AVbook, a high-frame-rate corpus of narrative audiovisual speech for investigating multimodal speech perception

Enrico Varano[1], Tobias Reichenbach[1,2,*]

1 Department of Bioengineering and Centre for Neurotechnology, Imperial College London, London, United Kingdom.
2 Department of Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-University Erlangen-Nuremberg, Erlangen, Germany. tobias.j.reichenbach@fau.de
** Contact author: tobias.j.reichenbach@fau.de*

Abstract

We present an audiovisual speech corpus that is designed for cognitive neuroscience studies and that can also be employed for research on audiovisual speech recognition. The corpus consists of 3.6 hours of audiovisual recordings of two speakers, one male and one female, reading passages from a narrative English text. The visual recordings were acquired at a high frame rate of 119.88 frames per second (fps) and exported at a high resolution of 528 x 718 pixels. The speech is pronounced with a neutral British accent and is directed at the camera. Both speakers read the same 59 passages of a book, for a total of 1h50' each. The passage scripts, largely contiguous within a non-fiction source book chosen for its compelling content, were selected and lightly edited to keep subjects who might listen to it interested and alert. As tools to test comprehension and attention, sets of four multiple-choice questions were written for each passage. A short written summary is also provided for each recording. To enable audiovisual synchronisation when presenting the stimuli, four videos of an electronic clapperboard were recorded in line with the corpus. Stimulus synchronisation of 0 ± 4 ms was achieved by pairing these with a high frame rate commercial monitor and a photo-sensor. The audiovisual speech material, the corresponding text, synchronization material, comprehension questions and written summaries set are available on the web for research use.

Introduction

Visual cues play a prominent role in natural communication and have been shown to improve speech comprehension with and without background noise (Reisberg et al., 1987, Ross et al., 2007). This audiovisual benefit occurs across the scale of linguistic units, from syllables (Bernstein et al., 2004) to words (Sumby and Pollack, 1954) and sentences (Grant and Seitz, 2000), and extends to automatic speech recognisers (Matthews et al., 2002).

Databases of audiovisual (AV) speech material play an important role in enabling studies on the neural mechanisms that underlie audiovisual integration regarding speech perception. Studies on this issue traditionally employed brain imaging techniques to investigate participants' responses to short speech tokens such as phonemes and syllables. Neural responses to such short stimuli were commonly analysed by averaging hundreds of response trials time-aligned to the onset of their repetitive stimuli to obtain event-related potentials (ERPs; Luck, 2004). The need

for stimuli suitable for such paradigms could be met by the researchers recording the material themselves (McGurk and MacDonald, 1976; Brown et al., 2018; Sumby and Pollack, 1954) or by employing one of many corpora published to support automatic speech recognition efforts.

The first audiovisual speech datasets, including TULIPS1 (Movellan, 1995) and AVletters (Matthews et al., 1998), accordingly consisted of few speakers reading digits and letters. Later datasets, for instance XM2VTS (Messer et al., 1999), AVICAR (Lee et al., 2004), VidTIMIT (Sanderson and Lovell, 2009), GRID (Cook et al., 2006), BL (Benezeth and Bachman, 2011) or MODALITY (Czyzewskiet al., 2017), provided complete sentences, a larger volume of recordings, and included several additional features such as high frame rates, head poses or lip highlighting. These databases were primarily designed to support the development of algorithms for audiovisual speech processing and recognition, speaker identification and detection, affective state recognition or talking head generation.

These audiovisual speech corpora have been used extensively to study neural mechanisms of audiovisual speech processing. However, the continuous and complex nature of natural speech is not entirely reflected in the short stimuli and limited lexicon of these speech materials (Sonkusare et al., 2019). Recent advances in analysis methodologies and computational power have allowed researchers to investigate neural responses to increasingly complex stimuli including ongoing natural speech (Marmarelis, 2004; Ding and Simon, 2014; Ding and Simon, 2012; Crosse et al., 2016). These studies have typically employed audiobooks as a practical solution to present ecologically valid speech to participants.

Continuous audiovisual speech corpora are, however, less common due to the resources required to assemble and process the material (Chitu and Rothkrantz, 2012). The most commonly employed option to date is a series of weekly addresses made by a well-known male talker speaking on contemporary social, political and economic issues (Supasorn et al., 2017; O'Sullivan et al., 2017). While several hours of audiovisual speech are available, the framing is not consistent and the content, tone and familiarity of the speaker could impact the obtained results in unintended ways.

Furthermore, when presenting multimodal material to subjects in an EEG or MEG study, the precise synchronisation of the different sensory streams must be tightly controlled to take advantage of the high temporal resolution of the recordings (Schultz et al., 2020; Crosse et al., 2014). The low frame rate of 30 frames per second (fps) of the majority of corpora, including the aforementioned male speaker corpus but with the notable exception of the short sentence MODALITY corpus, limits the precision in audiovisual alignment to 33 ms.

The AVbook corpus that we present here is a narrative audiovisual speech corpus that significantly extends the limited current pool of continuous speech corpora while providing additional features aimed at facilitating speech and neuroscience research. The material consists of 3.6 hours of high frame rate and high resolution recordings of two trained speakers, one male and one female, reading the same, curated narrative text. The framing is consistent due to the employment of a teleprompter and the speakers voice is recorded with a professional clip-on microphone, yielding high quality recordings that are also suitable for audio-only experiments. This may, in turn, enable closer comparison between uni-modal and multi-modal speech processing.

Corpus

Content Design

With the experimental participant's enjoyment and attention in mind, Alfred Lansing's book "*Endurance: Shackleton's Incredible Voyage to the Antarctic*" was selected as the source for the corpus content for its compelling narrative. The first eight chapters of the book, constituting

Part 1, were divided into 59 passages with an average word count of 335. The original text was lightly edited to partition it into contextually complete passages and to remove outdated vocabulary and direct speech. The latter action was taken to aid the speakers in producing a neutral tone without overly salient parts.

Collection

The audiovisual recordings were obtained in a studio at Imperial College London, UK, using a Sony A7S II camera (Sony Corporation, Japan) mounted in a teleprompter. The audio material was recorded concurrently through the camera's auxiliary sound port connected to a Sennheiser EW100 clip-on radio microphone (Sennheiser, Germany). The speakers wore the microphone on their clothing and sat on a chair at a distance of three meters from the teleprompter and the camera. The camera was oriented in landscape and framed the speaker's head, shoulders, and chest against a grey background.

The camera was set to record at the frame rate of 119.88 fps and with a resolution of 1280 x 720 pixels. The single channel audio was recorded at 48 kHz with 32-bit resolution.

The speakers, both professional actors, one female and one male, were recruited through the website StarNow.com and selected based on a video audition for their neutral accent and clear speech. They were directed to keep their head still facing the teleprompter, and to speak with a neutral tone of voice. The scroll speed of the teleprompter was adjusted to suit the speaker's preferred pace. Mistakes during production were re-recorded following a pause to be later cut in post-processing. Taking the post-production process of removing pauses and mistakes into account, this resulted in an average passage length of 111 and 104 seconds, and an average speech rate of 182 and 192 words per minute, for the female and male speakers respectively.

Post-processing

Editing of the footage was performed on Adobe Premiere CC 2019 (Adobe Inc., USA) with the aim of cutting speech production and pronunciation mistakes, and unnatural hesitations. Pauses over 0.8 seconds were cut down to the same threshold. Video jumps and audio clicks due to cuts were smoothed out with filters and transitions, and care was taken to minimise the overlap of these transitions with speech production. The framing was cropped around the speaker's face and neck to a vertical orientation with a resolution of 528 x 718 pixels. Five frames from each speaker are shown in Figure 1.

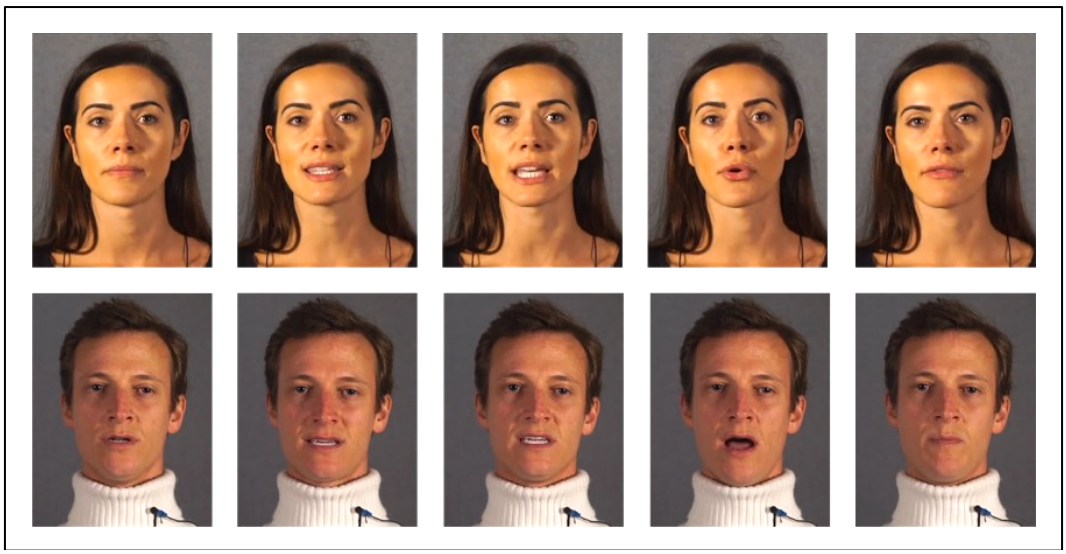

Figure 1 — Example frames from the AVbook corpus. The upper row shows five frames taken from the female speaker, and the bottom row five frames from the male speaker.

The resulting 59 videos were exported into mp4 container files using the h.265 video codec and the AAC audio codec at a frame rate of approximately 119.88 fps and 48 kHz with 16-bit resolution, respectively. The precise frame rate of the video recording was 120/1.001 fps, with each frame lasting about 8.34 ms.

Script, comprehension questions and summaries

The precise wording pronounced by each speaker was checked against the script by hand. A few inconsistencies and mispronunciations that could not be corrected during the editing process were annotated in the teleprompter script file.

A short summary text was also produced for each video segment. This text gave an overview of the content of the respective passage.

Furthermore, a set of multiple-choice questions and answers was written for each passage. Four questions, each with one correct and two incorrect options, were carefully drafted to avoid revealing the answers to other questions while minimising the number of correct answers identifiable by general knowledge, context or information contained in previous passages.

Synchronisation

Synchronisation material

To enable the precise synchronisation of the audio and video streams when presenting the audiovisual material to subjects, four videos of an electronic clapperboard were recorded in the same conditions and with the same equipment as the audiovisual speech material. These 60 second synchronisation videos were post-processed and exported in the same manner as the audiovisual speech material.

The electronic clapperboard, depicted schematically in Figure 2, consisted of a standard red LED and a passive breadboard buzzer, both powered in parallel by a GW Instek GFG-8219A signal generator producing a 2.5 kHz sine wave. A mechanical switch connected in series with the signal generator was operated irregularly to generate a non-periodic on/off signal to power the diode and buzzer intermittently. The buzzer was found to have a response time of 0.3 ms and LEDs are known to have response times of the order of single-digit nanoseconds. The image of the LED and the recording of the buzzer sound contained in the synchronisation videos can therefore serve to align the audio and the video signals of the AVbook corpus.

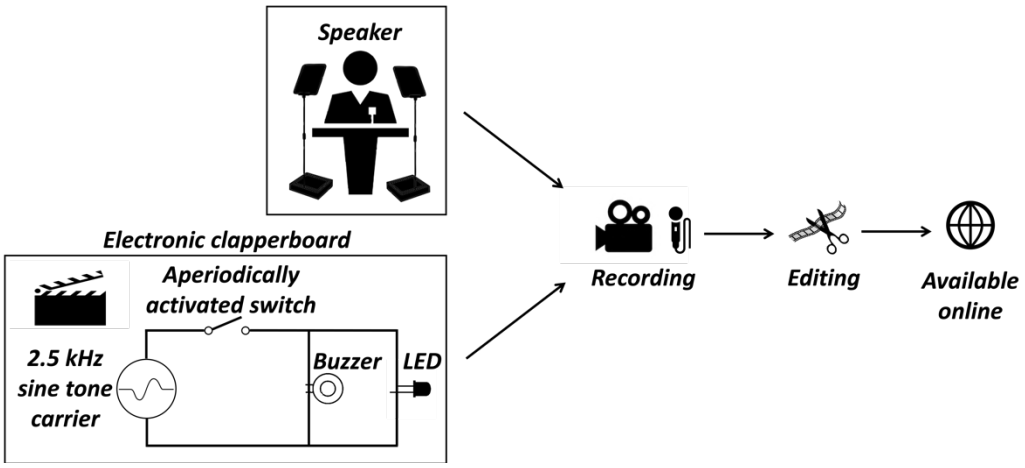

Figure 2 — Schematic depiction of the methodology for producing the audiovisual speech corpus (top) as well as the synchronisation material obtained through an electronic clapperboard (bottom).

Audiovisual alignment during presentation of the AVbook

When employing the AVbook material in behavioural or neuroscientific studies, the audiovisual synchronisation of the stimulus is essential. A diagram of the proposed stimuli presentation and alignment solution that we tested is shown in Figure 3. A custom-written stimulus presentation software module was employed to command the playback of the synchronisation files. The brightness of the image of the red LED was recorded simultaneously to the presented audio stream containing the buzzer sound and a cross-correlation was performed to determine the latency between the visual and the audio stimulus. A corresponding opposite latency shift can then be applied in the presentation software so that the resulting visual and audio stimuli are aligned.

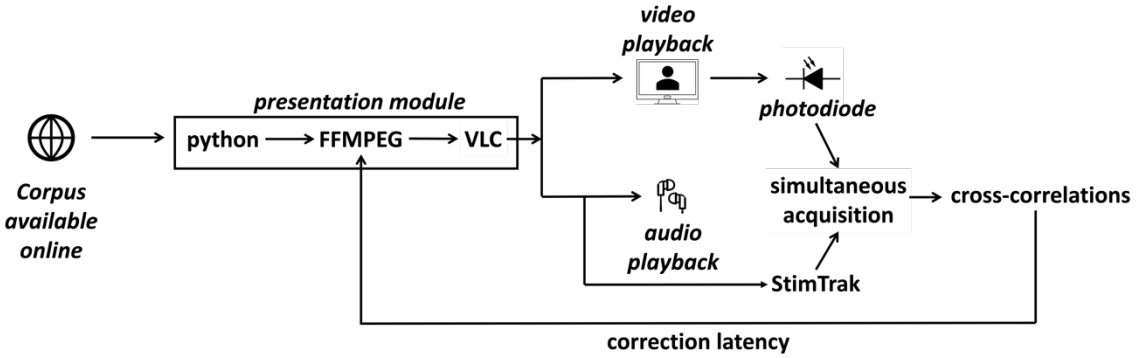

Figure 3 — Schematic depiction of the equipment and pipeline for presenting the audiovisual stimuli and for measuring and the latency between the visual and the audio signal.

Presentation hardware and module

In a test of the presentation setup, the audio stimulus was delivered diotically at a level of 70 dB(A) SPL using ER-3C insert earphones (Etymotic, USA) through a high-performance sound card (Xonar Essence STX, Asus, USA). The sound level was calibrated with a Type 4157 ear simulator (Brüel & Kjær, Denmark). The video component was delivered via a 144 Hz, 24-inch flat-screen monitor (24GM79G, LG, South Korea) set at a refresh rate of 119.88 Hz.

During an initial piloting phase, a Python 3.7 script calling the *python-vlc* library[1] was found to be the most reliable way to decode and present videos in the h.265 video codec synchronously to the audio material. Any residual timing difference between the audio and the video stimulus can then be corrected using the *-itoffset* FFMPEG flag[2] called through the *subprocess* python module.

Recording hardware

An actiCHamp electrophysiological recording amplifier (BrainProducts, Germany) was used to record the image of the red LED and the sound of the buzzer in the synchronisation videos. A photodiode (Photo Sensor, BrainProducts, Germany) and an acoustic adaptor (StimTrak, BrainProducts, Germany) were plugged into the auxiliary ports of the integrated amplifier. The combined data stream was then acquired through PyCorder (BrainProducts, Germany) at a sampling rate of 10 kHz, therefore allowing for temporal alignment of the audiovisual signal with a precision of 0.1 ms.

Results

A sliding window cross-correlation analysis (Figure 4) of the photodiode and acoustic adaptor signals for the synchronisation videos was performed to determine the delay of the audio signal

[1] Available at https://pypi.org/project/python-vlc/
[2] Available at https://ffmpeg.org

with respect to the video stimulus over the runtime of the video. It was found that the delay averaged 41.7 ms, was stable within 0.8 ms across time and jittered by a maximum of 8.1 ms between different trials. Correcting for the mean delay in the presentation software resulted in a mean delay of 0.6 ms.

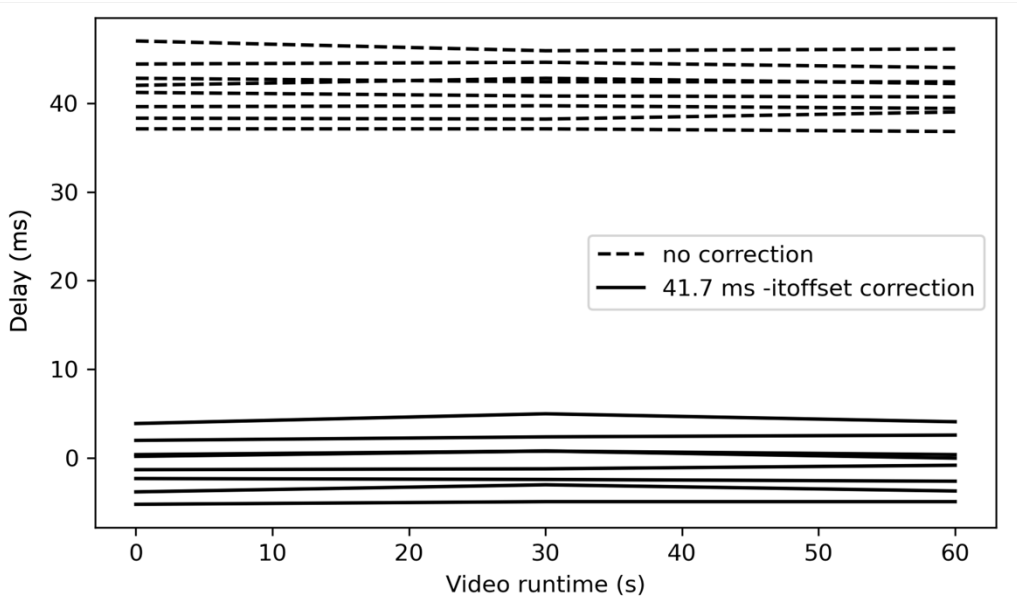

Figure 4 — Delay of the audio signal with respect the visual stimulus over the duration of a synchronisation video, for several trials. Before correcting for the delay in the presentation software (dashed), the average delay is 41.7 ms. After correction (solid) we obtain a negligible delay of 0.6 ms.

Comprehension Tests

Participants

To validate our multiple-choice comprehension questions, we conducted a behavioural experiment with seventeen native English speakers, ten of them female, with self-reported normal or corrected-to-normal vision and healthy hearing. The participants were between 18 and 29 years of age, with a mean age of 23 years. All participants were right-handed and had no history of mental health problems, severe head injury or neurological disorders. Before starting the experiment, participants gave informed consent. The experimental protocol was approved by the Imperial College Research Ethics Committee.

Stimuli Presentation

The experiment took place in an acoustically and electrically insulated room (IAC Acoustics, United Kingdom). The same equipment and settings employed in the audiovisual synchronisation experiment were used to control the audiovisual presentation and data acquisition.

We considered three types of stimuli. The first was audio-only speech in stationary, speech-shaped background noise. The second condition was audiovisual speech, also in background noise. In both conditions the speech was presented in a constant level of background noise, at a signal to noise ratio of -2 dB. The third condition was visual-only speech, with no sound presented. Subjects listened to 54 passages from the corpus and the three conditions were randomised between the different passages. Between each passage, the participants were tasked with answering the AVbooks comprehension questions and then to read the summary before

proceeding to the next task. A speech comprehension score was computed as the average percentage of correct answers for each of the three conditions.

Results

The comprehension scores (Figure 5) in the audio-only condition were found to be 55.3% ± 3.5% (mean and standard error of the mean) while they were 60.6% ± 2.8% in the audiovisual condition. The difference between the scores in the two conditions was not significant ($p = 0.18$, $t = 1.40$, paired Student's t-test with Benjamini–Hochberg FDR correction). On the other hand, subjects scored significantly lower in the video only condition (34.1% ± 2.9%) than when presented with the audio signal only or with the audiovisual stimuli ($p = 1.1 \times 10^{-5}$, $t = 6.51$ and $p = 3.5 \times 10^{-6}$, $t = 7.56$ respectively, FDR corrected paired Student's t-tests). Participants did not score better than the chance level (33%) in the video only condition ($p = 0.72$, $t = 0.37$, one sample Student's t-test).

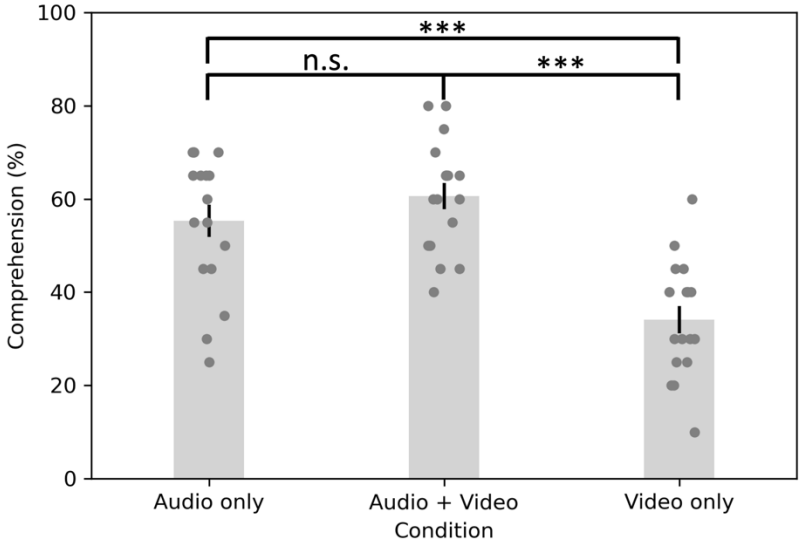

Figure 5 — Speech comprehension scores for the AVbook in background noise without a visual signal (audio only), for the audiovisual stimulus, as well as for the video only. Error bars represent the standard error of the mean, and the dark grey points show the average score per participant.

Conclusion

A narrated, continuous speech audiovisual corpus, AVbook, was collected to support the investigation of audiovisual speech perception through brain imaging and behavioural studies. Although the corpus is limited to two speakers, the high frame rate may make it an attractive option for developing and benchmarking computer algorithms of audiovisual speech processing as well.

The corpus passages were chosen to yield high attention and engagement of experimental subjects. In case the passages were not fully comprehensible to participants, for instance due to the inclusion of background noise, a short written summary was produced for each video segment. Providing this summary text before presenting the next passage will help to keep a subject's engagement at a high level.

We developed comprehension questions that accompany the AVbook corpus, and tested these in a behavioural experiment. Our finding that participants scored well above chance when they could hear the speech material, but that their performance was at chance level when they could

only lip-read suggests that the question set is fit for purpose. On the other hand, our behavioural experiment revealed no significant difference between the audio-only and the audiovisual conditions. This presumably reflected that the questions were designed to test participants' understanding of the stories and attention to it, and not for a highly accurate quantification of comprehension. The latter would benefit from single semantically unpredictable sentences that are assessed for the comprehension of key words.

We also developed a solution, an electronic clapperboard, to temporally align the audio and video stimuli precisely and reliably. Our synchronisation experiment, aided by the videos of the electronic clapperboard, yielded a negligible mean latency that remained consistent within trials and jittered by less than one frame (8.1 ms vs 8.3 ms) across trials.

Due to the precise audiovisual synchronization, the AVbook corpus can allow the neuroscience community to extend work done in the audio-only modality with traditional audio-book corpora to include the visual domain. Such studies could tackle the effect of audiovisual speech on neural responses related to attention (O'Sullivan et al., 2015) or linguistic factors such as surprisal (Weissbart et al., 2020). Furthermore, a consistent use of audiovisual speech material across geographies, equipment and experimental paradigms may help compare and reproduce results.

Further, due to its consistent framing and high number of frames, the AVbook corpus may be an attractive option for computational studies and therefore support the use of common material in language, speech perception and automatic speech recognition work, a result largely achieved by the GRID and VIDTIMIT corpora in the context of short sentences.

The complete corpus, synchronization material, teleprompter scripts, comprehension questions, summary texts, and raw video recordings are available to download for research use on zenodo.org.

Acknowledgements

This research was supported by the Royal British Legion Centre for Blast Injury Studies, EPSRC grants EP/M026728/1 and EP/R032602/1, as well as by the U.S. Army through project 71931-LS-INT.

References

Benezeth, Y., & Bachman, G. (2011). BL-Database: A French audiovisual database for speech driven lip animation systems. http://hal.inria.fr/inria-00614761/.

Bernstein, L. E., Auer, E. T. Jr., and Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech and Communication*. 44, 5–18. doi: 10.1016/j.specom.2004.10.011

Brown, V.A., Hedayati, M., Zanger, A., Mayn, S., Ray, L., et al. (2018) What accounts for individual differences in susceptibility to the McGurk effect?. *PLOS ONE* 13(11): e0207160. doi: 10.1371/journal.pone.0207160

Chitu, A.G., and Rothkrantz, L.J.M. (2007). Building a data corpus for audio-visual speech recognition. In L. J. M. Rothkrantz, & C. A. P. G. van der Mast (Eds.), Euromedia 2007 (pp. 88-92). Eurosis-ETI.

Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120 (5 Pt 1), 2421–2424. doi: 10.1121/1.2229005.

Crosse, M. J., & Lalor, E. C. (2014). Retraction. *Journal of neurophysiology*, 112(10), 2667. doi: 10.1152/jn.z9k-2710-retr.2014

Crosse, M.J., Di Liberto, G.M., Bednar, A., Lalor, E.C. (2016). The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. *Frontiers in Human Neuroscience.* 2016;10:604-604. doi: 10.3389/fnhum.2016.00604

Grant, K. W., and Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, 108, 1197–1208. doi: 10.1121/1.1288668

Matthews, I., Cootes, T.F., Bangham, J.A., Cox, S., Harvey, R. (2002). "Extraction of visual features for lipreading," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 198-213, doi: 10.1109/34.982900.

Lee, B., Hasegawa-johnson, M., Goudeseune, C., Kamdar, S., Borys, S., Liu, M., & Huang, T. (2004). AVICAR : Audio-Visual Speech Corpus in a Car Environment. *8th International Conference on Spoken Language Processing* pp. 8–11.

Luck, S. J. (2014). An Introduction to the Event-Related Potential Technique. Cambridge, MA: MIT press.

Marmarelis, V. Z. (2004). Nonlinear Dynamic Modeling of Physiological Systems. Hoboken, NJ: John Wiley & Sons.

McGurk, H., MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748. doi: 10.1038/264746a0

Messer, K., Matas, J., Kittler, J., & Jonsson, K. (1999). XM2VTSDB: The extended M2VTS database. In *Second International Conference on Audio and Video-based Biometric Person Authentication* (pp. 72–77).

Movellan, J. R. (1995). "TULIPS1 database" in *Visual Speech Recognition with Stochastic Networks* in *Advances in Neural Information Processing Systems*, Cambridge:MIT Pess, vol. 7.

Ding, N., Simon, J.Z., (2014). Cortical entrainment to continuous speech: Functional roles and interpretations. *Frontiers in Human Neuroscience*. 8, 311

Ding, N., Simon, J.Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences of the United States of America* 109, 11854–11859.

O'Sullivan, J.A., Power, A.J., Mesgarani, N., Rajaram, S., Foxe, J.J., Shinn-Cunningham, B.G., Slaney, M., Shamma, S.A., Lalor, E.C. (2015). Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG, *Cerebral Cortex*, Volume 25, Issue 7, July 2015, Pages 1697–1706, doi: 10.1093/cercor/bht355

O'Sullivan, A., Crosse, M.J., Di Liberto, G.M., Lalor, E.C. (2017). Visual Cortical Entrainment to Motion and Categorical Speech Features during Silent Lipreading. *Frontiers in Human Neuroscience.* doi: 10.3389/fnhum.2016.00679

Reisberg, D., Mclean, J., and Goldfield, A. (1987). "Easy to hear but hard to understand: a lip-reading advantage with intact auditory stimuli," in The Psychology of Lip-Reading, eds B. Dodd and R. Campbell (Hillsdale: Lawrence Erlbaum Associates), 97–114.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., and Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex* 17, 1147–1153. doi: 10.1093/cercor/bhl024

Sanderson, C., Lovell, B.C. (2009). Multi-Region Probabilistic Histograms for Robust and Scalable Identity Inference. In: *Tistarelli, M., Nixon, M.S. (eds) Advances in Biometrics. ICB 2009. Lecture Notes in Computer Science*, vol 5558. Springer, Berlin, Heidelberg. doi: 10.1007/978-3-642-01793-3_21.

Schultz B.G., Biau E., Kotz S.A. (2020). An open-source toolbox for measuring dynamic video framerates and synchronizing video stimuli with neural and behavioral responses. *Journal of Neuroscience Methods*, 343 (2020), Article 108830

Sonkusare, S., Breakspear, M., & Guo, C. (2019). Naturalistic Stimuli in Neuroscience: Critically Acclaimed. *Trends in cognitive sciences*, 23(8), 699–714. https://doi.org/10.1016/j.tics.2019.05.004

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, 26, 212–215. doi: 10.1121/1.1907309

Supasorn, S., Seitz, S.M., and Kemelmacher-Shlizerman, I. (2017). Synthesizing Obama: learning lip sync from audio. *Association for Computing Machinery Transactions on Graphics*. 36, 4, Article 95, 13 pages. doi: 10.1145/3072959.3073640

Weissbart, H., Kandylaki, K.D., Reichenbach; T. (2020). Cortical Tracking of Surprisal during Continuous Speech Comprehension. *Journal of Cognitive Neuroscience* 2020; 32 (1): 155–166. doi: 10.1162/jocn_a_01467