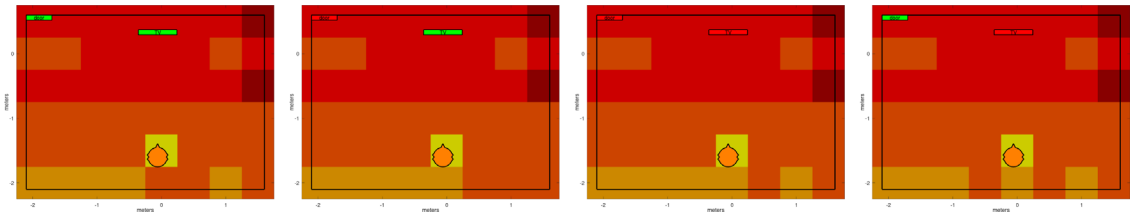

Figure 6. Illustration of the effect of switching on and off noise sources on the spatial speech recognition map with the hearing-impaired profile (indicated by the orange head symbol). Green boxes with text indicate a muted noise source, red boxes an active noise source. The legend of the color encoded speech levels is identical to the one in Figure 3.

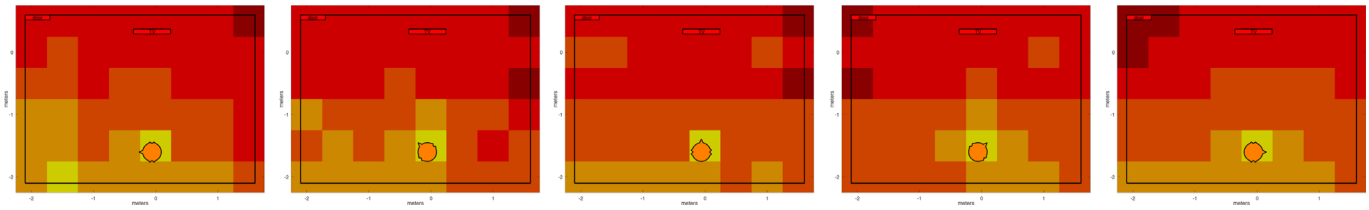

Figure 7. Illustration of the effect of head orientation on the spatial speech recognition map with the hearing-impaired profile (indicated by the orange head symbol). The legend of the color encoded speech levels is identical to the one in Figure 3.

sources at moderate levels are present, talkers will naturally produce higher speech levels [11], which might help listeners with impaired hearing.

Figure 7 illustrates the effect of head orientation on the spatial speech recognition map with the hearing-impaired listener profile in the demonstrator. In some regions, the required speech levels can be reduced by orienting the head towards the talker location. The effect is, however, rather limited. The standard audiogram N3, which was used in the simulations with an additional distortion component, results in much more elevated hearing thresholds at high frequencies than at low frequencies. The effect of the head shadow, simulated by the ortf receiver, is reduced at low frequencies compared to high frequencies. Possibly, the reduced use of high-frequency components for speech recognition with the impaired profile also reduced the effect of different head orientations in the simulated scene. An important question now is: *Will a hearing device help in that situation?* Probably yes, but, maybe more relevant: *Will it enable communication at normal or raised speech production levels?* The answer will depend on the acoustic scene, the hearing status of the listener, and the hearing device.

3.7 Aided impaired hearing

The simulated hearing device, which was configured to compensate a hearing loss with the standard audiogram N4 according the NAL-NL2 prescription rule, provides compression amplification to the simulated listener with the impaired hearing profile. Figure 8 illustrates the effect of impaired hearing and aided impaired hearing in the demonstrator. The simulations indicate that using the hearing device improves the required speech levels in the whole room, by about two color grades, that is, by approximately 6.7 dB. However, the levels of the normal hearing listener profile are not achieved, leaving the aided impaired listener with a communicative disadvantage, despite using a hearing device. This is one of the most relevant behavioral predictions, because it replicates that the distortion component of hearing loss cannot be compensated by amplification [3].

The simulation results suggest that with the aided impaired listener profile (blue head symbol), no conversation can be realized at speech levels produced with normal or raised effort from most talker positions. Now, the listener has several options to improve the situation. The target speaker could be asked to speak up or come closer, if feasible. This is, however, not always possible and it depends on the communication partners' willingness to cooperate. In addition, the listener could choose from the following options: turning the head, reducing the environmental noise, and obtaining a more suitable hearing device. All options could be interactively explored with the presented GUI, while only the first two (head orientation and noise control) were simulated and are available in the demonstrator.

Figure 9 illustrates the effect of head orientation on the spatial speech recognition map with the aided hearing-impaired listener profile in the demonstrator. While turning the head is predicted to help to lower the required speech levels for many talker positions, it is not sufficient to achieve the speech level equivalent to normal speech production effort. For a more thorough analysis, these levels would have to be contrasted with the speech levels that normal-hearing talkers would typically employ in that situation, which was not performed here.

Figure 10 illustrates the effect of switching the noise sources on and off with the aided hearing-impaired listener profile in the demonstrator. According to the simulation,

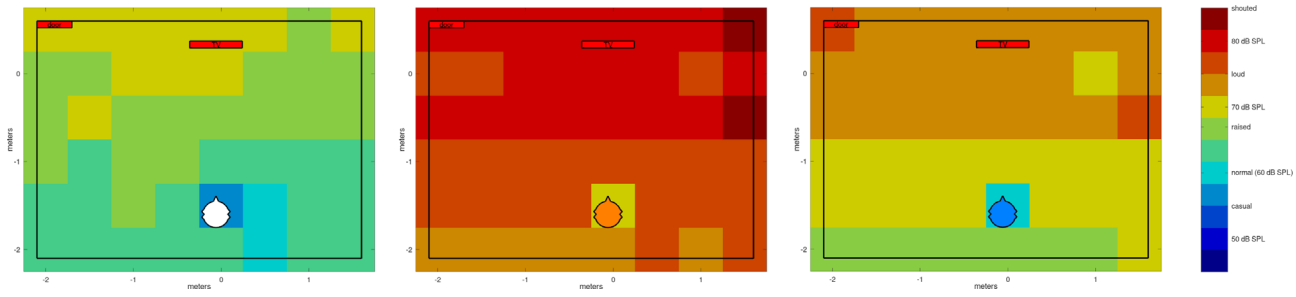

Figure 8. Illustration of the effect of impaired hearing and aided impaired hearing on the spatial speech recognition map. The listener profile is indicated by the head color, where white means normal hearing, orange means impaired hearing, and blue means aided impaired hearing.

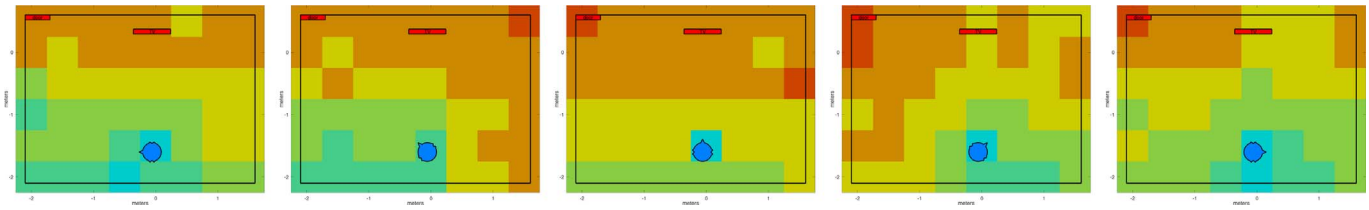

Figure 9. Illustration of the effect of head orientation on the spatial speech recognition map with the aided hearing-impaired profile (indicated by the blue head symbol). The legend of the color encoded speech levels is identical to the one in Figure 3.

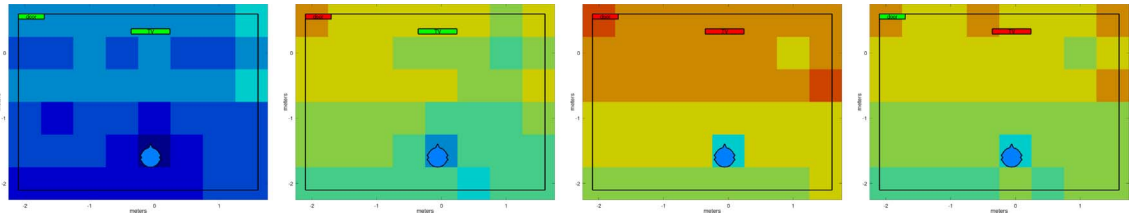

Figure 10. Illustration of the effect of switching on and off noise sources on the spatial speech recognition map with the aided hearing-impaired profile (indicated by the blue head symbol). The legend of the color encoded speech levels is identical to the one in Figure 3.

switching off the TV (second panel) as well as closing the door (last panel) would help to bring the required speech production levels approximately to the same levels that the normal-hearing profile achieves with both noise sources present (cf. first panel in Fig. 8). Interestingly, both noise sources, which had rather different spatial masking patterns for the normal-hearing listener profile (cf. Fig. 5), show a similar masking effect on the aided hearing-impaired listener profile (second and fourth panel). This observation suggests that already low noise levels can result in significant speech masking with aided impaired hearing. In other words, even low level noise can create challenging acoustic conditions for listeners with aided impaired hearing. A possibly relevant hypothesis which could be tested in listening experiments. Only in quiet (first panel), the profile with aided impaired hearing indicates that communication would be possible at speech levels which require casual to normal effort in speech production. This is in line with the notion that hearing devices provide the largest benefit for their users in quiet environments [29].

The last suggested option was to get a pair of suitable hearing devices. As mentioned earlier, additional amplification would most probably not help much, due to the limited benefit of amplification in noise [3]. In Figure 8, the benefits in SRT are spatially very consistent with the modeled hearing device, or in other words, the improvements do not depend on the talker position. For a hearing device with spatial noise reduction, such as, e.g. a beamformer, the results would probably look different. For example, possibly larger improvements could be expected in front of the listener at the cost of lower improvements in other directions. This behavior was not shown in the demonstration example, but would very likely be observed in a corresponding simulation, because multi-channel noise suppression aims to improve the SNR in a way that even listeners with normal hearing would benefit from the signal processing [14]. The directional pattern of a possibly adaptive noise reduction approach should be visible in the spatial speech recognition map, likely showing its spatial advantages as well as disadvantages in a given scene. The benefit of such

a hearing device would also very likely depend on the head direction which could be easily investigated with the interactive representation. While the presented simulations were not validated against the performance of human listeners, they show the potential of simulations to capture elementary limitations of human speech perception, and hence to show ecologically relevant behavior.

3.8 Temporal dynamics of a scene

In this first approach, a spatially static acoustic scene was used. For acoustic scenes with moving objects, e.g. passing cars, it would be possible to break it into a sequence of semi-static scenes which then could be simulated with the presented approach. The spatial representation could be extended with a slider which indicates the temporal position and which would also allow to manipulate it, just like the progress bar in common video players. For each "frame" of such a dynamic scene, all simulations would need to be repeated, which would increase the computational demand greatly. While the semi-static scene frames could be used to measure the speech recognition performance with human listeners, the results would possibly be different from measuring it in the corresponding dynamic scene. However, it is still subject to research how speech recognition performance of human listeners can be reliably measured in dynamic scenes.

3.9 Spatial resolution

The default mesh size of 50 cm was chosen to limit the number of simulations, which is approximately proportional to the square mesh size, but there is virtually no restriction to it. For example, Figure 11 compares a condition rendered with the default mesh size of 50 cm and with an reduced mesh size of 25 cm.

Due to the nearest-neighbor interpolation, the chosen mesh size is clearly visible. The representation would scale well down to very small mesh sizes of less than a centimeter. This is an advantage over more sophisticated interpolation techniques or visualizations, such as, e.g. contour plots. Maps with increasingly higher spatial resolutions can be calculated progressively by repeatedly halving the mesh size, that is, reusing the already calculated results in a higher resolution image.

3.10 Color scale

The simulated outcomes are SRT values on a continuous scale. The chosen color map with 12 colors between 45 dB SPL and 85 dB SPL results in a quantization of the SRTs, where one color grade step corresponds to $\frac{10}{3}$ dB. It was chosen, because it resolves the relevant level range with still easily distinguishable colors. Nonetheless, the quantization can also be chosen to resolve more or even less different SRT values. Figure 12 illustrates the effect of the number of color grades on the speech level representation.

With an increasing number of colors, more and smaller zones of similar SRTs can be distinguished, where, however, the colors of neighboring zones are increasingly difficult to distinguish. Less color steps also help in comparisons across different conditions to identify the same zone. For the example scene, the chosen color scale is a good compromise between discriminability and number of SRT zones. The 12 steps are also sufficient to distinguish levels which correspond to the approximate speech production effort classes casual, normal, raised, loud, and shouted [11].

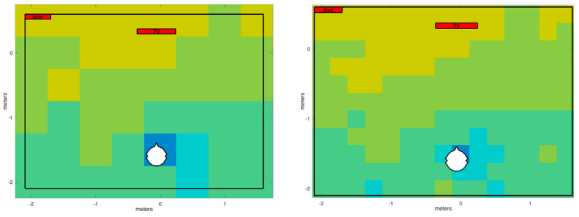

Figure 11. Comparison of spatial speech recognition maps rendered with a mesh size of 50 cm and 25 cm. The legend of the color encoded speech levels is identical to the one in Figure 3.

The quantization of the SRT values seems to result in a loss of information, because the exact values are not shown on the map. It would be true if all SRTs in the scene were so similar that they appeared as the same color (max 3.3 dB difference), which, however, was not observed even for the hearing impaired profile (cf. Fig. 6). But if there is a gradual spatial change in SRT which can be resolved with the chosen mesh size, then the borders of the color zones translate the information in speech level to the spatial dimension. Hence, a part of the information is not lost but translated to spatially discernible features; the zone borders.

3.11 Responsiveness

Because all simulation results were pre-simulated, GUI updates were almost instant. After clicking an element or touching the screen, there was no perceivable delay until the representation was redrawn. This enables the user to rapidly switch between conditions, and compare them from memory. To help comparisons, in the future, the addition of a reference button which sets the current map as a baseline could be implemented. Also, one could easily open two independent instances of the GUI side-by-side, which would allow arbitrary side-by-side comparisons.

For very fast simulation models or approximations thereof, the simulation data could be calculated on demand, which however would reduce the responsiveness of the GUI. In a wider meaning of responsiveness, the simulation approach allows to peek into simulation results and to follow the progress as the simulations run. It is also possible to render the maps first in a lower resolution and then successively halve the mesh size until the desired target resolution is reached. Reducing the time from changing simulation parameters to finishing the corresponding simulations might particularly help in the development of new communicative spaces.

3.12 Extensibility

GNU/Octave was used for the reference implementation due to the wide-spread use of the corresponding script

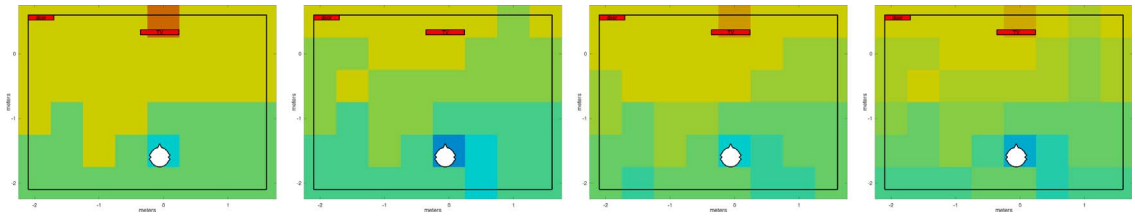

Figure 12. Comparison of representations of a spatial speech recognition map with different numbers of colors: first panel, 8, second panel 12, third panel 16, last panel 24. The only difference is the number of quantization steps between 45 dB SPL and 85 dB SPL.

language in R&D and its low-threshold accessibility. The whole implementation counts less that 300 lines of commented code. It serves as a reference for further research on the accessible interactive representation of abundant simulated spatial speech recognition performance data. Most importantly, comparatively little effort should be required to adapt the concept to a new scene. The visualization does not depend on a specific modeling toolchain. Hence, the employed demonstrator modeling toolchain can be easily extended or even completely replaced. Thus, the proposed accessible visualization should allow comparisons across models and modeling toolchains.

An obvious extension would be to control a real-time rendering instance of the corresponding scene, which would give the user an additional modality to capture the simulated acoustic condition. More sophisticated frameworks are available for visualizing data, and there is technically no reason to implement the visualization in a different environment, or embed the representation into existing applications.

3.13 Applications

As shown in the previous sections, the presented interactive spatial speech recognition map was well-suited to explore the abundant spatial SRT simulation data from the demonstrator with respect to the factors limiting speech recognition performance in a complex acoustic scene. Hence, such maps could be useful in virtually assessing hearing-related technology with respect to speech perception in controlled but possibly ecologically relevant settings. Whether a setting is ecologically valid needs to be investigated separately [25]. A particular strength of such a scene-based approach would be that different scenes can be used to compose sets of ecologically relevant challenges, and use these for objective evaluations. For this application, the modeling toolchain, that is the entire prediction model including all model components, must provide reliable predictions. But such a modeling toolchain currently does not exist or was at least not validated.

Even if the goal of a universal simulation model of speech communication that can be trusted without further validation in different acoustic settings is still not within reach, its creation is highly desirable. The main application of the spatial speech recognition map here is to enable comparability across models and with empirical data in controlled but complex acoustic conditions. The proposed demonstrator modeling toolchain, which consisted mainly of TASCAR, openMHA, and FADE, can be partially or completely exchanged. Predictions with other modeling approaches can be compared in detail, and specific samples also with empirical data, if available. It would also be possible to reconstruct laboratory conditions in simple scenes, e.g. free field with one point-source masker, to isolate certain aspects for validation and comparison purposes, which, however, might be ecologically less relevant. Nonetheless, the visualization of simulations in laboratory – as well as in more complex – conditions might be well-suited for teaching purposes.

Once accurate predictions of individual aided speech recognition performance of listeners with impaired hearing in complex and, more importantly, relevant acoustic scenes, will be possible, new applications may surface. For example, as suggested by [10], an application in room acoustic design could be attractive. At some point, with improving prediction accuracy, more information might be obtained from a model than could be obtained from a listener in speech recognition tests in a given amount of time. While this may sound paradoxical, it is not. If a model correctly predicts the effect of the factors which limit speech recognition performance in relevant acoustic conditions, this information does not have to be obtained from listening experiments. Only a few individual parameters would have to be inferred from listening experiments. This assumes that the speech recognition performance is mainly governed by fundamental principles which work the same for all listeners, and that these principles can be implemented in a model. With a focus on individualized models, which reflect the individual hearing status (e.g. [9]) and maybe individualized HRTFs (e.g. [30]), hearing devices could be individually optimized for complex and, more importantly, relevant acoustic scenes.

4 Conclusions

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

- *Interactive spatial speech recognition maps* were proposed to provide an intuitive representation of predicted spatial speech recognition performance data in complex listening conditions, which can include impaired hearing and hearing devices.
- "Browsing" such a map, which was generated by a modeling toolchain built for demonstration purposes, revealed relevant interactions between parameters which limited the speech recognition performance in typical ways.

- The predictions of the demonstrator model toolchain can be directly compared to the speech recognition performance of human listeners in the same task. This allows the direct validation of the employed model.
- The presented setup is versatile in the sense that it can be adapted to different acoustic scenes and rendering techniques, models of hearing devices, and also speech recognition models and models of impaired hearing. This allows the direct comparison of different modeling toolchains or model components.
- Two major application areas for the proposed representation were identified: First, as a suitable target to compare and validate different modeling approaches in ecologically relevant contexts, and subsequently as a tool to use validated models in the design of spaces and devices which take speech communication into account.

Conflict of interest

The author declared no conflict of interests.

Data availability statement

The employed modeling toolchain was published as a software package on zenodo.org (https://zenodo.org/record/4651595); DOI: https://doi.org/10.5281/zenodo.3734152; https://doi.org/10.5281/zenodo.4651595 [23].

Acknowledgments

The author thanks Julia Schütze for sharing a preliminary version of the living room scene.

Funding

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 352015383 – SFB 1330 A 3.

References

1. G. Keidser, H. Dillon, M. Flax, T. Ching, S. Brewer: The NAL-NL2 prescription procedure, Audiology Research 1, 1 (2011) 88–90. https://doi.org/10.4081/audiores.2011.e24.
2. B. Kollmeier, J. Kiessling: Functionality of hearing aids: State-of-the-art and future model-based solutions. International Journal of Audiology 57, sup3 (2018) S3–S28. https://doi.org/10.1080/14992027.2016.1256504.
3. R. Plomp: Auditory handicap of hearing impairment and the limited benefit of hearing aids. The Journal of the Acoustical Society of America 63, 2 (1978) 533–549. https://doi.org/10.1121/1.381753.
4. M. Vorländer: Auralization, Springer, Cham, 2020.
5. D. Schröder, M. Vorländer: RAVEN: A real-time framework for the auralization of interactive virtual environments, in Forum Acusticum, Aalborg, Denmark, January 2011, 1541–1546.
6. O. Buttler, T. Wendt, S. van de Par, S.D. Ewert: Perceptually plausible room acoustics simulation including diffuse reflections. The Journal of the Acoustical Society of America 143, 3 (2018) 1829–1830. https://doi.org/10.1121/1.5036004.
7. G. Grimm, J. Luberadzka, V. Hohmann: A toolbox for rendering virtual acoustic environments in the context of audiology. Acta Acustica united with Acustica 105, 3 (2019) 566–578. https://doi.org/10.3813/AAA.919337.
8. H. Kayser, T. Herzke, P. Maanen, C. Pavlovic, V. Hohmann: Open master hearing aid (openMHA) – an integrated platform for hearing aid research. The Journal of the Acoustical Society of America 146, 4 (2019) 2879–2879. https://doi.org/10.1121/1.5136988.
9. M.R. Schädler, D. Hülsmeier, A. Warzybok, B. Kollmeier: Individual aided speech-recognition performance and predictions of benefit for listeners with impaired hearing employing FADE. Trends in Hearing 24 (2020) 2331216520938929. https://doi.org/10.1177/2331216520938929.
10. M. Lavandier, S. Jelfs, J.F. Culling, A.J. Watkins, A.P. Raimond, S.J. Makin: Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources. The Journal of the Acoustical Society of America 131, 1 (2012) 218–231. https://doi.org/10.1121/1.3662075.
11. W.O. Olsen: Average speech levels and spectra in various speaking/listening conditions. American Journal of Audiology 7, 2 (1998) 21–25. https://doi.org/10.1044/1059-0889(1998/012).
12. B. Kollmeier, A. Warzybok, S. Hochmuth, M.A. Zokoll, V. Uslar, T. Brand, K.C. Wagener: The multilingual matrix test: Principles, applications, and comparison across languages: A review. International Journal of Audiology 54, Sup2 (2015) 3–16. https://doi.org/10.3109/14992027.2015.1020971.
13. M.R. Schädler, A. Warzybok, S.D. Ewert, B. Kollmeier: A simulation framework for auditory discrimination experiments: Revealing the importance of across-frequency processing in speech perception. The Journal of the Acoustical Society of America 139, 5 (2016) 2708–2722. https://doi.org/10.1121/1.4948772.
14. M.R. Schädler, A. Warzybok, B. Kollmeier: Objective prediction of hearing aid benefit across listener groups using machine learning: Speech recognition performance with binaural noise-reduction algorithms. Trends in Hearing 22 (2018) 2331216518768954. https://doi.org/10.1177/2331216518768954.
15. B. Kollmeier, M.R. Schädler, A. Warzybok, B.T. Meyer, T. Brand: Sentence recognition prediction for hearing-impaired listeners in stationary and fluctuation noise with FADE: Empowering the attenuation and distortion concept by Plomp with a quantitative processing model. Trends in Hearing 20 (2016) 2331216516655795. https://doi.org/10.1177/2331216516655795.
16. N. Bisgaard, M.S. Vlaming, M. Dahlquist: Standard audiograms for the IEC 60118–15 measurement procedure. Trends in Amplification 14, 2 (2010) 113–120. https://doi.org/10.1177/1084713810379609.
17. A.W. Bronkhorst: The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. Acta Acustica united with Acustica 861 (2000) 117–128.
18. J. Rennies, V. Best, E. Roverud, G. Kidd Jr: Energetic and informational components of speech-on-speech masking in binaural speech intelligibility and perceived listening effort. Trends in Hearing 23 (2019) 2331216519854597. https://doi.org/10.1177/2331216519854597.
19. J.A. Grange, J.F. Culling, B. Bardsley, L.I. Mackinney, S.E. Hughes, S.S. Backhouse: Turn an ear to hear: How hearing-impaired listeners can exploit head orientation to enhance their speech intelligibility in noisy social settings. Trends in

Hearing 22 (2018) 2331216518802701. https://doi.org/10.1177/2331216518802701.
20. M.R. Schädler, P. Kranzusch, C. Hauth, A. Warzybok: Simulating spatial speech recognition performance with an automatic-speech-recognition-based model, in: Proceedings of DAGA, Deutsche Gesellschaft für Akustik. 2020, pp. 908–911.
21. R. Beutelmann, T. Brand, B. Kollmeier: Revision, extension, and evaluation of a binaural speech intelligibility model. The Journal of the Acoustical Society of America 127, 4 (2010) 2479–2497. https://doi.org/10.1121/1.3295575.
22. T. Vicente, M. Lavandier, J.M. Buchholz: A binaural model implementing an internal noise to predict the effect of hearing impairment on speech intelligibility in non-stationary noises. The Journal of the Acoustical Society of America 148, 5 (2020) 3305–3317. https://doi.org/10.1121/10.0002660.
23. M.R. Schädler: m-r-s/fade-tascar-openmha: Release of version used for "interactive spatial speech recogition maps" [Data set]. Zenodo, 2021. https://doi.org/10.5281/zenodo.4651595.
24. T. Nuesse, B. Wiercinski, T. Brand, I. Holube: Measuring speech recognition with a matrix test using synthetic speech. Trends in Hearing 23 (2019) 2331216519862982. https://doi.org/10.1177/2331216519862982.
25. G. Keidser, G. Naylor, D.S. Brungart, A. Caduff, J. Campos, S. Carlile, M.G. Carpenter, G. Grimm, V. Hohmann, I. Holube, S. Launer, T. Lunner, R. Mehra, F. Rapport, M. Slaney, K. Smeds: The quest for ecological validity in hearing science: What it is, why it matters, and how to advance it. Ear and Hearing 41, Suppl 1 (2020) 5S–19S. https://doi.org/10.1097/AUD.0000000000000944.
26. P. Bottalico, I.I. Passione, S. Graetzer, E.J. Hunter: Evaluation of the starting point of the Lombard effect. Acta Acustica united with Acustica 103, 1 (2017) 169–172. https://doi.org/10.3813/AAA.919043.
27. G. Grimm, J. Luberadzka, V. Hohmann: Virtual acoustic environments for comprehensive evaluation of model-based hearing devices. International Journal of Audiology 57, sup3 (2018) S112–S117. https://doi.org/10.1080/14992027.2016.1247501.
28. J.A. Grange, J.F. Culling: The benefit of head orientation to speech intelligibility in noise. The Journal of the Acoustical Society of America 139, 2 (2016) 703–712. https://doi.org/10.1121/1.4941655.
29. H. Dillon: Hearing aids, Thieme Medical Publishers, 2012.
30. R. Pelzer, M. Dinakaran, F. Brinkmann, S. Lepa, P. Grosche, S. Weinzierl: Head-related transfer function recommendation based on perceptual similarities and anthropometric features. The Journal of the Acoustical Society of America 148, 6 (2020) 3809–3817. https://doi.org/10.1121/10.0002884.

Cite this article as: Schädler MR. 2022. Interactive spatial speech recognition maps based on simulated speech recognition experiments. Acta Acustica, **6**, 31.