Audio Engineering Society

Conference Paper

Presented at the 2022 International Conference on
Audio for Virtual and Augmented Reality
2022 August 15–17, Redmond, WA, USA



# Measuring audio-visual speech intelligibility under dynamic listening conditions using virtual reality

Alastair H. Moore[1], Tim Green[2], Mike Brookes[1], and Patrick A. Naylor[1]

[1]*Electrical and Electronic Engineering, Imperial College London, London, UK*
[2]*Speech, Hearing and Phonetic Sciences, University College London, London, UK*

Correspondence should be addressed to Alastair H. Moore (alastair.h.moore@imperial.ac.uk)

**ABSTRACT**

The ELOSPHERES project is a collaboration between researchers at Imperial College London and University College London which aims to improve the efficacy of hearing aids. The benefit obtained from hearing aids varies significantly between listeners and listening environments. The noisy, reverberant environments which most people find challenging bear little resemblance to the clinics in which consultations occur. In order to make progress in speech enhancement, algorithms need to be evaluated under realistic listening conditions. A key aim of ELOSPHERES is to create a virtual reality-based test environment in which alternative speech enhancement algorithms can be evaluated using a listener-in-the-loop paradigm. In this paper we present the sap-elospheres-audiovisual-test (SEAT) platform and report the results of an initial experiment in which it was used to measure the benefit of visual cues in a speech intelligibility in spatial noise task.

## 1 Introduction

The cocktail party problem posed by Cherry [1] has been widely studied over nearly seven decades and many factors have been shown to affect the intelligibility of speech in noisy enviroments. In particular Cherry's observation that spatial separation of a target and masker improves intelligibility, commonly known as spatial release from masking (SRM), is less effective in Hearing-Impaired (HI) listeners and in reverberant environments [2]. Whilst hearing aids (HAs) are very effective at restoring audibility, "trying to follow a conversation in the presence of noise" is the listening situation with the lowest satisfaction amongst HA users [3, 4].

As predicted in [1], lip reading can provide a substantial benefit to speech intelligibility [5]. Therefore, in the context of a group conversation, turning the head and/or shifting one's gaze to observe the target talker might be expected to be beneficial. However, it has also been shown [6, 7] that, in unaided conditions, orienting one's head away from the target can improve intelligibility, due to a higher Signal-to-Noise Ratio (SNR) at the better ear.

To date, the most effective noise suppression strategy in HAs is beamforming, in which sounds arriving from non-frontal directions are attenuated. If the user faces the active talker, which may be optimal for lip-reading, the onsets of other talkers may be suppressed, until such time has the head is turned. If the user turns slightly away from the active talker, which may maximise the

SNR at one ear, the beamformer may actually suppress the target.

Since head motion is an integral part of normal listening behaviour and a listener does not necessarily face the target talker, adapting the speech enhancement to account for the head movement would seem logical. Indeed the sensors and algorithms required to track head orientation/position and identify potential source directions are already available in several commercially available virtual reality (VR) headsets and augmented reality (AR) devices are also nascent [8]. Several multichannel acoustic signal processing algorithms which incorportate knowledge of the changing array orientation have recently been proposed [9, 10, 11, 12] and objective metrics based on models of speech intelligibility suggest they would be beneficial.

Measuring the real-world benefit of rotation-aware beamforming algorithms for head-worn microphones arrays, whether in the context of hearing aids or AR, is challenging and is the focus of the current work. Listening to binaural audio which has been recorded/processed under different head motion conditions to those encountered during playback is not helpful because the incongruent cues lead to a confused sense of space. This problem can only be solved by modifying the microphone signals, and therefore also the subsequent processing, according to the listener's own head movement — a paradigm which has been called "listener-in-the-loop" (LITL).

In this paper we present a new research tool for conducting LITL experiments. Its name derives from the Speech and Audio Processing (SAP) lab's convention of prefixing 'sap' to our repository names and the project name — Environment and Listener Optimised Speech Processing for Hearing Enhancement in Real Situations (ELOSPHERES). The sap-elospheres-audiovisual-test (SEAT) platform is open source and available from [13].

The remainder of this paper is organised as follows. Section 2 reviews current approaches to LITL evaluation of hearing aid speech enhancement and motivates the proposed approach. Section 3 presents an overview of the SEAT platform while Section 4 describes the implementation of the audiovisual rendering component. Section 5 describes a listening experiment conducted to validate the platform. The paper ends with a discussion of the software roadmap and conclusions in Section 6.

## 2 Listener-in-the-loop (LITL)

The goal of LITL experiments is to evaluate the combined performance, in this case speech intelligibility, of the listener and any speech enhancement algorithm together, since the behaviour of either will affect the behaviour of the other. In all cases the speech enhancement algorithm must operate online, i.e. in real time, and with an acceptibly small latency. Note that acceptible latency could be a parameter under test in a LITL experiment.

LITL experiments can be broadly categorised according to the nature of the sound field encountered by the microphone array.

**Natural** The sound field is naturally occuring due to one or more live talkers. Conversation dynamics are natural and interlocutors can react to the listener's behaviour. This is as close to real life as an experiment can get. However it is neither repeatable nor straightforward to measure important parameters, such as SNR of the talker(s).

**Controlled** The sound field is created by playing known anechoic signals through loudspeakers where each loudspeaker represents an independent source. One can easily control Signal-to-Interference Ratio (SIR) by varying the presentation level of each loudspeaker but to vary the spatial properties of the scene requires physical changes to the environment, or to repeat the experiment in a different location.

**Reconstructed** A sound scene which has been recorded in advance, simulated in advance, or simulated in real-time is rendered using an array of loudspeakers. Computation of the driving signal for each loudspeaker can vary depending on how the sound field is represented. The accuracy of the reconstructed sound field is potentially limited by the density of the loudspeaker array and the acoustics of the room. However, as with Natural and Controlled experiments, the inputs to the speech enhancement algorithm are sampled live from the environment and so inherently include the effect of head motion.

**Simulated** The microphone signals encountered in a sound scene are calculated directly from the sound scene representation, taking into account the listener's head movement. As with Reconstructed scenes, that representation could originate from a recording of a real environment or a simulation. Using simulated

microphone signals offers huge flexibility, since there is no dependence on the listener's local environment. Of course, simulated acoustics are used extensively in VR/AR systems. Whereas those systems deliver binaural (i.e. 2-channel) output and can make perceptually-validated simplifications to reduce computational complexity, simulated acoustics in this context requires several output channels. Moreover, the simulation must be sufficiently accurate that processing of those signals by the speech enhancement algorithm is equivalent to processing of signals obtained in a real environment.

For the ELOSPHERES project, we have chosen to develop our platform using the Simulated approach. The primary motivation for this is the flexibility it offers in terms of where experiments can be conducted and therefore the potential for more researchers to conduct experiments using the platform. Ultimately we hope that audiology clinics will be able to use these tools during hearing aid fitting in their existing premises.

## 3 System Overview

The SEAT platform is designed as an overarching framework for audiovisual listening experiments. The expectation is that these experiments will be conducted in VR but the platform does not depend on or require this. The main SEAT program is a Python script which co-ordinates running a block of trials, each of which is presented within the context of a scene. This is achieved using three main modules, which abstract the implementation details

**AVRendererControl** Deals with the actual presentation of stimuli.

**ResponseMode** Deals with obtaining responses, either from the listener or the experimenter, depending on the test design.

**ProbeStrategy** Defines how test conditions (e.g. SIR) should be varied according to the listener's responses.

A fourth module takes care of logging the trial-by-trial results as the block progresses. Each module has a defined application programming interface (API) so that alternative implementations of each module can be used, according to the requirements of the experiment. A top-level configuration file defines which implementation of each module should be used, along with any settings.

At present, ResponseMode has implementations for a signal detection task, where the listener reports true or false for each trial, and a speech intelligibility task, where the listener reports what words they heard and the experimenter enters which keywords were correctly identified. Implementations for ProbeStrategy include fixed SNR, an adaptive procedure targeting a specific accuracy and a dual-track adaptive procedure, where trials for each track are interspersed.

## 4 AVRendererControl implementation

A block diagram of the main software components of SEAT platform are shown in Fig. 1. In an effort to follow sustainable software development practices, many components of the system are re-used from existing software resources. The major contribution here is to provide the necessary glue and flow control to build a VR-based listening test.

All the materials required for a scene are contained within a single folder. In the context of a whole experiment, this may lead to duplication of assets, but ensures that this one folder is self-contained. The assets can be assembled by hand but it is more convenient to generate them using a script. Example scripts can be obtained by contacting the first author.

### 4.1 Visual rendering

Visual components of the scene are rendered using a VR headset tethered to a PC running Windows 10. The VR Application (VRApp) is developed in Unity [14] and is a modified version of ListeningEffortPlayer [15], which was developed by colleagues in a parallel project with similar aims. The application provides four configurable canvases upon which videos can be projected. One is a 'skybox' which completely surrounds the virtual space. Projecting a 360° video or photo onto the skybox gives the user the impression of being in a particular room. The remaining canvasses are rectangular and are intended for projecting talking head videos corresponding to target/interferers. These canvases might, for example, appear to the user as TV screens within the scene. The videos to be played and their locations are controlled using Open Sound Control (OSC) [16] commands. The VRApp takes care of updating the visual rendering according to the orientation of the user's head. It also broadcasts this orientation using OSC, enabling the audio rendering to be dynamically updated.

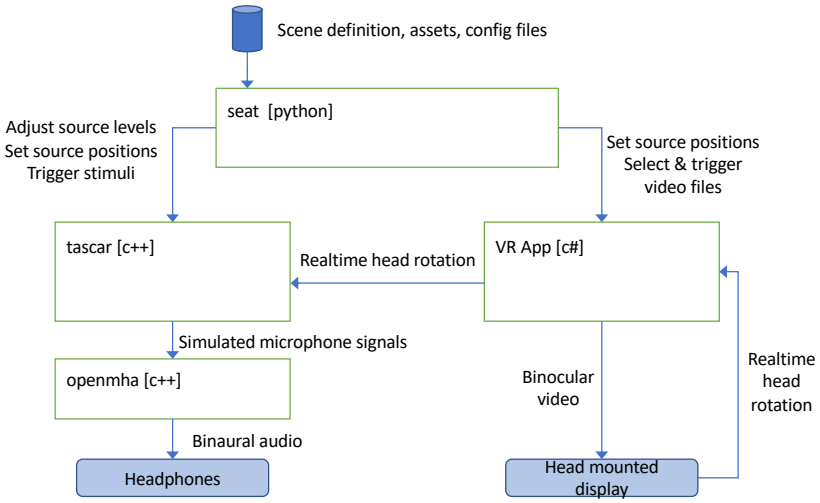



**Fig. 1:** System diagram showing main software blocks

Our modifications [17] transform the head orientation into the correct format before sending it to the required OSC address for our audio renderer.

### 4.2 Audio rendering

Audio rendering is achieved using Toolbox for Acoustic Scene Creation And Rendering (TASCAR) [18] running on Ubuntu under Windows Subsystem for Linux (WSL). A scene is defined using an xml-formatted configuration file. Briefly, a scene consists of one or more audio sources and one or more receivers. The positions and levels of these can be time varying according to parameter values specified in the file and/or in real-time using OSC.

In the current implementation, three point sources are defined, corresponding to the three rectangular video canvases. Audio files corresponding to talking head videos are loaded at the start of the scene and positioned/triggered using OSC. Additionally, comfort noise is implemented as a fourth point source, present throughout the whole scene, with a very low level.

The receiver uses a spherical head model with parametric filters to simulate the Head-Related Transfer Function (HRTF) giving a binaural output. The microphone signals are passed via the Jack Audio Connection Kit (JACK) [19] to OpenMHA [20] where speech enhancement can be performed.

Audio samples are transmitted from the JACK server running on WSL to a JACK server running on Windows using JackTrip [21]. On Windows, JACK uses low-latency ASIO drivers to output the audio via the sound card's headphone port.

## 5 System evaluation

In order to validate the SEAT platform an experiment was conducted which investigated factors that affect how much benefit to speech-in-noise recognition is obtained from seeing the talker in addition to hearing them. Previous evidence [5] suggests that audio-visual (AV) benefit is greater when the masker consists of two interfering talkers than when it is speech-shaped noise (SSN). It has been inferred that, as well as providing phonetic information, the visual signal aids in segregating the target speech stream from interfering speech. This benefit might be expected to be smaller when auditory spatial cues can be used to segregate sound sources and, consistent with this, [22] found less AV benefit when the precedence effect was used to perceptually separate target speech from 2 interfering talkers, compared to when maskers and target were co-located. Here, we look for a similar effect but in this case using the SEAT platform to present the stimuli in VR.

### 5.1 Methods

17 normally hearing adults participated (14 female, mean age 23.5, range 18-51). Visual stimuli were pre-

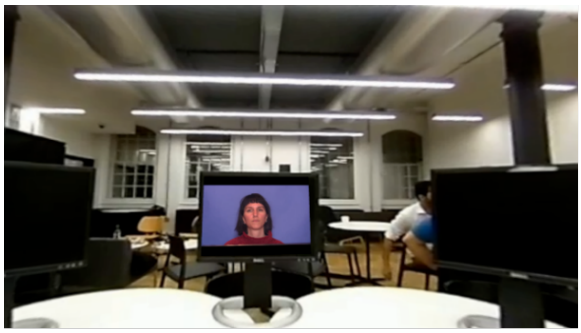


**Fig. 2:** Screenshot from café video presented via Vive headset

sented via a HTC Vive Pro Eye headset. Sound was presented via Sennheiser HDA-200 headphones. In all conditions a 360° video recording from a café was played. Three computer monitors were located on a table within the café. In conditions which included visual presentation of target speech, the video of the talker was presented on the middle of the three monitors (Fig 2). The other monitors remained blank throughout.

Target speech consisted of IEEE sentences spoken by a female talker of Southern British English. Masking noise was either segments of connected passages from two other female talkers (2T) or two independent segments of speech-shaped noise (SSN). On each trial maskers began 1.5 s before target speech and lasted for 7 s. Maskers were either both co-located with the target speech (0° azimuth) or symmetrically spaced either side of the target at ±40° azimuth. Presentation of target speech was either audio-visual (AV) or audio-only (AO), resulting in a total of 8 conditions. Presentation of maskers was always audio-only. Two runs of 30 trials were completed in each condition with each run containing two interleaved adaptive tracks tracking 20 % and 80 % correct key word recognition. Speech reception thresholds (SRTs, signal-to-noise ratio for 50 % correct) were derived from fitted psychometric functions.

### 5.2 Results

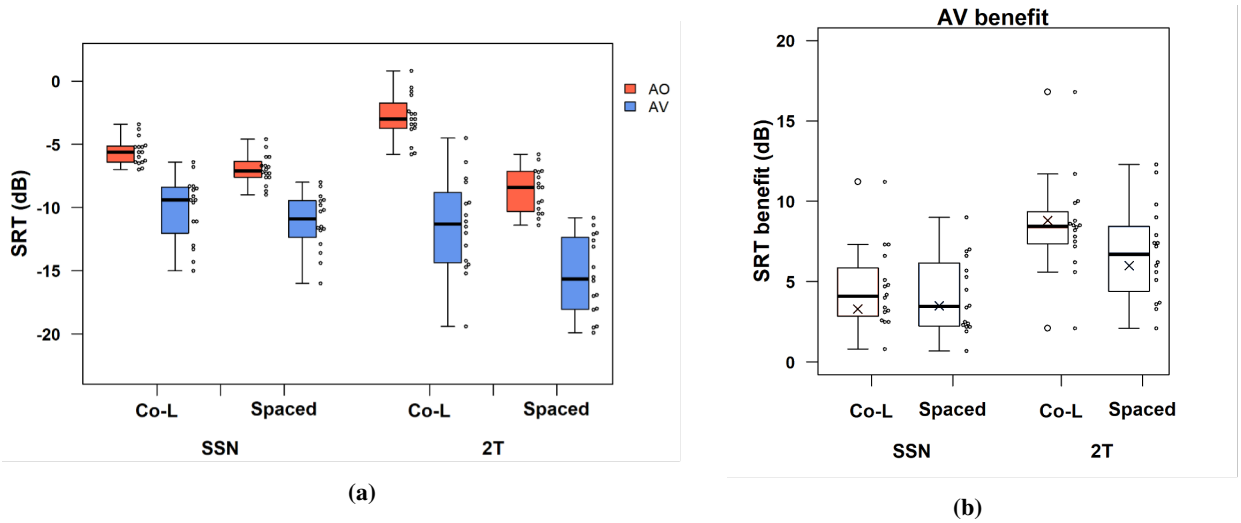
Data from one outlier were excluded due to exceptionally low SRTs in AV conditions. SRTs for all conditions for the remaining participants are shown in Figure 3(a). One notable aspect of the data is that, consistent with previous comparisons between energetic and informational masking, for AO presentation there is a substantially bigger effect of spatial separation for 2T maskers than for SSN. It is also noticeable that, for both masker types, relative to co-located AO performance the addition of visual information gives greater benefit than the availability of auditory spatial cues.

Figure 3(b) shows differences in SRT between AO and AV presentation for each combination of masker type and spatial configuration. Consistent with expectations, AV benefit was larger with competing speech than with SSN, but was reduced when auditory spatial cues were available. The crosses in Figure 3(b) show mean AV benefit from the similar conditions in [22] which closely match the present data.

AV benefit data were analysed with a linear mixed model using Satterthwaite's approximation for degrees of freedom. There were significant main effects of masker type ($p < 0.001$) and spatial configuration ($p = 0.038$), but no significant interaction ($p = 0.22$).

### 5.3 Discussion

A key goal of the study was to validate the SEAT platform by replicating previous findings relating to audio-visual speech intelligibility. The similarity between the present data and that of [22] is therefore encouraging. As suggested in [22], two possible (non-exclusive) explanations for the differences in AV benefit across conditions are 1) that in the presence of informational masking, visual cues not only give additional phonetic information beyond that available in the audio signal, but also provide information about the amplitude contour of the target speech which aids segregation of the target speech stream from the interfering sources, and 2) that the contribution of phonetic information derived from lip-reading may be greater for fluctuating than for steady-state maskers. Speech perception in real environments typically is audio-visual, particularly for the hearing-impaired, so that it is important to assess how information from both modalities contributes to intelligibility. However, a limitation of the present study is that maskers were never presented audio-visually. The SEAT platform provides the capability to test speech understanding in more complex and realistic situations in which there are multiple possible talkers who may be speaking at the same time with uncertainty about which talker should be attended at particular times.




(a) (b)

**Fig. 3:** (a) SRTs for each combination of masker type (SSN: speech shaped noise, 2T: two talkers) and spatial configuration (Co-L: co-located, Spaced: $\pm 40°$) for audio-only (red) and audio-visual (blue) presentation. Open circles immediately to the right of each box plot show individual data points. (b) Audio-visual benefit (decrease in SRT). Crosses show mean benefit in the similar conditions of [22].

### 5.4 Reproducibility

To faciliate reproduction of the present experiment, instructions for setting up JackTrip and WSL are available from https://github.com/alastairhmoore/test-jacktrip-wsl, software versions are reported in Table 1 and materials are available from [23].

## 6 Software roadmap and conclusions

This paper has reported on the design of the SEAT platform and the specific implementation used for the present validation experiment. It has been shown that our VR-based speech intelligibility test delivers very similar SRM and visual benefit as has been reported in real-world studies.

There are many more pyschophysical factors which could be evaluated with only small modifications to the configuration files, for example, the effect of masker spacing, the effect of visible versus audio-only maskers, the effect of non-stationary and/or diffuse background noise.

An important capability for the study of speech enhancement in hearing aids is the simulation of microphone array signals. This is already possible in TASCAR using several receiver types. We are currently investigating the accuracy/computational cost trade offs associated with different rendering pipelines [24]. With multichannel signals output from TASCAR it will be possible to enable the existing beamforming algorithms within OpenMHA.

The current study used a relatively simple scene. In the future more complex listening situations, including early reflections and reverberation will be realised.

The SEAT software [13] is open source and suggestions for improvements are welcome.

## Acknowledgement

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/S035842/1].

## References

[1] Cherry, E. C., “Some experiments on the recognition of speech, with one and with two ears,” *J. Acoust. Soc. Am.*, 25(5), pp. 975–979, 1953, doi: 10.1121/1.1907229.

**Table 1:** Software respositories and versions

| Name | Host | Installation method | Version/tag |
|---|---|---|---|
| SEAT | Windows 10 | source [13] | v0.3 |
| ListeningEffortPlayer | Windows 10 | source [17] | seat_v0.3 |
| JACK | Windows 10 | binary [19] | 1.9.11 |
| JackTrip | Windows 10 | binary [21] | 1.2.1 |
| TASCAR | WSL Ubuntu 20.04 | apt | 0.214 |
| JACK | WSL Ubuntu 20.04 | apt | 1.9.12 |
| JackTrip | WSL Ubuntu 20.04 | source [21] | 1.3.0 |

[2] Marrone, N., Mason, C. R., and Kidd, G., "The effects of hearing loss and age on the benefit of spatial separation between multiple talkers in reverberant rooms," *J. Acoust. Soc. Am.*, 124(5), pp. 3064–3075, 2008, ISSN 0001-4966, doi: 10.1121/1.2980441.

[3] Kochkin, S., "MarkeTrak VIII: Consumer satisfaction with hearing aids is slowly increasing," *The Hearing J.*, 63(1), pp. 19–20, 2010, ISSN 0745-7472, doi:10.1097/01.HJ.0000366912.40173.76.

[4] Abrams, H. B. and Kihm, J., "An Introduction to MarkeTrak IX: A New Baseline for the Hearing Aid Market," 2015.

[5] Avivi-Reich, M., Puka, K., and Schneider, B. A., "Do age and linguistic background alter the audiovisual advantage when listening to speech in the presence of energetic and informational masking?" *Atten. Percept. Psychophys.*, 80(1), pp. 242–261, 2018, ISSN 1943-393X, doi:10.3758/s13414-017-1423-5.

[6] Kock, W. E., "Binaural localization and masking," *J. Acoust. Soc. Am.*, 22(6), pp. 801–804, 1950, ISSN 0001-4966, doi:10.1121/1.1906692.

[7] Grange, J. A. and Culling, J. F., "The benefit of head orientation to speech intelligibility in noise," *J. Acoust. Soc. Am.*, 139(2), pp. 703–712, 2016, ISSN 0001-4966, doi:10.1121/1.4941655.

[8] Donley, J., Tourbabin, V., Lee, J.-S., Broyles, M., Jiang, H., Shen, J., Pantic, M., Ithapu, V. K., and Mehra, R., "EasyCom: An augmented reality dataset to support algorithms for easy communication in noisy environments," *arXiv:2107.04174 [cs, eess]*, 2021.

[9] Tourbabin, V. and Rafaely, B., "Direction of arrival estimation using microphone array processing for moving humanoid robots," *IEEE/ACM Trans. Audio, Speech, Language Process.*, 23(11), pp. 2046–2058, 2015, ISSN 2329-9290.

[10] Zohourian, M. and Martin, R., "Binaural speaker localization and separation based on a joint ITD/ILD model and head movement tracking," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, pp. 430–434, 2016, doi:10.1109/ICASSP.2016.7471711.

[11] Moore, A. H., Lightburn, L., Xue, W., Naylor, P. A., and Brookes, M., "Binaural mask-informed speech enhancement for hearing aids with head tracking," in *Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, pp. 461–465, Tokyo, Japan, 2018, doi:10.1109/IWAENC.2018.8521361.

[12] Moore, A. H., Xue, W., Naylor, P. A., and Brookes, M., "Noise covariance matrix estimation for rotating microphone arrays," *IEEE/ACM Trans. Audio, Speech, Language Process.*, 27(3), pp. 519–530, 2019, ISSN 2329-9304, doi:10.1109/TASLP.2018.2882307.

[13] "SEAT: sap-elospheres-audiovisual-test," https://github.com/ImperialCollegeLondon/sap-elospheres-audiovisual-test, 2022.

[14] "Unity 3d." https://unity3d.com, 2019.

[15] Murray-Browne, T., "Listeningeffortplayer," https://github.com/timmb/ListeningEffortPlayer, 2020.

[16] Freed, A., "Open sound control: A new protocol for communicating with sound synthesiz-

ers," in *International Computer Music Conference (ICMC)*, 1997.

[17] Murray-Browne, T. and Moore, A. H., "Listeningeffortplayer (fork)," https://github.com/alastairhmoore / timmbListeningEffortPlayer, 2020.

[18] Grimm, G., Luberadzka, J., and Hohmann, V., "A toolbox for rendering virtual acoustic environments in the context of audiology," *Acta Acustica united with Acustica*, 105(3), pp. 566–578, 2019, doi:10.3813/AAA.919337.

[19] "Jack audio connection kit," https://jackaudio.org, 2021.

[20] Kayser, H., Herzke, T., Maanen, P., Zimmermann, M., Grimm, G., and Hohmann, V., "Open community platform for hearing aid algorithm research: Open Master Hearing Aid (openMHA)," *SoftwareX*, 17, 2022, ISSN 23527110, doi:10.1016/j.softx.2021.100953.

[21] "JackTrip," https://jacktrip.github.io/jacktrip/, 2020.

[22] Helfer, K. and Freyman, R., "The role of visual speech cues in reducing energetic and informational masking," *J. Acoust. Soc. Am.*, 117, pp. 842–849, 2005.

[23] "Materials used in "measuring audio-visual speech intelligibility under dynamic listening conditions using virtual reality"," 2022, doi:10.5281/zenodo.6889160.

[24] Moore, A. H., Vos, R. R., Naylor, P. A., and Brookes, M., "Processing pipelines for efficient, physically-accurate simulation of microphone array signals in dynamic sound scenes," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2021.