






A Data Collection Protocol, Tool and Analysis for the Mapping of Speech Volume to Avatar Facial Animation

R. Miyawaki¹ , M. Perusquia-Hernandez¹ , N. Isoyama¹ , H. Uchiyama¹  and K. Kiyokawa¹ 

¹Nara Institute of Science and Technology, Japan

Abstract

Knowing the relationship between speech-related facial movement and speech is important for avatar animation. Accurate facial displays are necessary to convey perceptual speech characteristics fully. Recently, an effort has been made to infer the relationship between facial movement and speech with data-driven methodologies using computer vision. To this aim, we propose to use blendshape-based facial movement tracking, because it can be easily translated to avatar movement. Furthermore, we present a protocol for audio-visual and behavioral data collection and a tool running on WEB that aids in collecting and synchronizing data. As a start, we provide a database of six Japanese participants reading emotion-related scripts at different volume levels. Using this methodology, we found a relationship between speech volume and facial movement around the nose, cheek, mouth, and head pitch. We hope that our protocols, WEB-based tool, and collected data will be useful for other scientists to derive models for avatar animation.

CCS Concepts

• **Human-centered computing** → **Visualization toolkits**;

1. Introduction

In recent years, communicating via video chat tools using a 3D avatar has become commonly used. Accordingly, avatar communication platforms have become popular. For instance, Mesh for Microsoft Teams [Mic], where people can use their avatars in conversational or collaborative situations. In addition, high-fidelity and customized avatar-creation tools are becoming more accessible. For example, MetaHuman [Epia] and its creation tool, which does not require any rigging or modeling expertise. This social tendency and technological development have the potential to establish a new style of Human-Avatar Interaction in a virtual environment or even a common video chat tool [Zoo]. For example, people can improve their public speaking skills by watching their presentation video as a reflection [ZFKK21].

To advance these technologies, it is necessary to increase our understanding of human behavior, and how it can be mapped to avatars in a computer-vision manner. Furthermore, we should better understand the interaction between humans and avatars. Most of the knowledge used to animate an avatar's facial movement is derived using the Facial Action Coding System (FACS) [EF82]. The FACS is a facial movement descriptor framework. The advantage of using the FACS is that it only describes units of movement, or Action Units (AU), without making inferences about the movement's meaning. Interpretation can then be left to an expert, or to the perceivers of the movement. Due to its precise movement descriptions, experts in affective sciences and computer graphics communicate

their research on facial movement using the FACS as a framework. However, estimating AUs precisely is a computationally expensive and difficult task to do with off-the-shelf software such as OpenFace toolkit 2.0 [BZLM18] and Py-Feat [CXBC21]. An alternative to using the FACS is to represent facial behaviors with blendshape values. Blendshapes are geometric meshes [LAR*14] that are used as movement units in three-dimensional (3D) avatars. They are used as rigs for the animation of virtual humans. Previous work has explored the importance of using blendshapes related to AUs perceptual units to convey information using facial displays [CZDM20]. A combination between both FACS and blendshape can be used as an animation method [ABMS15]. Blendshape interpolation is used to create a natural face, and the edited face was created by combining 3D models of four facial expressions: anger, fear, happy, and sad. These expressive 3D models were created based on FACS. Anger was depicted with AU (26 + 4 + 17 + 10 + 9 + 20 + 2); fear with AU (2 + 4 + 5 + 26 + 15 + 20 + 1); happy with AU (14 + 12 + 6 + 1); and sad with AU (23 + 1 + 15 + 4).

The facial movement also occurs during speech. Whereas phonemes are recognized as the audio unit of speech, visemes are the visual equivalent [Fis68]. Visemes have been previously defined as a set of phonemes that have an identical appearance on the lips [BH17]. Even though the visual appearance of speech has been extensively researched, there is still work to do in regard to how visual representations are mapped to different perceptual characteris-

tics, such as emotion and volume levels, to communicate the same perceptual message using animation in virtual humans. An example of lip-specific animations is **JALI** [ELFS16]. They attempted to build the jaw and lip action model and validated it for end-to-end automated speech-synchronized animation.

The relationship between speech and visual appearance has also been deemed relevant by previous research. Voice realism had an impact in the sense of presence in VR, but not on empathic responses [HZC*22]. Additionally, voice similarity between an avatar and its user is related to increased performance, time spent, identification, competence, relatedness, and immersion [KRMM21]; and voice customization moderates the effect of avatar appearance customization [KRM*22]. These works show the importance of addressing the interplay between facial movement and speech sound.

Using existing datasets is beneficial to analyze the relationship between facial movement and speech in multiple contexts. There are several existing datasets of audio-visual speech. Examples are the LRW-1000 dataset is a popular resource for lip reading data in-the-wild [YZF*19]; the LRS2 [ACS*18] is a dataset extracted from BBC television; and the LRS3 [ACZ18] contains data extracted from TED and TEDx videos. There are also datasets that are not in English and not publicly available. For unreleased Japanese language datasets, Shirakata and Saitoh originally collected facial video and speech audio to predict sentences from silent videos on talking [SS21] with *ATR phoneme balance sentences (ATR)* [KTS*90] and *Inter-field Task Accelerating (ITA) Corpus* [JOK*]. The ATR contains 503 sentences and was created with careful attention to phonemic balance from randomly selected Japanese literature. ITA is a relatively new Japanese corpus under the public domain and was created with both phonemic balance and easiness to read. Datasets published so far that include facial movement, especially the mouth and lips, hint at the importance of understanding the relationship between facial movement and speech. Owing to this, researchers need datasets in various languages and multi-modal data to further improve avatar facial animation during speech. Nevertheless, each dataset has used different elicitation protocols and the data collection apparatus is often not released. This hinders the ability of other research groups to collect data that can be appended to increase the understanding of speech production and use it for avatar design and customization. To this aim, we need a simple tool and a standard protocol to collect multi-modal data using various speech corpus.

The multi-modal datasets can be used in multi-modal human-avatar or human-agent interaction studies as well. In a recent study, Higgins et al. [HZC*22] investigated preferred voice types with a photorealistic agent. To create stimuli in their part of the experiments, they captured the body, face, and voice of a single female actor. There are also several toolboxes to animate blendshapes. Worth mentioning is HeadBox [VOJ*22], an open-source facial animation tool for Microsoft RocketBox avatar library [GFOP*20]. The Headbox has a total of 15 visemes and 48 FACS AUs, from which 30 can be used with the VIVE Facial Tracker. Users can use this tool with Unity as a platform, OpenFace as a facial feature detector, and Oculus Lipsync as a viseme. Furthermore, they also introduced a Python script for Maya to generate standardized ARKit Blend-

shapes [App]. This work has the potential that allows researchers to manipulate the facial blendshapes of the wide variety of rigged avatars.

In this paper, we introduce a protocol and a WEB-based recording system of speech at different volume levels. Facial blendshapes and head and eyes rotation are recorded along with speech audio and facial video. Using relative displacements from the neutral face as descriptors has the advantage that the values can be mapped into an avatar's facial movement directly. In this manner, we do not need to rely on intermediary AU detection [VOJ*22]. We collected data from nine participants to demonstrate analysis possibilities for both head and facial movement and audio at different speaking volume levels. To our knowledge, existing literature did not describe the relationship between standardized blendshape-based facial movement tracking, head and eyes movement, and speech produced at different volume levels. Most research is done following the FACS. Hence, we expect our protocol and findings to be helpful to continue exploring the relationship between facial movement tracked with blendshape values and speech.

In summary, our contributions are:

- Development of an open-source WEB-based system and an experimental protocol for the collection of a multi-modal dataset of speech and speech-related head, eyes, and facial movement.
- The open-source code of the aforementioned WEB-based system.
- A database of Japanese speakers reading emotion-related sentences at four different perceptual volume levels. Synchronized blendshape tracking values, head and eyes tracking values, video, and audio are provided.

2. WEB-based recording system

We developed a recording system running on the WEB browser as shown in Fig. 1. The objective of this system is to record both speech audio and facial videos on a laptop, along with head, eyes, and facial movement data at different levels of speech. The facial movement was recorded with an iOS device. This system ensures that the recorded data is synchronized despite it being captured by different devices. For speech, our system is ready to be used different types of corpora to be able to adapt the data collection to different languages. It is also possible to use other corpora by following the format of an ITA corpus or modifying the programs of the system. In addition, thanks to its WEB-based implementation, it can be executed independently of the platform, making it easy to use. The system is available at [our GitHub open-source repository](#) distributed under the MIT license.

This system consists of five components as shown in Fig. 2. In our protocol, participants filled out their assigned anonymous ID and selected a condition (Fig. 2-A). For audio calibration, they watched an image of a lecture hall (Fig. 2-B) and produced a voiced phoneme at the loudest volume level they could, and we captured their volume values five times. The average of the Root Mean Square (RMS) of the five loud phonemes was stored, together with the minimum voice volume value corresponding to the background sound (Fig. 2-C). During the recording, the participants watched

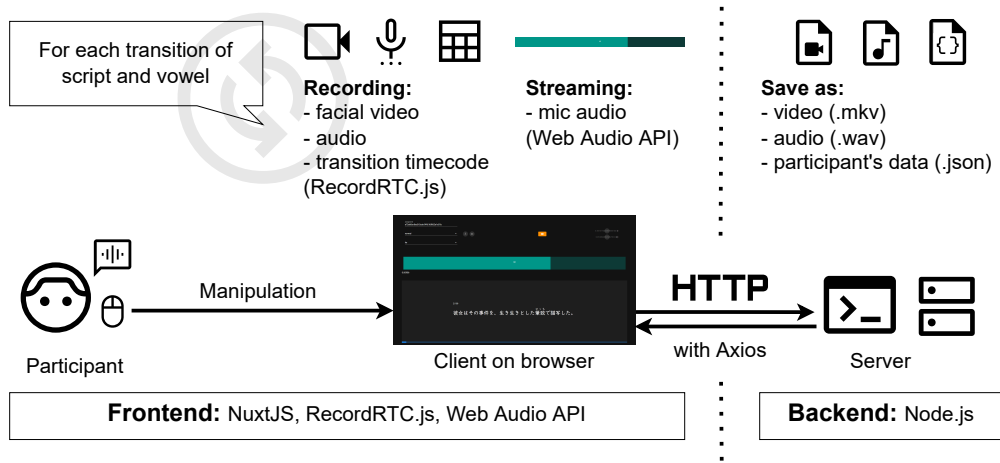


Figure 1: Overview of WEB-based recording system



Figure 2: Recording system running on a WEB browser

the volume indicator (Fig. 2-D) and read the text on each slide. After reading the script, they pressed next to go through a deck of scripts corresponding to the selected corpus (Fig. 2-E).

The output of the system is a set of multi-modal data containing speech, facial video, and blend shape displacement. Speech audio and a facial video per script per condition are stored in the back-end server in Fig. 1. Time codes of transition per script are stored to synchronize audio and videos with the facial movement data. The time series of head, eyes, and facial movement is recorded by Live Link Face [Epib]. In this external app, the time codes are recorded frame by frame, which enables us to synchronize with the other data with small errors.

2.1. Implementation and technical details

We created the whole front end and its components on NuxtJS and used the media libraries, RecordRTC.js, and Web Audio API. The front end enables the calibration and displays a volume indicator with streaming speech audio to provide feedback to the participant. The interface also shows the corpus scripts, records speech audio and facial videos; and sends data to a back-end server every

time the users change the script. The back-end implementation was done with Node.js to store Binary Large Objects (BLOB) which are recorded with the front-end implementation.

Volume indicator. The speech volume is displayed as a bar in real-time (Fig. 2-D). We disabled `echoCancellation`, `noiseSuppression`, and `autoGainControl` of `AudioContext` in Web Audio API so that the indicator can display its bar as it is. Its range is set to average both minimum and maximum RMS recorded during the initial calibration. To provide real-time feedback, the indicator's bar is animated according to the speech RMS value of every captured audio segment. Each segment was defined as a window of 32768 audio samples. The animation had two parameters: the reaction speed of animation and color. The reaction speed depends on the Fast Fourier Transform (FFT) window size (32768 samples/window) and sampling rate (48000 Hz). The FFT was calculated using the Mozilla Web API function `AnalyserNode.fftSize`. The default color of the volume indicator is gray, and it changes to green when participants reach the desired volume range set for each of the four experimental conditions (see Sec. 3.2). The desired ranges are defined as $0 < RMS \leq 25$ on muffled, $25 < RMS \leq 50$ on low volume, $50 < RMS \leq 75$ on normal volume, and $75 < RMS \leq 100$ on high volume condition. We note that the RMS values are normalized using the minimum and maximum values recorded during calibration and multiplied by 100.

Script slider. We set 2 s interval between transitions on clicking the "Next" and "Prev" buttons (Fig. 2-E) to prevent an insufficient length media caused by the automatic transition. We also added a progress bar at the bottom part of the slide for the participant to see how many scripts are left. Immediately after clicking, the back end handles HTTP requests to save audio, video, and participant data.

3. Methods

This protocol for recording a multi-modal dataset can be useful to reproduce data collection in other languages. Protocols, data, and analysis scripts are available in [this work's Open Science Framework \(OSF\) repository](#).



Figure 3: Recording environment

Participants faced a monitor displaying the WEB-based recording system (Fig. 2). A separator was used to give a sense of privacy and minimize contact with the experimenter. The screen of the iPad was turned off to minimize distractions.

3.1. Participants

Nine Japanese native speakers (5 females) agreed to take part in the experiment. Their average age was 26.3 years old (SD = 9.90). In addition, six out of nine participants (2 females) agreed to release their data (22.3 years old, SD = 0.82).

3.2. Experimental design

A within-subjects experiment was conducted. All participants went through two tasks with four conditions. **Task 1** is to speak out 100 different Japanese scripts, and **Task 2** is long-tone of vowels in Japanese, /a/, /i/, /u/, /e/, and /o/. The conditions correspond to four volume levels: (1) normal, (2) high, (3) low, and (4) muffled. To be able to imagine differences among the levels, they first tried (1) and then the other randomized levels. Randomization was applied to both conditions and Japanese scripts to prevent biases due to their fatigue over time.

3.3. Apparatus

We set up a recording environment as shown in Fig. 3.3. The equipment consisted of an iPad Pro (2021), Logitech Webcam C980 (60 fps), AKG Perception 170 Condenser Microphone (48000 Hz), a mouse and a keyboard, a laptop connected with a single display, and GSYXERGILES photography kit with a green background and two light bulbs placed in front of the participant to avoid shadows on the face. A partition was also used to give a sense of privacy to participants. We also intended to reduce the social effects of the experimenter being in the same room as the participant. Tracked facial blendshapes, head/eyes rotations, and their timecodes on 60 fps were recorded by Live Link Face with the iPad Pro which is compatible with ARKit and has a depth sensor. We separately ran Live Link Face and a WEB recording system (see Sec. 2). These were later synchronized with their timecodes connected to the same NPT server (see Sec. 4.1).

3.4. Stimuli

We chose the ITA Corpus [JOK*] as a relatively new Japanese and public domain corpus. Thanks to its entropically phoneme-balanced scripts, we expected to collect a variety of phonemes with equal probability of facial movement values regarding speech audio and viseme. In addition, this corpus considers easy to read. It includes short sentences extracted from Japanese literary works and other corpora. To avoid lengthening the experiment, we preselected the 100-sentence of *ITA-emo*, but not the 424-sentence of *Ita-recit*. According to their previous work, the extended entropy, an indicator of the uniformity of the frequency of occurrence of a phoneme chain of the ITA-emo scripts is 21.85 bits. This value is similar to another popular Japanese corpus, the *ATR503* [KTS*90], whose extended entropy is 22.49 bits. We considered both for our experiment, and eventually, we selected ITA Corpus considering its availability and easiness to read.

3.5. Measurements

We recorded speech audio, facial video, and tracked facial movement including blendshapes and rotations of head and eyes with the iPad Pro. The audio formatted as wave and video as .mkv were recorded on a WEB browser per script. The head and facial tracking data were recorded and used as the degree of their face when relaxed (see Sec. 3.6). We chose widely-used blendshapes, ARKit Blendshapes [App] as a target for analysis because blendshapes are not used only for facial movement descriptors like FACS. The number of tracked blendshapes with ARKit is 52. Additionally, Live Link Face has head and eyes rotation tracking in three axes for a total of 61 tracked variables. The blendshape displacement is tracked from 0 to 1, except for head and eyes rotation, when it is tracked from -1 to 1.

3.6. Procedure

Participants were welcomed and the experiment was explained. If they agreed to participate, they signed an informed consent form. Next, participants were provided with a written explanation of the task. Written instruction was preferred to avoid biasing their behavior as much as possible, as we aimed to investigate natural facial behavior. Participants were provided with the opportunity to ask questions. If everything was clear, they put up their hair with a clip and the chair height was adjusted. This was to prevent occlusion of the face, and therefore, to be able to collect accurate values of facial movement. Next, data for calibration was recorded. For the facial behavior data, the participant's neutral face was recorded with Live Link Face while facing toward a display and gazing at the top point of the camera. This expression was used as an initialization of the head and eye rotation, and facial blendshapes. Subsequently, an audio calibration was performed. The highest volume level of their voice and a fragment of the silent background noise at that moment were recorded. This data was used to personalize the real-time volume feedback indicator (Fig. 2-D). To equalize all participants' understanding of what "the loudest voice volume" means, they were provided with a reference image. They were asked to look at a lecture hall picture (Fig. 2). Then they were asked to produce a voice volume level that would be heard by the audience in

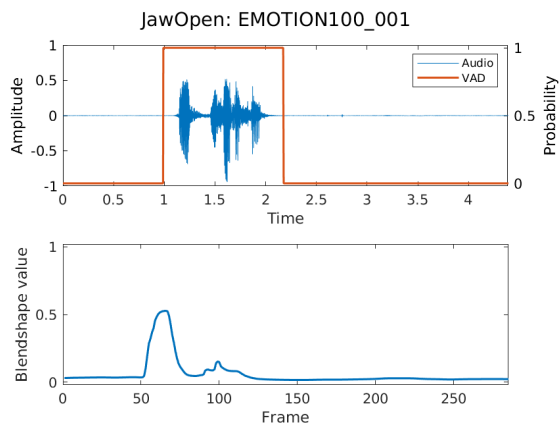


Figure 4: Example of facial movement preceding speech

The upper part shows the whole audio signal (blue) and its result of VAD (orange), while the bottom part shows the whole value of blendshape, “JawOpen”. This script was EMOTION100_001, recorded under the high volume condition (P3).

the last row (“please imagine the volume is loud enough to reach people seated at the far end of the lecture hall”). Then, they spoke out a single phoneme, /a/, for several seconds. After finishing the calibration, participants started to manipulate the slides with each script and vowel shown in Fig. 2-E as practice. We then asked any questions and solved any misunderstandings that the participants had. Participants were then asked not to move largely their head pose, not to touch their mouth while speaking, and not to stop the trials halfway. They were asked to keep speaking on one trial, to assume they were speaking in public, and to keep a constant speed. As a recording trial, participants went through two trials for each of the four-volume conditions. They were also allowed to drink water in between trials. Finally, we disclosed that facial movement data was also recorded during the trials. An additional consent form was used to ask which of the recorded data they allow us to share with others. We then gave a pastry as a reward.

4. Analysis and results

4.1. Pre-processing

First, we segmented the entire series of facial movement data recorded with Live Link Face per script per condition. Besides, the speech audio was also segmented. The average duration of the script segments was 4.49 s (270 frames). The pre-processing was done in Python 3.10, and MATLAB22a was used for plotting and for statistical analyses.

Audio pre-processing. We applied a noise reduction and a normalization with a sound processing software *SoX*. We inferred voiced and unvoiced audio regions for each script with a Voice Activity Detection (VAD) method: *rVAD* [TkSD20]. *rVAD* is an unsupervised segment-based binary classification for voice detection.

Blendshape pre-processing. The movements were tracked directly as relative displacements from the neutral pose calibrated

before the experiment. The facial and head movement per script was processed to extract the segments of facial movement going from their onset to the end of the voiced section. We removed the data after the completion of the voiced section from both audio and head and facial movement data to remove irrelevant movements produced after the participant finished reading the scripts. We kept the onset of the facial movement as the start of the segment because the participants’ facial movement occasionally preceded their actual voice (see Fig. 4.1). This was observed in previous studies as well [LTR09]. Hence, we determined the onset of the facial movement as a starting point and the end frame of the voiced region as a stopping point. The starting point of the facial movement was defined as the first frame when the value of head and facial movement exceeded 2 SD from the beginning of each trial. That baseline of SD was calculated by averaging the first 500 msec of each condition (i.e., averaged first 500 msec head and facial series of data recorded with Live Link Face). The selected head and facial data had different lengths because participants spoke at different speeds and the scripts had different lengths. When the facial movement was minimal, the length of the calculated segments was very small or zero, leading to problematic re-sampling. Therefore, we removed these segments for further analysis. Therefore, all facial movement data for the figure plots were resampled from the original length to the average script length. This process enabled us to average the number of frames of facial and head movement data in each script and then added them all together to obtain an average. Averaging of the time series was applied twice. One was done for each facial movement data per script and then per condition.

4.2. Validation

In process of selecting the starting and stopping points, sometimes zero or very short segment of facial movement data was generated. A short segment might occur when the participants only moved by the end of the speech trial, which was eliminated for analysis. On resampling, we removed segments with less than 10 frames of data. As for the first averaging per script per condition, if any data with zero or very short segments was found in each participant’s script, the corresponding data was removed across the conditions for each type of facial movement data, from the average and therefore from further analyses. For example, in the “muffled” condition, the lowest volume would have a lower Signal-to-Noise Ratio (SNR) compared with the other conditions because the face did not move as much as in other conditions. Therefore, we carefully checked valid frames. The result of valid frames among the conditions is shown in Table 1. As the table shows, the muffled condition had the highest invalid ratio.

4.3. Head and facial movement

We examined the shapes of the distributions of facial movement because each of their parameters has different characteristics while speaking. First, we looked into the shapes of distribution and found them skewed. Therefore, we chose to plot log-scaled histograms around the y-axis (Fig. 5). Solid lines represent calculated means and dashed lines standard deviation around the mean. In addition, we ran Friedman tests to investigate the effect of volume level on

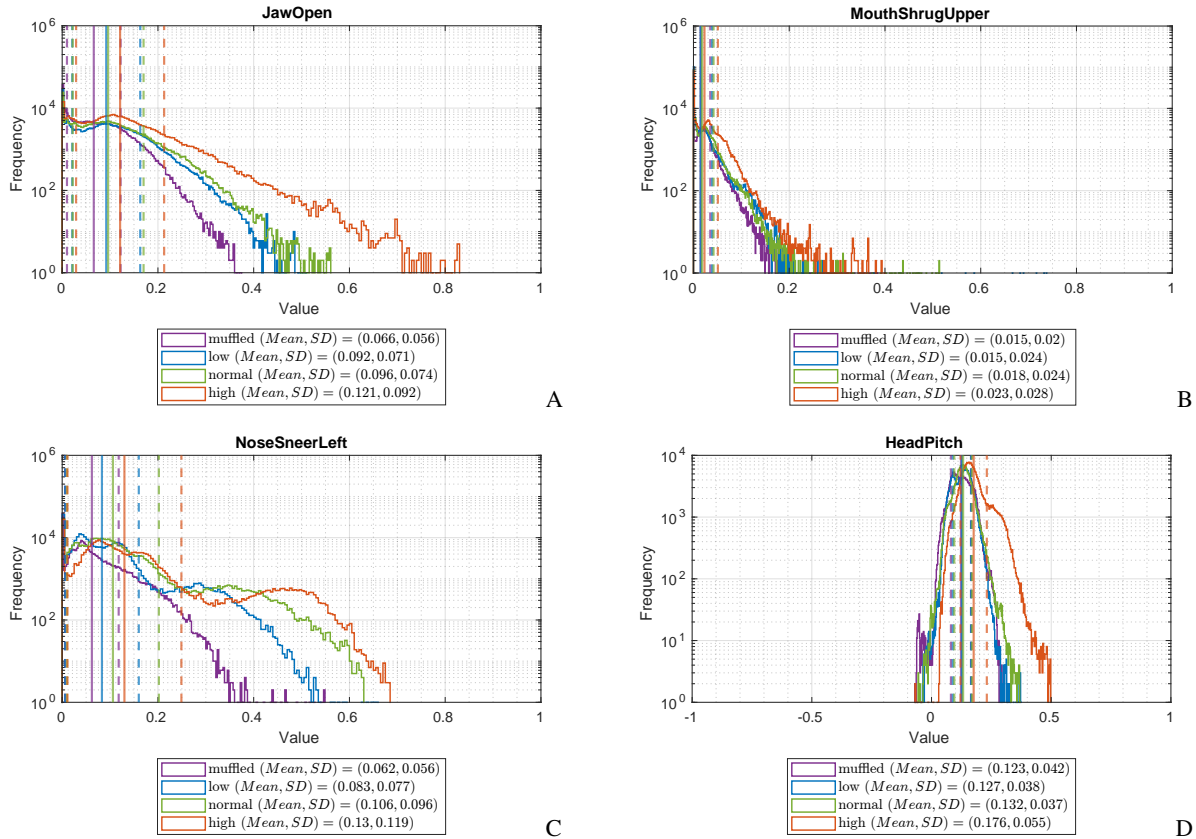


Figure 5: Example of log-scaled histograms of facial and head movement among the four conditions

The conditions are muffled voice, low volume, normal volume, and high volume. Vertical solid lines represent the condition’s mean value, and the dashed lines represent the standard deviation from the mean per condition. High volume was defined as the perceived necessary volume to reach the last row in an auditory hall depicted in a provided image.

Table 1: Total number and ratio of invalid frames on facial movement in Task 1 after pre-processing

Condition	Invalid data	Invalid ratio
muffled	6,047	11.0 %
low	4,696	8.6 %
normal	4,777	8.7 %
high	5,391	9.8 %
total	20,911	9.5 %

Number of entire data for each condition was 54,900 as following:
 $54,900 = \text{scripts} \times \text{participants} \times \text{data types}$

facial movement. We applied pairwise comparisons with Bonferroni correction to provide insight into what pairs and types of movement are meaningful for different speech volume levels.

Here, we took some examples from 61 types of data (52 of ARKit Blendshapes [App], three of head rotations, and six of eyes rotations) and show the result of histogram differences in Fig. 5. There were significant differences on A) “JawOpen” ($\chi^2(3) = 22.73$) between muffled-high ($p < .0001$) and low-high ($p < .05$). Even though B) “MouthShrugUpper” ($\chi^2(3) = 10.05$)

is also a blendshape around the mouth, we found smaller differences of average between the volume levels on the muffled-normal ($p < .05$). In terms of a non-mouth but around a nose blendshapes, C) “NoseSneerLeft” ($\chi^2(3) = 16.60$), significant differences were confirmed on the muffled-normal ($p < .05$) and muffled-high ($p < .001$). Further, the mean of D) “HeadPitch” ($\chi^2(3) = 13.40$) in the high condition was higher than the others and significant differences were found on the muffled-low ($p < .05$) and muffled-high ($p < .01$).

A comparison result for the facial data is shown in 2. Speaking of pairs of conditions, we have six pairs: muffled-low, muffled-normal, muffled-high, low-normal, low-high, and normal-high. Finally, we chose muffled-normal, muffled-high, and low-high conditions with at least one significant difference for each facial movement data. An exception to this was “HeadPitch”, which has one significant difference in muffled-low as we already mentioned in the histogram. We note the facial movements are described in the Apple’s document [App] except for the head and eye rotations.

Table 2: The extracted result of pairwise comparisons after Friedman test and Bonferroni correction for multiple comparisons

Facial and Head Movements	$\chi^2(3)$	muffled-normal <i>p</i>	muffled-high <i>p</i>	low-high <i>p</i>
CheekSquintLeft	21.00	0.01**	0.00***	0.02*
CheekSquintRight	20.60	0.04*	0.00***	0.00**
HeadPitch	13.40	0.17	0.00**	1.00
JawForward	24.73	0.02*	0.00***	0.01**
JawOpen	22.73	0.11	0.00***	0.04**
MouthLowerDownLeft	22.47	0.02*	0.00***	0.02*
MouthLowerDownRight	23.80	0.04*	0.00***	0.01*
MouthPressLeft	15.80	0.04*	0.00***	0.27
MouthPressRight	13.93	0.11	0.00**	0.27
MouthStretchLeft	21.00	0.17	0.00***	0.01**
MouthStretchRight	17.93	0.27	0.00***	0.01*
MouthUpperUpLeft	21.67	0.01*	0.00***	0.01*
MouthUpperUpRight	21.67	0.01*	0.00***	0.01*
NoseSneerLeft	16.60	0.02*	0.00***	0.17
NoseSneerRight	17.27	0.01*	0.00**	0.17

****p* < .001, ***p* < .01, **p* < .05All *p* values are rounded to two decimal places.Outside the table, HeadPitch was significantly different in the muffled-low condition (*p* = .04).

Note that only significant differences are shown. The other 46 facial movements did not differ between volume levels.

5. Discussion and future work

We presented a protocol for audio-visual and behavioral data collection for both well-balanced phonemes and 61 types of facial movement data including 52 types of blendshapes [App] and nine types of head and eyes rotations. Unlike FACS, we argued that using standardized blendshapes and head rotations directly is helpful to animate virtual avatars that can convey perceptual messages in a visual manner. This is because FACS-based detection is often computationally expensive and not directly transferable to mesh representations. We also created a WEB-based system that aids synchronized multi-modal data collection of speech and facial movement. Researchers can use and customize the open-source system to analyze data or create models to meet their individual research purposes. This is important to further analyze facial movement characteristics to convey a sound-related message with virtual human animations. As a start, we provided a database of six Japanese participants reading emotion-related scripts as an example. We also explored the characteristics of this dataset by checking both speech audio and facial movement regarding avatar animation.

Out of 61 different facial movements, we were able to extract 15 significantly different facial movements for each pair of conditions. As we expected, our analyses showed that facial movements around the jaw and mouth are the most relevant when producing speech at different volume levels as shown in 2. It is quite natural to think that a higher level of voice increase values of the jaw movement. Interestingly, the significantly different blendshape displacements are not limited only to the jaw movement, as suggested by previous work [ELFS16]. For example, “CheekSquintLeft” and Right, upward movements of the cheek, are significantly different. It was pointed out in the previous literature [BDY*04] that cheek areas also give valuable information for emotion classification, which provides relevance to speaking emotion-related sentences. In addition, movement around the nose is also relevant and should be

considered for avatar animation. Surprisingly, the head pitch was also different between different volume conditions. A possible explanation is that the participants needed to take more air in when speaking out loud.

A limitation of this analysis is that to analyze the facial movement data, we focused mostly on voiced segments, and it is possible that the facial movement is not perfectly synchronized with the voice. Future work should explore in more detail the delay between facial movement and speech production. Furthermore, we only analyzed volume levels. The dataset also contains information about the emotion conveyed by the scripts read by the participants. We hope that the scientific community will use our WEB-based tool to collect more data following this protocol and that our dataset will be used to do more analyses in this direction.

References

- [ABMS15] ALKAWAZ M., BASORI A. H., MOHAMAD D., SABA T.: Blend shape interpolation and facs for realistic avatar. *3D Research* 6 (01 2015), 1–10. doi:10.1007/s13319-015-0038-7. 1
- [ACS*18] AFOURAS T., CHUNG J. S., SENIOR A., VINYALS O., ZISSERMAN A.: Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018), 1–1. doi:10.1109/TPAMI.2018.2889052. 2
- [ACZ18] AFOURAS T., CHUNG J. S., ZISSERMAN A.: LRS3-TED: a large-scale dataset for visual speech recognition. *CoRR abs/1809.00496* (2018). URL: <http://arxiv.org/abs/1809.00496>, arXiv:1809.00496. 2
- [App] APPLE: ARKit blendShapes - Apple Developer Documentation. Accessed: 2022-09-17. URL: <https://developer.apple.com/documentation/arkit/faceanchor/2928251-blendshapes>. 2, 4, 6, 7
- [BDY*04] BUSSO C., DENG Z., YILDIRIM S., BULUT M., LEE C. M., KAZEMZADEH A., LEE S., NEUMANN U., NARAYANAN S.: Analysis of emotion recognition using facial expressions, speech

- and multimodal information. In *Proceedings of the 6th International Conference on Multimodal Interfaces* (New York, NY, USA, 2004), ICMI '04, Association for Computing Machinery, p. 205–211. URL: <https://doi.org/10.1145/1027933.1027968>, doi:10.1145/1027933.1027968. 7
- [BH17] BEAR H. L., HARVEY R.: Phoneme-to-viseme mappings: the good, the bad, and the ugly. *Speech Communication* 95 (2017), 40–67. URL: <https://www.sciencedirect.com/science/article/pii/S0167639317300286>, doi:<https://doi.org/10.1016/j.specom.2017.07.001>. 1
- [BZLM18] BALTRUSAITIS T., ZADEH A., LIM Y. C., MORENCY L.-P.: Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (2018), pp. 59–66. doi:10.1109/FG.2018.00019. 1
- [CXBC21] CHEONG J. H., XIE T., BYRNE S., CHANG L. J.: Py-feat: Python facial expression analysis toolbox. *CoRR abs/2104.03509* (2021). URL: <https://arxiv.org/abs/2104.03509>, arXiv:2104.03509. 1
- [CZDM20] CARRIGAN E., ZIBREK K., DAHYOT R., MCDONNELL R.: Investigating perceptually based models to predict importance of facial blendshapes. In *Proceedings of the 13th ACM SIGGRAPH Conference on Motion, Interaction and Games* (New York, NY, USA, 2020), MIG '20, Association for Computing Machinery. URL: <https://doi.org/10.1145/3424636.3426904>, doi:10.1145/3424636.3426904. 1
- [EF82] EKMAN P., FRIESEN W. P.: Measuring facial movement with the Facial Action Coding System. In *Emotion in the human face*, Ekman P. (Ed.), second ed. Cambridge University Press, 1982, ch. 9, pp. 178–211. 1
- [ELFS16] EDWARDS P., LANDRETH C., FIUME E., SINGH K.: Jali: An animator-centric viseme model for expressive lip synchronization. *ACM Trans. Graph.* 35, 4 (jul 2016). URL: <https://doi.org/10.1145/2897824.2925984>, doi:10.1145/2897824.2925984. 2, 7
- [Epic] EPIC GAMES: MetaHuman - high-fidelity digital humans made easy. Accessed: 2022-09-17. URL: <https://www.unrealengine.com/en-US/metahuman>. 1
- [Epic] EPIC GAMES: Recording facial animation from an ios device. Accessed: 2022-08-20. URL: <https://docs.unrealengine.com/5.0/en-US/recording-face-animation-on-ios-device-in-unreal-engine/>. 3
- [Fis68] FISHER C. G.: Confusions among visually perceived consonants. *Journal of Speech and Hearing Research* 11, 4 (1968), 796–804. URL: <https://pubs.asha.org/doi/abs/10.1044/jshr.1104.796>, arXiv:<https://pubs.asha.org/doi/pdf/10.1044/jshr.1104.796>, doi:10.1044/jshr.1104.796. 1
- [GFOP*20] GONZALEZ-FRANCO M., OFEK E., PAN Y., ANTLEY A., STEED A., SPANLANG B., MASELLI A., BANAKOU D., PELECHANO N., ORTS-ESCOLANO S., ORVALHO V., TRUTOIU L., WOJCIK M., SANCHEZ-VIVES M. V., BAILENSON J., SLATER M., LANIER J.: The rocketbox library and the utility of freely available rigged avatars. *Frontiers in Virtual Reality* 1 (2020). URL: <https://www.frontiersin.org/articles/10.3389/frvir.2020.561558>, doi:10.3389/frvir.2020.561558. 2
- [HZC*22] HIGGINS D., ZIBREK K., CABRAL J., EGAN D., MCDONNELL R.: Sympathy for the digital: Influence of synthetic voice on affinity, social presence and empathy for photorealistic virtual humans. *Computers & Graphics* 104 (2022), 116–128. URL: <https://www.sciencedirect.com/science/article/pii/S0097849322000474>, doi:<https://doi.org/10.1016/j.cag.2022.03.009>. 2
- [JOK*] JUNYA, OGUCHI F., KANAI Y., ODA T., SAITO M., MORISE: Ita-corporus. Accessed: 2022-08-19. URL: <https://github.com/mmorise/ita-corporus>. 2, 4
- [KRM*22] KAO D., RATAN R., MOUSAS C., JOSHI A., MELCER E. F.: Audio matters too: How audial avatar customization enhances visual avatar customization. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2022), CHI '22, Association for Computing Machinery. URL: <https://doi.org/10.1145/3491102.3501848>, doi:10.1145/3491102.3501848. 2
- [KRMM21] KAO D., RATAN R., MOUSAS C., MAGANA A. J.: The Effects of a Self-Similar Avatar Voice in Educational Games. *Proceedings of the ACM on Human-Computer Interaction* 5, CHI PLAY (2021), 238:1–238:28. URL: <https://doi.org/10.1145/3474665>, doi:10.1145/3474665. 2
- [KTS*90] KUREMATSU A., TAKEDA K., SAGISAKA Y., KATAGIRI S., KUWABARA H., SHIKANO K.: Atr japanese speech database as a tool of speech recognition and synthesis. *Speech Communication* 9, 4 (1990), 357–363. URL: <https://www.sciencedirect.com/science/article/pii/016763939090011W>, doi:[https://doi.org/10.1016/0167-6393\(90\)90011-W](https://doi.org/10.1016/0167-6393(90)90011-W). 2, 4
- [LAR*14] LEWIS J. P., ANJYO K., RHEE T., ZHANG M., PIGHIN F., DENG Z.: Practice and Theory of Blendshape Facial Models. In *Eurographics 2014 - State of the Art Reports* (2014), Lefebvre S., Spagnuolo M., (Eds.), The Eurographics Association. doi:10.2312/egst.20141042. 1
- [LTR09] LIVINGSTONE S. R., THOMPSON W. F., RUSSO F. A.: Facial Expressions and Emotional Singing: A Study of Perception and Production with Motion Capture and Electromyography. *Music Perception* 26, 5 (06 2009), 475–488. URL: <https://doi.org/10.1525/mp.2009.26.5.475>, arXiv:https://online.ucpress.edu/mp/article-pdf/26/5/475/564372/mp_2009_26_5_475.pdf, doi:10.1525/mp.2009.26.5.475. 5
- [Mic] MICROSOFT: Mesh for microsoft teams aims to make collaboration in the 'metaverse' personal and fun. Accessed: 2022-09-17. URL: <https://news.microsoft.com/innovation-stories/mesh-for-microsoft-teams/>. 1
- [SS21] SHIRAKATA T., SAITOH T.: Japanese sentence dataset for lip-reading. In *2021 17th International Conference on Machine Vision and Applications (MVA)* (2021), pp. 1–5. doi:10.23919/MVA51890.2021.9511353. 2
- [TKSD20] TAN Z.-H., KR. SARKAR A., DEHAK N.: rvad: An unsupervised segment-based robust voice activity detection method. *Computer Speech & Language* 59 (2020), 1–21. URL: <https://www.sciencedirect.com/science/article/pii/S0885230819300920>, doi:<https://doi.org/10.1016/j.csl.2019.06.005>. 5
- [VOJ*22] VOLONTE M., OFEK E., JAKUBZAK K., BRUNER S., GONZALEZ-FRANCO M.: Headbox: A facial blendshape animation toolkit for the microsoft rocketbox library. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* (2022), pp. 39–42. doi:10.1109/VRW55335.2022.00015. 2
- [YZF*19] YANG S., ZHANG Y., FENG D., YANG M., WANG C., XIAO J., LONG K., SHAN S., CHEN X.: Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)* (2019), pp. 1–8. doi:10.1109/FG.2019.8756582. 2
- [ZFKK21] ZHOU H., FUJIMOTO Y., KANBARA M., KATO H.: Virtual reality as a reflection technique for public speaking training. *Applied Sciences* 11, 9 (2021), 3988. 1
- [Zoo] ZOOM VIDEO COMMUNICATIONS: Using avatars in meetings and webinars. Accessed: 2022-09-17. URL: <https://support.zoom.us/hc/en-us/articles/4642184011917-Using-Avatars-in-meetings-and-webinars/>. 1