

# AuViST - An Audio-Visual Speech and Text Database for the Heard-Text-Recall Paradigm

Cosima A. Ermert<sup>1</sup> , Chinthusa Mohanathanan<sup>2</sup> , Jonathan Ehret<sup>3</sup> ,  
Sabine J. Schlittmeier<sup>2</sup> , Torsten W. Kuhlen<sup>3</sup> , Janina Fels<sup>1</sup> 

<sup>1</sup> Institute for Hearing Technology and Acoustics, RWTH Aachen University, Germany

<sup>2</sup> Work and Engineering Psychology, Institute of Psychology, RWTH Aachen University, Germany

<sup>3</sup> Visual Computing Institute, RWTH Aachen University, Germany

Corresponding Author: [cosima.ermert@akustik.rwth-aachen.de](mailto:cosima.ermert@akustik.rwth-aachen.de)

**Abstract:** The Audio-Visual Speech and Text (AuViST) database provides additional material to the heard-text-recall (HTR) paradigm by Schlittmeier et al. [1]. German audio recordings in male and female voice as well as matching face tracking data are provided for all texts.

**Keywords:** heard-text recall paradigm, HTR, audio recordings, face tracking, text recall, text comprehension, memory, conversational content, two-talker conversation

## 1 Introduction

When verbal short-term memory performance is investigated in cognitive psychology, oftentimes unrelated digits or single words are employed as stimuli (cf. [2]). In everyday scenarios, however, people are predominantly challenged to memorize coherent content instead of isolated digits or words. To fill this gap, Schlittmeier et al. [1] developed the heard-text-recall (HTR) paradigm for investigating text comprehension and memory. Their database provides 34 German texts (Set<sub>2</sub>). These texts can be presented as a two-talker conversation to assess short-term memory for close-to-real-life listening situations as Fintor et al. [3] have done. Here, participants listened to a short conversation between two speakers (male and female) and are asked content-related questions afterwards. Listening effort was measured simultaneously via a dual-task design.

The present database provides audio-recordings and face tracking data from a male and a female speaker for the text recall paradigm and text database provided by [1] (Set<sub>2</sub>). Our material allows to realize the HTR paradigm (cp. [1]) either for running speech spoken by one speaker (male or female) or as a conversation between these two speakers. Audio recordings with both a male and a female speaker were performed (see Section 2.3). To add visual stimuli in the form of speaking virtual humans in a virtual reality (VR) environment in the future, face tracking data was collected during the recordings (see Section 2.4). All texts and the collected data can be downloaded as the Audio-Visual Speech

and Text (AuViST) database. In [1], completed and blinded genograms are available for each text.

## 2 Procedure

### 2.1 Description of Content and Presentation as Running Speech from one Talker or as a Two-Talker Conversation

Each text describes three generations of a family (grandparents, parents, and children) considering different aspects such as age, profession, or hobbies of the individual family members, as well as their relationship with each other [1]. In each text, 5-6 people are mentioned by name. The texts consist of 120-131 words and 10 sentences, with at most one subordinate clause in a sentence. Thus, the texts are comparable in length. In this publications, suggestions are given on how the 10 sentences of one text can be distributed between two conversational partners (e.g., a female and a male talker) to simulate a conversation. The turn-taking between the two conversational partners aims to simulate a typical conversation. Therefore, successive sentences that form a unit of meaning are assigned to one conversational partner before the other partner takes over. The speaking time between the two conversational partners is approximately the same.

In the HTR paradigm, for each text, nine questions are asked about the names of family members, their relationship to each other, and further information (e.g., profession, locations, hobbies, age) [1]. In



the `AuViST_HTR_TextMaterial.txt` file, the first three questions refer to a specific person, the next three questions to a relationship between two family members, and the final three questions ask for a fact, age, or location. Most questions are so-called indirect questions since information must be integrated across sentences to answer the questions. Besides, there are direct questions which can be answered on the basis of the information given in a single sentence. Each question can be answered with 1-2 words. No question can be answered with a simple yes/no. Only one question is asked for each name, fact, or relationship. It was ensured that names, facts, or events that were asked about were not mentioned in previous questions. The questions were not recorded and are thus only included in the provided text files (see Section 3). For further details on the texts and questions, see [1].

## 2.2 Speakers

The texts were recorded sentencewise by one male (21 years) and one female (34 years) German native speaker. They were both experienced in stage performance and asked to use normal intonation and volume during the recordings. Their fundamental frequencies were 120 Hz for the male and 175 Hz for the female speaker.

The Acoustic Voice Quality Index (AVQI) after Maryn et al. [4] was computed for the two voices using the plugin by PHONANIUM<sup>1</sup> (Version 02.03) for PRAAT [5] (Version 6.1.47). The analysis revealed an AVQI of 2.49 for the male and 2.39 for the female speaker. As AVQI values below 2.7 are considered normal for German language [6], both speakers were found to have a healthy voice.

## 2.3 Audio Recordings

The recordings were performed in the hemi-anechoic chamber of the RWTH Aachen University ( $l \times w \times h = 12.6 \times 7.57 \times 5.3 \text{ m}^3$ ), with a lower frequency limit of 100 Hz [7]. The approximately rigid floor ( $\alpha = 0.02$ ) was covered with porous acoustic absorbers in a radius of approximately 1.5 m around the speaker. During the measurement, speakers were seated on bar chair with a height of 0.76 m. The microphone, a Neumann KM 184 mt with cardioid characteristics, and a pop filter were kept at a distance of approximately 1 m from the speakers. The signals were transmitted to Reaper (Version 5.80/x64) via an Octamic RME and a Hammerfall DSP Multiface II. Recordings were done with a sampling frequency of 48 kHz (24 bit).

The acoustic post-processing was done in the audio editing software *Audacity(R)* [8] (Version 3.0.5). Each sentence was extracted by cutting at the respective beginning and end. If multiple recordings existed of the same sentence, a selection was made regarding

subjective preference. Normalization of the perceived loudness was performed according to European Broadcasting Union (EBU) standard R128 [9] towards an average loudness of  $-23$  Loudness units relative to Full Scale (LUFS).

## 2.4 Face Tracking

While recording audio, facial movements of the speakers were captured to enable animation of corresponding speaking virtual agents' faces. For this Apple's *TrueDepth* sensor was used since it worked better in pretests than purely RGB-based solutions, e.g., *OpenFace 2.0* [10]. The recordings were done in the *Live Link Face* app<sup>2</sup> (Version 1.1.1 (1)) with an *iPhone XR* (Version iOS 14.7.1), which records face animations at 100 Hz and writes timestamps and the activation values of 61 blend shapes into a `.csv` file. Additionally it also synchronously captures an RGB video. At the beginning of the recording session a calibration was performed using the built in functionality of the app. The `.csv` files were cut to exactly match the timings of the cut audio files.

## 3 Files

All datasets described in this report can be downloaded from <https://doi.org/10.18154/RWTH-2023-05543>.

The download consists of the audio recordings in `.wav` format, the face tracking data in `.csv` format, and all text material in written form [1] including suggested turns as well as questions and answers in a `.txt` file and in a machine-readable `.json` file. All audio files and 34 texts with corresponding questions are in German language. The audio recordings and face tracking data are labeled as follows: `HTR_[modality]_[gender of speaker]_t[text number]s[sentence number]`, with the modality being *audio* for the audio recordings and *face* for the face tracking data; and the gender of the speaker abbreviated as *m* for the male and *f* for the female speaker. As an example, the audio recording for text 1, sentence 4 with the male voice can be found in `HTR_audio_m_t1s4.wav` and the corresponding face tracking data in `HTR_face_m_t1s4.csv`. The sampling rate for both the audio and the face tracking data is 48 kHz.

## Acknowledgments

The authors want to thank the two speakers for recording the stimuli, Nils Rummler for support regarding the audio recording setup, Moritz Bender and Oliver Renaldi for assisting with the audio post-processing, and Isabel Schiller for carrying out the voice analysis.

## Funding

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): SPP2236 -

<sup>1</sup> <https://www.phonanium.com/product/acoustic-voice-quality-index/>

<sup>2</sup> <https://apps.apple.com/us/app/live-link-face/id1495370836/>

444724862; Listening to, and remembering conversations between two talkers: Cognitive research using embodied conversational agents in audiovisual virtual environments.

### ORCID iD

Cosima A. Ermert  <https://orcid.org/0000-0002-4884-817X>

Chinthusa Mohanathanasan  <https://orcid.org/0000-0001-6916-1425>

Jonathan Ehret  <https://orcid.org/0000-0001-6270-5119>

Sabine J. Schlittmeier  <https://orcid.org/0000-0001-9051-4547>

Torsten W. Kuhlen  <https://orcid.org/0000-0003-2144-4367>

Janina Fels  <https://orcid.org/0000-0002-8694-7750>

### References

- [1] S. J. Schlittmeier, C. Mohanathanasan, I. S. Schiller, and A. Liebl, “Measuring text comprehension and memory: A comprehensive database for Heard Text Recall (HTR) and Read Text Recall (RTR) paradigms, with optional note-taking and graphical displays,” *RWTH Publications*, 2023. DOI: [10.18154/RWTH-2023-05285](https://doi.org/10.18154/RWTH-2023-05285).
- [2] J. T. E. Richardson, “Measures of Short-Term Memory: A Historical Review,” *Cortex*, vol. 43, no. 5, pp. 635–650, Jan. 2007. DOI: [10.1016/S0010-9452\(08\)70493-3](https://doi.org/10.1016/S0010-9452(08)70493-3).
- [3] E. Fintor, L. Aspöck, J. Fels, and S. J. Schlittmeier, “The role of spatial separation of two talkers’ auditory stimuli in the listener’s memory of running speech: Listening effort in a non-noisy conversational setting,” *International Journal of Audiology*, vol. 61, no. 5, pp. 371–379, May 2022. DOI: [10.1080/14992027.2021.1922765](https://doi.org/10.1080/14992027.2021.1922765).
- [4] Y. Maryn, M. De Bodt, and N. Roy, “The Acoustic Voice Quality Index: Toward improved treatment outcomes assessment in voice disorders,” *Journal of Communication Disorders*, vol. 43, no. 3, pp. 161–174, 2010 May-Jun. DOI: [10.1016/j.jcomdis.2009.12.004](https://doi.org/10.1016/j.jcomdis.2009.12.004).
- [5] P. Boersma and D. Weenink, *Praat: Doing phonetics by computer [Computer application]*, 1992.
- [6] B. Barsties and Y. Maryn, “Der Acoustic Voice Quality Index in Deutsch,” *HNO*, vol. 60, no. 8, pp. 715–720, Aug. 2012. DOI: [10.1007/s00106-012-2499-9](https://doi.org/10.1007/s00106-012-2499-9).
- [7] F. Pausch, “Documentation of the experimental environments and hardware used in the dissertation ”Spatial audio reproduction for hearing aid research: System design, evaluation and application”,” Lehrstuhl für Hörtechnik und Akustik, Tech. Rep. RWTH-2022-01536, 2022. DOI: [10.18154/RWTH-2022-01536](https://doi.org/10.18154/RWTH-2022-01536).
- [8] A. Team, *Audacity(R): Free Audio Editor and Recorder [Computer application]*, 2021.
- [9] *Loudness Normalisation and Permitted Maximum Level of Audio Signals (EBU R128-2020)*. Geneva: European Broadcasting Union, Aug. 2020.
- [10] T. Baltrušaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, “OpenFace 2.0: Facial Behavior Analysis Toolkit,” in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2018.