JASA ARTICLE

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

Enrico Varano,[1] Pierre Guilleminot,[1] and Tobias Reichenbach[2,a)]
[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-Nürnberg, Erlangen, Germany*

ABSTRACT:
Seeing a speaker's face can help substantially with understanding their speech, particularly in challenging listening conditions. Research into the neurobiological mechanisms behind audiovisual integration has recently begun to employ continuous natural speech. However, these efforts are impeded by a lack of high-quality audiovisual recordings of a speaker narrating a longer text. Here, we seek to close this gap by developing AVbook, an audiovisual speech corpus designed for cognitive neuroscience studies and audiovisual speech recognition. The corpus consists of 3.6 h of audiovisual recordings of two speakers, one male and one female, each reading 59 passages from a narrative English text. The recordings were acquired at a high frame rate of 119.88 frames/s. The corpus includes phone-level alignment files and a set of multiple-choice questions to test attention to the different passages. We verified the efficacy of these questions in a pilot study. A short written summary is also provided for each recording. To enable audiovisual synchronization when presenting the stimuli, four videos of an electronic clapperboard were recorded with the corpus. The corpus is publicly available to support research into the neurobiology of audiovisual speech processing as well as the development of computer algorithms for audiovisual speech recognition.
 https://doi.org/10.1121/10.0019460
(Received 23 December 2022; revised 17 April 2023; accepted 1 May 2023; published online 30 May 2023)
[Editor: Bastian Epp] Pages: 3130–3137

I. INTRODUCTION

Visual cues play a prominent role in natural communication. Seeing a talker's face can have a large benefit on speech comprehension, in particular in the presence of 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). Automatic speech recognizers can similarly benefit from the additional information contained in a speaker's facial movements (Matthews *et al.*, 2002).

Studies on the neural mechanisms that yield audiovisual integration 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 analyzed by averaging hundreds of response trials time-aligned to the onset of their repetitive stimuli to obtain event-related potentials (ERPs) (Luck, 2014). The requirement for suitable stimuli in such paradigms can be fulfilled by researchers recording the material themselves (Brown *et al.*, 2018; Hauswald *et al.*, 2018; McGurk and MacDonald, 1976; Mégevand *et al.*, 2020; Sumby and Pollack, 1954) or by employing one of many corpora published to support the development of automatic speech recognition.

The first audiovisual speech datasets, including TULIPS1 (Movellan, 1995) and AVletters (Matthews *et al.*, 2002), accordingly consisted of few speakers reading digits and letters. Later datasets, for instance XM2VTS (Messer *et al.*, 1999), AVICAR (Lee *et al.*, 2004), GRID (Cooke *et al.*, 2006), VidTIMIT (Sanderson and Lovell, 2009), BL (Benezeth *et al.*, 2011), TCD-TIMIT (Harte and Gillen, 2015), MODALITY (Czyzewski *et al.*, 2017), LRS (Chung *et al.*, 2017), or RoomReader (Reverdy *et al.*, 2022) provided complete sentences, a larger volume of recordings, and included several additional features such as high frame rates, different camera angles, lip highlighting, spontaneous conversation, multimodal cues of conversational engagement and behavioral aspects of collaborative interaction. These databases were primarily designed to support the development of algorithms for audiovisual speech processing and recognition, speaker identification and detection, affective state recognition, and talking head generation.

These audiovisual speech corpora have been used extensively to study neural mechanisms of audiovisual speech processing as well. 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 (Crosse *et al.*, 2016; Ding and Simon, 2012, 2014; Giraud and

a)Electronic mail: tobias.j.reichenbach@fau.de

 0001-4966/2023/153(5)/3130/8/$30.00

https://doi.org/10.1121/10.0019460

Poeppel, 2012; Marmarelis, 2004; Moses *et al.*, 2016). These studies have typically employed audiobooks as a practical solution to present ecologically valid speech to participants.

Audiovisual corpora of continuous speech are less common, however, due to the resources required to assemble and process the material (Chitu and Rothkrantz, 2007). 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 (O'Sullivan *et al.*, 2017; Suwajanakorn *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. Other works recommend employing established corpora of short sentences such as TCD-TIMIT (Crosse *et al.*, 2021).

Furthermore, when presenting multimodal material to subjects in an electroencephalography (EEG) or magnetoencephalography study, the precise synchronization of the different sensory streams must be tightly controlled to take advantage of the high temporal resolution of the recordings (Crosse and Lalor, 2014; Schultz *et al.*, 2020). While many electrophysiological studies investigate low-frequency neural oscillations, the precise latency of responses is important to establish the order of hierarchical neural processes and to improve the signal-to-noise ratio by reducing jitter across and within trials. The low frame rate of 30 frames/s (fps) of the majority of corpora, including the aforementioned corpus of speeches by a male politician, but with the notable exception of the short sentence MODALITY corpus, limits the precision in audiovisual alignment to 33 ms. In addition, high-rate videos may also be useful for extracting features from the video for feature and stimuli reconstruction.

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 language and neuroscience research. The material consists of 3.6 h 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 multimodal speech processing.

II. THE CORPUS

The book, *"Endurance: Shackleton's Incredible Voyage to the Antarctic"* by Lansing (1959), was selected as the source for the corpus content for its compelling narrative, with the experimental participants' enjoyment and attention in mind. 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.

A. Collection

The audiovisual recordings were obtained in a quiet studio at Imperial College London, UK. The studio was enclosed in solid walls but was not further sound-proofed. Recordings were obtained 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 3 m from the teleprompter and the camera. The camera was oriented in landscape and framed the speaker's head, shoulders, and chest against a gray background.

The camera was set to record at the frame rate of 119.88 fps (precisely, 120/1.001) and with a resolution of 1280×720 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 online and selected based on a video audition for their neutral British accent and clear speech. They were directed to keep their heads still, face the teleprompter, and speak with a neutral tone of voice. The scroll speed of the teleprompter was adjusted to suit the speakers' preferred pace. If mistakes in speaking occurred during the production, the corresponding segment was re-recorded following a pause that was later cut in post-processing. This resulted in an average passage length of 111 and 104 s, with an average speech rate of 182 and 192 words/min, for the female and male speakers, respectively.

B. Post-processing

The recordings were edited on Adobe Premiere CC 2019 (Adobe Inc., San Jose, CA). Inhalation sounds, hesitations, and speech production or pronunciation mistakes were corrected by cutting the audio during pauses in the speech and by introducing re-recorded material. Video jumps and audio clicks due to cuts were smoothed out with filters and transitions, and care was taken to minimize the overlap of these transitions with speech production. Since such edits may influence neural responses depending on experimental design and analysis technique, the raw recordings are also made available with the corpus. Researchers wanting full control of the stimuli may thus perform their own editing. The framing was cropped around the speaker's face and neck to a vertical orientation with a resolution of 528×718 pixels. Five exemplary frames from each speaker are shown in Fig. 1.

The resulting 59 videos were exported into mp4 container files using the h.265 video codec and the advanced audio coding (AAC) audio codec at the recording frame rate of 120/1.001 fps and 48 kHz with 16 bit resolution, respectively. No

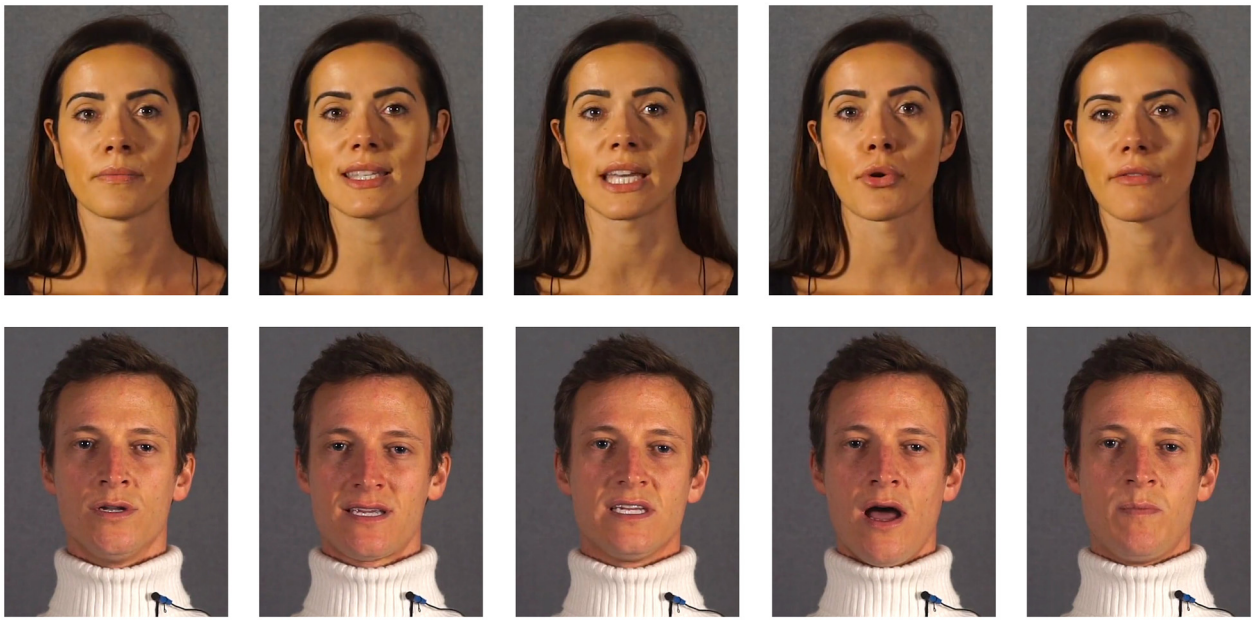

FIG. 1. (Color online) Example frames from the AVbook corpus. The upper row shows five frames taken from the female speaker and the bottom row shows five frames taken from the male speaker.

clipping was observed in the audio and no further processing was performed.

The audio was also separately exported as a WAVE file. An *a priory* signal-to-noise ratio (SNR) was estimated using a voice activity detection (VAD) script (Brookes, 2022) following the ITU-T P.56 standard (ITU-T, 2011). The average SNR was found to be 34.0 ± 2.9 dB and 44.5 ± 7.1 dB for the female and male speakers, respectively (mean and standard deviation). Similar results were obtained when employing a custom-tuned VAD script or by calculating the SNR by comparing silent video segments which were cut out of the final corpus (less than 1 dB difference in both cases for both speakers).

To investigate the spectral properties of the voices, we computed the average frequency content of the speech signals for both speakers. Figure 2 depicts the magnitude of the normalized frequency content on a logarithmic frequency scale. This analysis shows differences in the spectral characteristics between the female and male speakers.

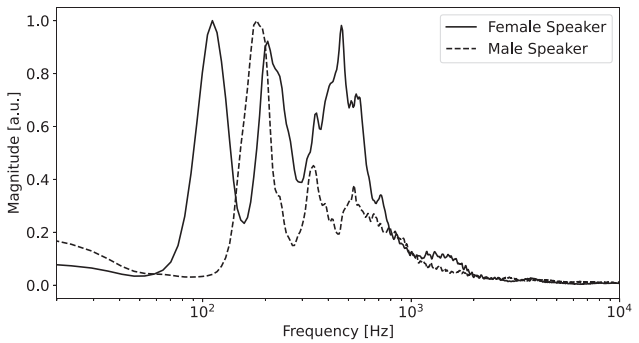

FIG. 2. Normalized frequency content of the speech signals for the female and male speakers, plotted on a logarithmic frequency scale.

In particular, it highlights the different fundamental frequencies of the two voices (about 111 Hz on average for the male voice and about 176 Hz on average for the female voice).

C. 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, four multiple-choice questions were written for each passage, each with one correct and two incorrect options. The questions were carefully drafted to avoid revealing the answers to those about later passages while minimizing the number of correct answers identifiable by general knowledge, context, or information contained in previous passages.

D. Words, phonemes, and timing

In addition to the original script for each passage, phonetic information was extracted from the corpus using the Montreal forced aligner (McAuliffe *et al.*, 2017). The output consists of TextGrid files containing the word and phone level segmentation for each passage and speaker, and is provided with the corpus. The alignment between these files and the corresponding audio files were then manually checked using Praat (Styler, 2013).

The analysis of word and phonetic features has been useful in linguistic and neuroscience studies, allowing, for

https://doi.org/10.1121/10.0019460

example, to obtain phonetic and word-level information from EEG responses to naturalistic speech (O'Sullivan *et al.*, 2021; Weissbart *et al.*, 2020). In addition, such information can also be used to design stimuli based on phonetic markers rather than acoustic features (Guilleminot and Reichenbach, 2022).

III. SYNCHRONIZATION

When employing the AVbook material in behavioral or neuroscientific studies, the synchronization of the audio and the visual signal is essential. To enable precise synchronization, four videos of an electronic clapperboard were recorded under the same conditions and with the same equipment as the audiovisual speech material. These 60 s synchronization videos were post-processed and exported in the same manner as the audiovisual speech material.

A. Synchronization material

The electronic clapperboard, depicted schematically in Fig. 3(A), consisted of a red light-emitting diode (LED) and a passive breadboard buzzer, both powered in parallel by a GW Instek GFG-8219A (Instek, Taiwan) 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 negligible constant response time of 0.3 ms, while LEDs have even shorter response times of a few nanoseconds. The image of the LED and the recording of the buzzer sound contained in the synchronization videos can therefore serve to align the audio and the video signals of the AVbook corpus.

To ensure that the latency correction translates well to the longer passage files, a further check signal was included. This consisted of a small 200 × 100 pixel video containing a square that pseudo-randomly flipped between black and white [Fig. 3(B)]. This video was overlaid onto a corner of the corpus videos and the videos of the electronic clapperboard using the *-filter_complex* command in FFmpeg (https://www.ffmpeg.org/).

B. Audiovisual alignment for presentation of the AVbook

A diagram of the proposed AVbook presentation solution, and of a method for aligning the audio and the visual signals, is shown in Fig. 4. A custom-written stimulus presentation software module was employed to control the playback of the video of the electronic clapperboard. The brightness of the image of the red LED was recorded simultaneously to the presented audio stream containing the buzzer sound. The cross correlation between the envelopes of the visual and the audio signals was then computed to determine the latency between both stimuli. The envelopes were computed from the absolute value of the Hilbert

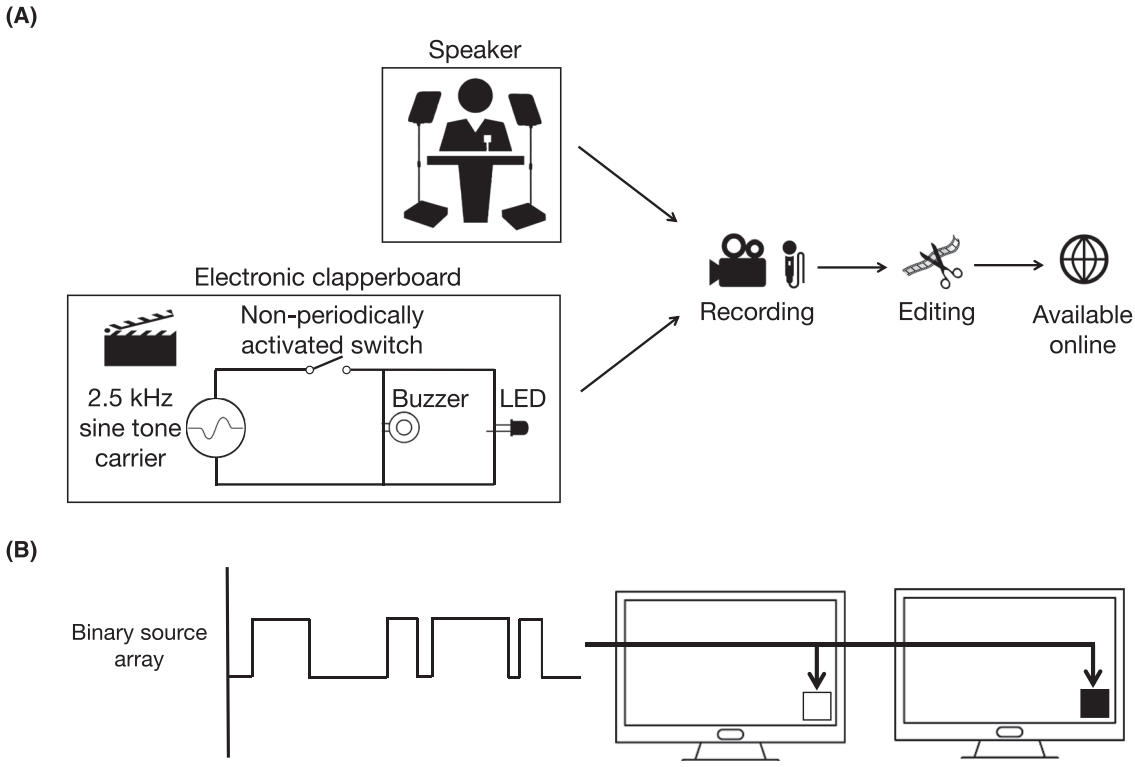

FIG. 3. (A) Schematic depiction of the methodology for producing the audiovisual speech corpus and the synchronization material obtained by filming an electronic clapperboard. The same protocol was employed to record videos of the speaker and electronic clapperboard. The latter consisted of occasional, correlated emissions of sound and light. The videos were then edited and shared on an online database. (B) A second synchronization signal consisted of a small 200 × 100 pixel video. The video contained a square which pseudo-randomly flipped between black and white. This video can be overlaid onto a corner of the corpus videos to enable an additional check of synchronization.

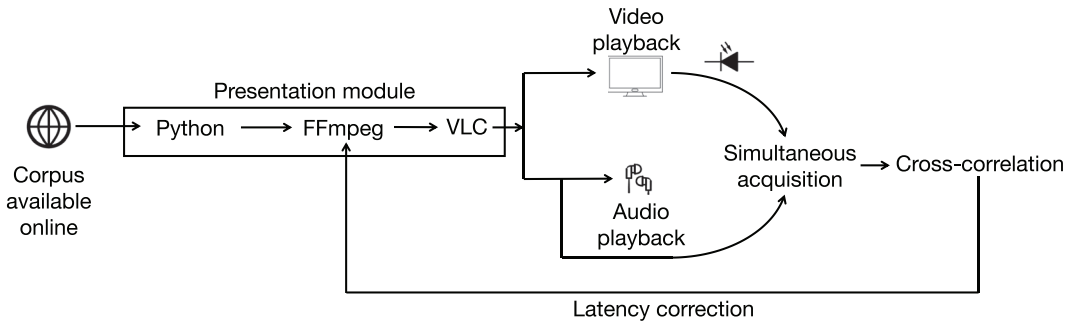

FIG. 4. Schematic depiction of the equipment and pipeline for presenting the audiovisual stimuli and for measuring the latency between the visual and the audio signal. The AVbook corpus can be edited for presentation with a Python routine using the FFmpeg software through the *subprocess* module, which allows for audiovisual manipulations such as latency correction or noise masking. This is then followed by a playback call to the VLC media player through the *python-vlc* bindings module.

transform of the signal and through filtering with a zero-lag forward–backward filter. A corresponding inverted latency shift can then be applied in the presentation software so that the resulting visual and audio stimuli are aligned.

Once the audiovisual latency correction is determined and applied, any residual latency during playback of the narrative AVbook videos can be measured by recording the brightness of the screen area corresponding to the overlaid tracking signal. To this end, one can first determine the onset time of the audio component of the passage (with respect to overall recording start time) by computing a cross correlation between the simultaneous recordings of the presented audio (as obtained with an audio cable splitter plugged into the headphone input) and the original audio waveform (as extracted from the WAVE files). The onset time of the video component of the passage can similarly be determined by cross correlating the recording of the brightness of the overlaid portion of the video with the binary array from which the tracking overlaid video was generated. The difference between the start times of the audio and video streams yields the residual latency.

C. 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 through ER-3C insert earphones (Etymotic, USA) and a high-performance sound card (Xonar Essence STX, Asus, Tapei, Taiwan). The sound level was calibrated with a type 4157 ear simulator (Brüel & Kjaer, Denmark). The video component was presented through a Radeon Pro WX 3100 graphics card (Advanced Micro Devices, Inc., USA) and a 24 in. flat-screen monitor (24GM79G, LG, South Korea) with a maximal refresh rate of 144 Hz set at a refresh rate of 119.88 Hz. The test was conducted on a HP desktop computer running Windows 10 version 21H2 64-bit operating system (Microsoft Corporation, Palo Alto, CA) equipped with a four-core Intel Xeon W-2102 CPU running at 2.90 GHz (Intel Corporation, USA) and 32 GB of random-access memory.

A Python 3.7 script calling the VideoLan Client (VLC) media player (Videolan, https://www.videolan.org/) through the *python-vlc* library (Python-vlc, https://pypi.org/project/python-vlc/) was found to be the most reliable way to decode and present videos in the h.265 video codec synchronously to the audio material. The residual timing difference between the audio and the video stimulus was corrected using the *-itoffset* FFmpeg flag called through the *subprocess* Python module (see Fig. 4).

When piloting the setup, a 30 ms offset of the residual timing difference (but no difference in jitter standard deviation) was found between the Asus Xonar Essence STX sound card and the builtin HP Realtek High Definition Audio sound card of the desktop computer employed.

D. Recording hardware

An integrated electrophysiological recording amplifier (actiCHamp, BrainProducts, Germany) was used to record the image of the red LED and the sound of the buzzer in the synchronization videos. The visual signals were measured with a photodiode (Photo Sensor, BrainProducts, Germany) and the acoustic stimulus through an acoustic adapter (StimTrak, BrainProducts, Germany) which 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 an estimation of the temporal delay between the visual and the audio stimulus with a precision of 0.1 ms.

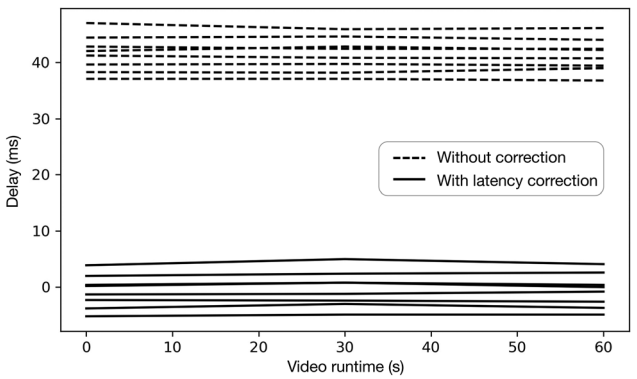

FIG. 5. Delay of the audio signal with respect to the visual stimulus over the duration of a video of the electronic clapperboard, for several trials. Before correcting for the delay in the presentation software (dashed), the average delay was 41.7 ms. After correction (solid), we obtained a negligible mean delay of 0.6 ms.

https://doi.org/10.1121/10.0019460

E. Results

A sliding window cross correlation analysis of the envelopes of the recordings of the photodiode and acoustic adapter signals from the synchronization videos was performed to determine the delay of the audio signal with respect to the video stimulus over the runtime of the video (Fig. 5). The average offset over eight trials (two for each of the four electronic clapperboard recordings) was found to be 41.7 ms. This offset was stable within trials, with a maximum difference of 1.08 ms detected across the 60 s runtime, and jittered by a maximum observed amplitude of 9.19 ms between different trials (or roughly ± 0.5 frames from the mean value). Despite extensive hardware and software optimization, this jitter was not accounted for and is assumed to arise from limitations in the precision of audio and video binding due to the relationship between the jitter and the video frame rate. It is, however, worth noting that the proposed *python-vlc* playback solution was found to be the most stable protocol by a significant margin. Increases in jitter appeared when the video frame rate did not match the frame rate of the monitor (an additional within trials jitter of ± 1.5 frame periods on top of the jitter of the optimal solution, and ± 0.5 frame periods across trials), when the video was exported at 120 rather than 120/1.001 fps and played with the monitor on the 120 fps setting (an additional ±1 frame periods within trials and ±1.5 frame periods across trials), when played through tkvlc (available from PyTkinterVLC, https://pypi.org/project/tkvlc/, an additional ±0.5 frame periods across trials), and when played through the video player integrated within the PsychoPy presentation application (https://www.psychopy.org/, ±15 frame periods within trials, see discussion at https://discourse.psychopy.org/t/moviestim3-issues-regarding-high-fps/6468).

Correcting for the mean delay in the presentation software and repeating the latency measurement with the electronic clapperboard signal resulted in a negligible mean offset of 0.6 ms, with a maximum observed jitter amplitude of 8.87 ms across trials. As expected, the post-correction latency difference measured in the overlaid signal was found to be 40.4 ms, a value within residual measurement error of the prescribed latency correction of 41.7 ms.

IV. COMPREHENSION TESTS

To validate our multiple-choice comprehension questions, we conducted a behavioral experiment in which subjects were presented with stimuli from the AVbook corpus and subsequently with the corresponding multiple-choice questions.

A. Participants

Seventeen native English speakers, 10 female, with self-reported normal or corrected-to-normal vision and normal hearing volunteered to take part. The participants were between 18 and 29 yrs of age, with a mean age of 23 yrs. 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.

B. Stimuli presentation

The experiment took place in an acoustically and electrically insulated room (IAC Acoustics, United Kingdom). The same equipment and settings as employed in the audiovisual synchronization 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 speech-shaped background noise. In both conditions, the speech was presented at a constant 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. The three conditions were randomized between the different passages. Between each passage, the participants were tasked with answering the corresponding AVbooks comprehension questions and then to read the summary before proceeding to the next passage. The speech comprehension score was computed as the average percentage of correct answers for each of the three conditions.

C. Results

The comprehension score (Fig. 6) was found to be 55.3% ± 3.5% (mean and standard error of the mean) in the audio-only condition, 60.6% ± 2.8% in the audiovisual condition, and 34.1% ± 2.9% in the visual-only condition. Participants scored better than chance level (33.$\overline{3}$ %) in the audio-only and in the audiovisual condition ($p = 1.7 \times 10^{-5}$, $t = 6.0$ and $p = 2.0 \times 10^{-7}$, $t = 8.8$, respectively, one sample Student's t-test with two-stage Benjamini–Hochberg false discovery rate

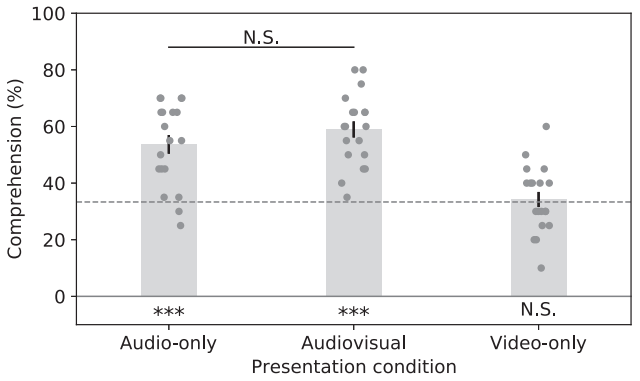

FIG. 6. Speech comprehension scores for the AVbook in background noise without a visual signal (audio only), for the audiovisual stimulus, as well as for lip-reading of the video only. Error bars represent the standard error of the mean, and the dark gray points show the average score per participant. The comprehension scores in the audio-only and the audiovisual condition are significantly above the chance level (dashed line; $***, p < 0.001$) while the average comprehension score in the visual-only condition is very near the chance level.

correction). On the other hand, subjects did not perform above chance level in the visual-only condition ($p = 0.75$, $t = 0.3$).

The difference between the scores in the audio-only and audiovisual conditions was not significant ($p = 0.14$, $t = 1.6$, two-sided, paired Student's t-test).

V. CONCLUSION

An audiovisual corpus, AVbook, of narrated, continuous speech was collected to support the investigation of audiovisual speech perception in neuroimaging and behavioral 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 from experimental subjects. In case the passages were not fully comprehensible to participants, for instance due to the inclusion of background noise or due to presentation of the visual stimulus alone, a short written summary was produced for each video segment. Providing this summary text before presenting the next passage may help keep a subject's engagement at a high level.

We developed comprehension questions that accompany the AVbook corpus, and tested these in a behavioral 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.

While piloting work showed that all questions can be answered correctly in quiet listening conditions or by reading the script of the corresponding passage, some questions may represent a significant memory test and not all subjects could answer them correctly in the presence of background noise. The question set can, however, still differentially evidence attention deficits and audiovisual integration gain across participants, as is reinforced by the observation that participants scoring at chance level in the audio-only condition were equally likely to score significantly above chance or again at chance level in the audiovisual condition.

On the other hand, our behavioral 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 aim may be better pursued using single semantically unpredictable sentences that are assessed for the comprehension of key words, such as in the GRID corpus (Cooke *et al.*, 2006).

We also developed a method, an electronic clapperboard, to temporally align the audio and video stimuli precisely and reliably. Our synchronization experiment, aided by the videos of the electronic clapperboard, yielded a negligible mean latency that remained consistent within trials and jittered by roughly one frame (8.34 ms) across trials.

Due to the precise audiovisual synchronization, the AVbook corpus will allow the neuroscience community to extend work done in the audio-only modality with traditional audiobook 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 English-speaking geographies, equipment, and experimental paradigms may help compare and reproduce results.

Further, due to its consistent framing and high frame rate, 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, similarly to what the GRID and VIDTIMIT corpora achieved in the context of short sentences (Varano *et al.*, 2022).

Last, the AVbook corpus has the potential to guide the design of speech enhancement algorithms in hearing aids: audiovisual speech has recently been shown to enhance auditory attention decoding, a technique used to identify the attended speaker in an auditory selective attention task using analysis of EEG data (Fu *et al.*, 2019).

The complete corpus, phone-level segmentation files, synchronization material, teleprompter scripts, comprehension questions, summary texts, and raw video recordings are available to download for research use (Zenodo, https://zenodo.org/record/7387047).

ACKNOWLEDGMENTS

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

Benezeth, Y., Bachman, G., Le-Jan, G., Souviraà-Labastie, N., and Bimbot, F. (**2011**). "BL-Database: A french audiovisual database for speech driven lip animation systems," Research Report RR-7711.

Bernstein, L. E., Auer, E. T., and Takayanagi, S. (**2004**). "Auditory speech detection in noise enhanced by lipreading," Speech Commun. **44**(1), 5–18.

Brookes, M. (**2022**). "Speech processing toolbox for MATLAB," https://github.com/ImperialCollegeLondon/sap-voicebox.

Brown, V. A., Hedayati, M., Zanger, A., Mayn, S., Ray, L., Dillman-Hasso, N., and Strand, J. F. (**2018**). "What accounts for individual differences in susceptibility to the McGurk effect?," PLoS ONE **13**(11), e0207160.

Chitu, A. G., and Rothkrantz, L. J. M. (**2007**). "Building a data corpus for audio-visual speech recognition," in *Thirteenth Annual Scientific Conference on Web Technology, New Media, Communications and Telematics Theory, Methods, Tools and Applications (Euromedia)*, edited by L. Rothkrantz and C. van der Mast (Eurosis-ETI, Delft – Gent, Belgium), pp. 88–92.

Chung, J. S., Senior, A., Vinyals, O., and Zisserman, A. (**2017**). "Lip reading sentences in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, pp. 3444–3453.

Cooke, M., Barker, J., Cunningham, S., and Shao, X. (**2006**). "The GRID Audio-Visual Speech Corpus (1.0)" [Data set]. Zenodo, http://10.5281/zenodo.3625687

Crosse, M. J., and Lalor, E. C. (**2014**). "Retraction: The cortical representation of the speech envelope is earlier for audiovisual speech than audio speech," J. Neurophysiol. **112**(10), 2667.

Crosse, M. J., Liberto, G. M. D., and Lalor, E. C. (**2016**). "Eye can hear clearly now: Inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration," J. Neurosci. **36**, 9888–9895.
Crosse, M. J., Zuk, N. J., Di Liberto, G. M., Nidiffer, A. R., Molholm, S., and Lalor, E. C. (**2021**). "Linear modeling of neurophysiological responses to speech and other continuous stimuli: Methodological considerations for applied research," Front. Neurosci. **15**, 705621.
Czyzewski, A., Kostek, B., Bratoszewski, P., Kotus, J., and Szykulski, M. (**2017**). "An audio-visual corpus for multimodal automatic speech recognition," J. Intell. Inf. Syst. **49**(2), 167–192.
Ding, N., and Simon, J. Z. (**2012**). "Emergence of neural encoding of auditory objects while listening to competing speakers," Proc. Natl. Acad. Sci. U.S.A. **109**(29), 11854–11859.
Ding, N., and Simon, J. Z. (**2014**). "Cortical entrainment to continuous speech: Functional roles and interpretations," Front. Hum. Neurosci. **8**, 311.
Fu, Z., Wu, X., and Chen, J. (**2019**). "Congruent audiovisual speech enhances auditory attention decoding with eeg," J. Neural Eng. **16**(6), 066033.
Giraud, A. L., and Poeppel, D. (**2012**). "Cortical oscillations and speech processing: Emerging computational principles and operations," Nat. Neurosci. **15**, 511–517.
Grant, K. W., and Seitz, P.-F. (**2000**). "The use of visible speech cues for improving auditory detection of spoken sentences," J. Acoust. Soc. Am. **108**, 1197–1208.
Guilleminot, P., and Reichenbach, T. (**2022**). "Enhancement of speech-in-noise comprehension through vibrotactile stimulation at the syllabic rate," Proc. Natl. Acad. Sci. U.S.A. **119**(13), e2117000119.
Harte, N., and Gillen, E. (**2015**). "TCD:TIMIT: An audio-visual corpus of continuous speech," IEEE Trans. Multimedia **17**(5), 603–615.
Hauswald, A., Lithari, C., Collignon, O., Leonardelli, E., and Weisz, N. (**2018**). "A visual cortical network for deriving phonological information from intelligible lip movements," Curr. Biol. **28**(9), 1453–1459.
ITU-T (**2011**). "ITU-T P.56: Objective measurement of active speech level (12/2011)," in *International Telecommunication Union Recommendation E 37305, Series P: Terminals and Subjective and Objective Assessment Methods - Objective Measuring Apparatus* (Geneva, Switzerland).
Lansing, A. (**1959**). *Endurance: Shackleton's Incredible Voyage* (Hodder and Stoughton, London, UK).
Lee, B., Hasegawa-Johnson, M., Goudeseune, C., Kamdar, S., Borys, S., Liu, M., and Huang, T. (**2004**). "AVICAR: Audio-visual speech corpus in a car environment," in *8th International Conference on Spoken Language Processing (ICSLP)* (Jeju, Jeju Island, Republic of Korea), pp. 2489–2492.
Luck, S. J. (**2014**). *An Introduction to the Event-Related Potential Technique*, 2nd ed. (MIT Press, Cambridge, MA).
Marmarelis, V. Z. (**2004**). *Nonlinear Dynamic Modeling of Physiological Systems*, iEEE Press Series on Biomedical Engineering (Wiley Hoboken, NJ).
Matthews, I., Cootes, T. F., Bangham, J. A., Co, S., and Harvey, R. (**2002**). "Extraction of visual features for lipreading," IEEE Trans. Pattern Anal. Mach. Intell. **24**(2), 198–213.
McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (**2017**). "Montreal forced aligner: Trainable text-speech alignment using Kaldi," Interspeech **2017**, 498–502.
McGurk, H., and MacDonald, J. (**1976**). "Hearing lips and seeing voices," Nature **264**, 746–748.
Mégevand, P., Mercier, M. R., Groppe, D. M., Golumbic, E. Z., Mesgarani, N., Beauchamp, M. S., Schroeder, C. E., and Mehta, A. D. (**2020**). "Crossmodal phase reset and evoked responses provide complementary mechanisms for the influence of visual speech in auditory cortex," J. Neurosci. **40**(44), 8530–8542.
Messer, K., Matas, J., Kittler, J., Luettin, J., and Maître, G. (**1999**). "XM2VTSDB: The extended M2VTS database," in *Proceedings of the Second International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, Washington, DC.
Moses, D. A., Mesgarani, N., Leonard, M. K., and Chang, E. F. (**2016**). "Neural speech recognition: Continuous phoneme decoding using spatio-temporal representations of human cortical activity," J. Neural Eng. **13**(5), 056004.
Movellan, J. R. (**1995**). "TULIPS1 database," in *Advances in Neural Information Processing Systems 7* (MIT Press, Cambridge, MA).
O'Sullivan, A. E., Crosse, M. J., Di Liberto, G. M., de Cheveigné, A., and Lalor, E. C. (**2021**). "Neurophysiological indices of audiovisual speech processing reveal a hierarchy of multisensory integration effects," J. Neurosci. **41**(23), 4991–5003.
O'Sullivan, A. E., Crosse, M. J., Liberto, G. M. D., and Lalor, E. C. (**2017**). "Visual cortical entrainment to motion and categorical speech features during silent lipreading," Front. Hum. Neurosci. **10**, 679.
O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., and Lalor, E. C. (**2015**). "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," Cereb. Cortex **25**, 1697–1706.
Reisberg, D., McLean, J., and Goldfield, A. (**1987**). "When half a face is as good as a whole: Effects of simple substantial occlusion on visual and audiovisual speech perception," in *The Psychology of Lip-Reading* (MIT Press, Hillsdale, NJ), pp. 97–114.
Reverdy, J., Russell, S. O., Duquenne, L., Garaialde, D., Cowan, B. R., and Harte, N. (**2022**). "Roomreader: A multimodal corpus of online multiparty conversational interactions," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 2517–2527.
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," Cereb. Cortex **17**, 1147–1153.
Sanderson, C., and Lovell, B. C. (**2009**). "Multi-region probabilistic histograms for robust and scalable identity inference," in *Advances in Biometrics*, edited by M. Tistarelli and M. S. Nixon (Springer, Berlin, Heidelberg), pp. 199–208.
Schultz, B., Biau, E., and Kotz, S. (**2020**). "An open-source toolbox for measuring dynamic video framerates and synchronizing video stimuli with neural and behavioral responses," J. Neurosci. Methods **343**, 108830.
Sonkusare, S., Breakspear, M., and Guo, C. (**2019**). "Naturalistic stimuli in neuroscience: Critically acclaimed," Trends Cognit. Sci. **23**(8), 699–714.
Styler, W. (**2013**). "Using Praat for Linguistic Research," University of Colorado at Boulder Phonetics Lab.
Sumby, W. H., and Pollack, I. (**1954**). "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Am. **26**, 212–215.
Suwajanakorn, S., Seitz, S. M., and Kemelmacher-Shlizerman, I. (**2017**). "Synthesizing Obama: Learning lip sync from audio," ACM Trans. Graph. **36**(4), 1–13.
Varano, E., Vougioukas, K., Ma, P., Petridis, S., Pantic, M., and Reichenbach, T. (**2022**). "Speech-driven facial animations improve speech-in-noise comprehension of humans," Front. Neurosci. **15**, 781196.
Weissbart, H., Kandylaki, K. D., and Reichenbach, T. (**2020**). "Cortical tracking of surprisal during continuous speech comprehension," J. Cognit. Neurosci. **32**, 155–166.