Visual Hearing Aids: Artificial Visual Speech Stimuli for Audiovisual Speech Perception in Noise

Zubin Datta Choudhary
zubin.choudhary@ucf.edu
University of Central Florida
Orlando, Florida, USA

Gerd Bruder
bruder@ucf.edu
University of Central Florida
Orlando, Florida, USA

Gregory F. Welch
welch@ucf.edu
University of Central Florida
Orlando, Florida, USA

ABSTRACT

Speech perception is optimal in quiet environments, but noise can impair comprehension and increase errors. In these situations, lip reading can help, but it is not always possible, such as during an audio call or when wearing a face mask. One approach to improve speech perception in these situations is to use an artificial visual lip reading aid. In this paper, we present a user study ($N = 17$) in which we compared three levels of audio stimuli visualizations and two levels of modulating the appearance of the visualization based on the speech signal, and we compared them against two control conditions: an audio-only condition, and a real human speaking. We measured participants' *speech reception thresholds* (SRTs) to understand the effects of these visualizations on speech perception in noise. These thresholds indicate the decibel levels of the speech signal that are necessary for a listener to receive the speech correctly 50% of the time. Additionally, we measured the usability of the approaches and the user experience. We found that the different artificial visualizations improved participants' speech reception compared to the audio-only baseline condition, but they were significantly poorer than the real human condition. This suggests that different visualizations can improve speech perception when the speaker's face is not available. However, we also discuss limitations of current plug-and-play lip sync software and abstract representations of the speaker in the context of speech perception.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in visualization**; • **Computing methodologies** → **Computer graphics**.

KEYWORDS

Speech perception, background noise, hearing, speechreading, visualizations, virtual humans, user study

ACM Reference Format:
Zubin Datta Choudhary, Gerd Bruder, and Gregory F. Welch. 2023. Visual Hearing Aids: Artificial Visual Speech Stimuli for Audiovisual Speech Perception in Noise. In *VRST '23: 29th ACM Symposium on Virtual Reality Software and Technology, October 09–11, 2023, Christchurch, NZ*. ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

VRST '23, October 09–11, 2023, Christchurch, NZ

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Humans leverage and integrate both auditory and visual sensory information to understand speech, making speech perception a multi-modal process. For example, people who are deaf or hard of hearing depend on the visual component of speech. Lip reading, or speech reading, is our ability to understand speech by interpreting the facial movements of the speaker. However, there are situations where we cannot access the visual component of speech, such as when we make an audio call or the speaker wears a face mask. For the listener, it may become difficult to understand the speaker if the listener's and/or speaker's environment is noisy, which can be functionally equivalent to a decrease in the acoustic signal-to-noise ratio (*SNR*). In similar situations, where the *SNR* is low, humans typically benefit from the visual component significantly more than when the *SNR* is high [31]. However, in audio calls, we cannot see the speaker. Those situations present an opportunity to improve speech perception by adding *visual hearing aids*, i.e., artificial visual speech stimuli to improve speech perception in noise in the absence of the speaker's face.

Researchers in the past have proposed different artificial visual stimuli or alternatives to human lips that could possibly aid speech perception, such as Bernstein et al. [4], who evaluated an oval shape, or the "Lissajous" shape, that would modulate its vertical extent based on the acoustic speech amplitudes. They found that when the shape was modulated in synchrony with the audio signal, participants' speech perception was higher than when no visual stimuli were presented. This indicates that some artificial visual stimuli are viable alternatives to human lip movements for speech perception. Another potential visual aid could be the use of virtual humans (VHs) with accurate lip synchronization. VHs can provide users a strong sense of *social presence* or *co-presence*, which denote how much users feel that a VH is socially connected or co-located and present in the same space with them [19, 27]. A strong sense of social and co-presence can make users perceive the VH more realistic and human-like [7]. It is expected that higher fidelity of social interactions will provide related benefits and increase engagement with the VH [18, 28]. Among the different characteristics of a VH such as appearance and behavior, speech and lip synchronization are important aspects to creating believable, embodied conversational VHs [12]. However, in terms of lip reading, it is not known if and how current VH lip synchronization implementations affect our ability to understand speech.

In this paper, we simulate a noisy environment and measure user's speech perception when presented different visual stimuli. We prepared three visualizations; the first is akin to Bernstein et al.'s[4] oval condition, which modulated using the speech signal's amplitude or via amplitude-based modulation. The other two are a

VH and the VH lips alone and feature lip sync or viseme-based modulation. Therefore, we ran an user study to evaluate our three visualizations (Oval, VH and VH lips) that are modulated synchronously with speech in two ways (amplitude-based or viseme-based) on participants' speech perception. We compare them to two control conditions: audio only (no visual stimuli) and a real human speaking. In terms of speech perception in noise, we posited the following research questions:

- **RQ1**: How do *artificial visual speech stimuli* compare to an *audio-only* condition?
- **RQ2**: How do *artificial visual speech stimuli* compare to a *real human speaking*?
- **RQ3**: How do *artificial visual speech stimuli* compare to each another?

We additionally measured the usability of the system, the user experience, and asked participants subjective questions about their preferences and the reasoning behind them.

The remainder of this paper is structured as follows. Section 2 provides an overview of related work. Section 3 describes our experiment. The results are presented in Section 4 and discussed in Section 5. Limitations and future work are discussed in Section 6. Section 7 concludes the paper.

2 RELATED WORK

In this section, we discuss related work on human audio-visual speech perception and relevant findings about virtual humans, their lip movements for increased social presence, and alternative visualizations for speech reading.

2.1 Audio Visual Speech Perception

To understand how human listeners recognize speech and use it to understand spoken language, researchers have been extensively studying human speech perception. It has been traditionally assumed that speech perception in face-to-face contexts is a purely uni-modal (i.e., auditory) process and that visual cues are independent or additive. Since then, seminal work done on the McGurk Effect [17] and by McDonald et al. [16] demonstrated the importance of visual cues for human speech processing.

Today, it is known that lip reading, or speech reading, is part of our daily lives, and we depend on visual cues to clearly understand another person speaking [11, 23, 32]. We naturally take advantage of the patterns created by our lips, tongue, teeth, etc. as we speak. While vision has proven beneficial, especially for people who are deaf or hard of hearing, work done by Fowler et al. [9] and Sam et al. [25], showed that speech can also be *felt*, providing further evidence for the multi-modal nature of speech perception [24].

On a daily basis, humans are expected to clearly listen to, understand, and respond to a wide variety of visual and aural conditions. Some conditions might aid communication, while others may have adverse effects, such as environmental aural noise or the absence of visuals [3]. In such adverse or difficult listening situations, it can be considered to be functionally equivalent to a decrease in the acoustic SNR. Humans have demonstrated significant tolerance for audiovisual speech perception at low SNRs by relying on visual cues significantly more than in situations with higher SNRs [31]. Furthermore, Jordan et al. [14] showed that we are tolerant of substantial occlusions on different parts of the speaker's face as well. This suggests that audiovisual speech perception overcomes significant losses in aural clarity and facial visibility in everyday life and employs speech reading cues from across the face, which, when combined with experience-based processes, contribute to a robust and adaptable system of visual and audiovisual speech perception that can accommodate a wide range of visual needs.

2.2 Virtual Humans and Artificial Visual Stimuli

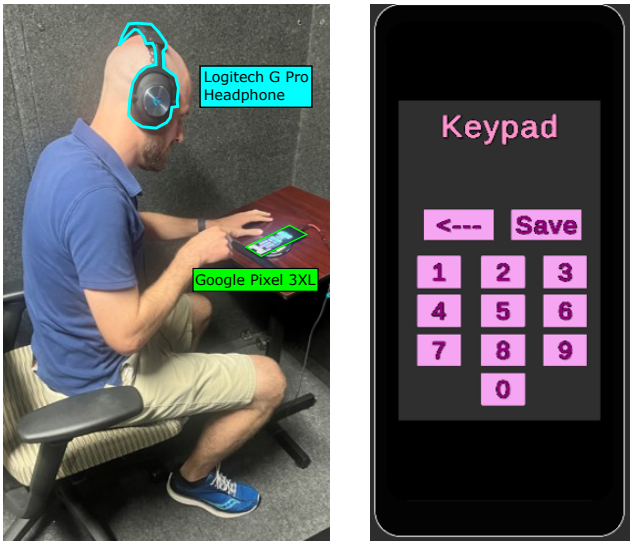
In situations with low SNR and no visual cues, we cannot benefit from the visual modality. These scenarios are fairly common, particularly when we make audio calls or wear face masks, and they may become even more common as head-mounted displays (HMDs) become more widely available and popular for social VR experiences. HMDs can share audio and deliver realistic virtual environments, but tracking of facial movements, in particular lip movements, for the virtual humans we embody is not widely available yet.

Across various domains, such as animation and gaming, different types of lip-sync tools have emerged. These include plug-and-play software, deep learning approaches [33], and hand-drawn animations. The difficulty of implementation and the accuracy of speech reading cues vary among these approaches. Plug-and-play programs are typically easier to set up, but they may offer less accurate speech reading cues than hand-drawn animations [20].

Researchers have investigated different human and non-human visualizations that could provide speech reading cues similar to lips. Thomas et al. [34] compared different face representations such as lips only, eyes and lips etc., along with different modulations with speech, such as whole face movement, oral area movement and extraoral movement. Their results showed that visual speech and the type of modulation substantially influence auditory speech recognition.

Similarly, Bernstein et al. [4] tested an oval-like shape, or a Lissajous shape, which is modulated based on the acoustic speech amplitudes. This means that the shape scaled up as the speech became louder, and scaled down as the speech became softer. This study highlighted the potential of artificially generated visual speech stimuli by showing that participants' perception of speech was improved by a straightforward graphic with an amplitude-based modulation over no visuals at all. While the oval shape does not provide much information about the speech itself, work done by Yuan et al. [37] and Vroomen et al. [36] demonstrated that the timing of the audiovisual interaction plays a crucial role, and provided enough speech reading cues to show an improvement.

In our study, we consider the same scenario where the speech signal is noisy and the visuals are not available to the listener. Broadly we attempt to compare a non-human visualization (Bernstein et al.'s oval-like shape) and a more human visualization (VH with lip sync). More specifically, in these scenarios we propose three visualizations – *oval*, *virtual human lips*, and *virtual human face* – with two modulation types – *amplitude-based* and *viseme-based*. We investigate their potential when compared to no visuals and a real person speaking. We describe our experiment in the following section.



(a) Setup (b) Keypad

Figure 1: (a) Annotated photo showing a participant completing the experiment on a Google Pixel 3XL wearing the Logitech G Pro Headphones. (b) The keypad interface participants used to input the three digits they heard in each trial.

3 EXPERIMENT

In this section, we describe the speech perception experiment we ran to look into the research questions we posed in Section 1.

3.1 Participants

We ran an a-priori power analysis using G*Power [8] application version 3.1. Based on previous work [6], we assumed an effect size of 0.35, and used default values for power and error probability of 80% and 5%. The analysis recommended a total sample size of 16 participants. We obtained participant's demographics using ACM's DIE Demographics Questionnaire [1]. We recruited 17 participants from our university community, 9 male, 7 female, and 1 non-binary. 13 were born between 1981 and 2000, while 4 participants were born after 2000. None of the participants reported any visual, motor, or cognitive disabilities. All of the participants had normal or corrected-to-normal vision and none of them reported any visual or vestibular disorders, such as color or night blindness, dyschromatopsia, or a displacement of balance. The participants were either students or non-student members of our university community who responded to open calls for participation, and received monetary compensation of $15 for their participation. The experiment took participants on average 75 minutes to complete.

3.2 Materials

The setup and stimuli we used for our study are described below.

3.2.1 Setup. Participants were seated in a quiet "whisper room" booth in our laboratory, see Figure 1a. They used the Google Pixel 3XL, which has a 6.3 inch display, a native resolution of 1440×2960 pixels, and runs on Android 12. Based on an article [1], the average smartphone display size from 2021 to 2022 was 6.3 inches, hence, we selected a handheld device with the same display size. Additionally, participants wore Logitech G Pro VR Headphones [2], characterized as full bandwidth with passive noise cancellation. To develop the audio-visual environment, we prepared an Android application in the Unity Engine version 2020.3.f1. The application was deployed onto the smartphone and the data was logged locally.

3.2.2 Audio Stimuli. In this study, we used audio stimuli similar to what Smiths et al. [29] and Krishnamurthy et al. [15] had done previously. The audio stimuli consisted of a speech signal based on triplets of digits presented simultaneously with background noise. During the trials, the decibel levels of the speech signal were varied while the background noise remained constant at 60 dBA. The speech signal and the background noise are described below.

(1) **Speech Signal**: The speech signal was a triplet of digits spoken by a female speaker in American English and digitally recorded by a professional Blue Yeti USB microphone[2]. Only monosyllabic digits (i.e., 1, 2, 3, 4, 5, 6, 8, and 9) were included in the study. These captured audio clips were then cleaned up with the Audacity audio software's noise reduction filter. We next created a look-up table with the adjustments needed for each of the eight audio clips to duplicate relative changes of +4,dB, +2,dB, and -2,dB. To ensure that these exact decibel changes were received by participants, we calibrated them using an SLM25TK Sound Level Meter[3]. It has a measurement range of 30–130 dBA, 0.1 dB resolution, and a frequency response of 31.5 Hz–8.5 KHz. The same speech recordings were used throughout the experiment.
(2) **Background Noise**: To simulate background noise, we created an 8-talker *babble*, following Krishnamurthy et al. [15] guidelines. We decided to go for multi-talker babble over other noises, such as white noise, because it can be expected in a real-life context, and would also lie within a typical environmental noise spectrum [13]. Similar to Choudhary et al [6], to create the 8-talker babble, we generated four female and four male speaker audio clips using Google Cloud's text-to-speech API[4], and superimposed them. All 8 clips played simultaneously, and at the end of the created babble clip, it was looped. This was played during the presentation of all audio-visual stimuli, and was calibrated on the headphones to the level of 60 dBA.

3.2.3 Visual Stimuli. During this study, all visual stimuli were presented in portrait mode within the 6.3 inches display of the handheld device. The background was set to black at all times to not distract participants from what was presented to them. All visual stimuli are shown in Figure 2. We used two methods to modulate the visual stimuli synchronously with the verbal signal: *amplitude-based* and *viseme-based*.

(1) **Amplitude-based:** The stimuli are modulated based on the acoustic speech amplitudes. This was obtained by using a Fourier transform to perform a spectral analysis on the audio

[1] https://community.acm.org/demographics/welcome.cfm

[2] https://www.bluemic.com/en-us/products/yeti/
[3] https://www.tekcoplus.com/products/slm25tk
[4] https://cloud.google.com/text-to-speech/

(a) Audio-Only (b) Oval (c) VH Lips (d) VH Full Face (e) Real Human

Figure 2: Visual stimuli used in the experiment conditions: (a) audio-only, (b) oval, (c) virtual human's lips, (d) virtual human's full face, and (e) real human (black bar added for de-identification, participants experienced without the black bar).

data. This helped us interpret the amplitudes by revealing the spikes in the signal.

(2) **Viseme-based:** The visual stimuli modulate based on the viseme(s) of the speech material. A viseme is the visual description of a phoneme in spoken language. Over time, these visemes are interpolated to simulate natural mouth motion.

We describe the implementations of the visual stimuli and the speech modulation below.

(1) **Oval Visualizer:** This comprised of a oval which had a fixed width of 1.5 cm, and had a minimum and maximum height of 0.5 cm and 1.5 cm respectively. Bernstein et al. [4] found that participants could extract some speech reading cues to improve their speech perception when a similar oval shape, known as the Lissajous shape in their paper, was modulated based on the speech amplitude.
As a result, for the *amplitude-based* modulation, similar to their implementation, the oval shape scales up and down along the acoustic amplitudes of the speech.
For the *viseme-based* modulation, the oval shape changes analogous to the humans mouth. So when the mouth is closed, the shape is flat. Similarly when the mouth opens, the shape size increases vertically by the same amount.

(2) **Virtual Human:** We used a hi-poly Caucasian female Unity Multipurpose Avatar (UMA)[5] with black hair and brown clothes. The avatar blinked and made slight eye saccades when she was not speaking.
For the *amplitude-based* modulation, the mouth of the virtual human opened and closed based on the acoustic speech amplitudes. We did so by preparing a custom script to modulate the blendshapes offered by the UMA virtual human.
For the *viseme-based* modulation, we used UMA's OneClick preset from the SALSA LipSync Suite [6]. This provided real-time lip-sync for our UMA virtual human.

(3) **Real Human:** We captured a Caucasian female actor pronouncing the different digits in American English in front of a Logitech 4K Brio HDR webcam and a green screen with shadow-free illumination so that the lip, jaw, and tongue movements were clearly visible. All video edits were done using the DaVinci video editing software [7].

3.3 Study Design

We performed a partial-factorial within-subjects design with the following factors and control conditions (see Figure 2).

- **Visualization (3 levels):** All visualizations were synced to the speech signal based on the modulation type.
  (1) **Oval:** Along with the speech, participants would see an oval that scaled vertically based on the modulation type.
  (2) **VH Full Face:** Participants would see the full face of the VH.
  (3) **VH Lips:** Participants would only see the VH's lips.
- **Modulation Type (2 levels):** These modulation types were implemented for the three visualizations in sync with the speech.
  (1) **Amplitude-based:** Modulation based on the acoustic speech amplitudes.
  (2) **Viseme-based:** Modulation based on human lip movements.
- **Control Conditions (2 levels):**
  (1) **Audio-Only:** In this condition, the human was not visible to participants. They only heard the speech.

[5]https://assetstore.unity.com/packages/3d/characters/uma-2-unity-multipurpose-avatar-35611

[6]https://crazyminnowstudio.com/unity-3d/lip-sync-salsa/

[7]https://www.blackmagicdesign.com/products/davinciresolve/

Table 1: Table illustrating the six experimental conditions, including the three levels of visualizations and the two levels of modulations. The two control conditions are not shown in this table.

	Modulation Type	
Visualization	*Amplitude-based*	*Viseme-based*
Oval	Oval_{Amp}	Oval_{Vis}
VH Lips	VHLips_{Amp}	VHLips_{Vis}
VH Full Face	VH_{Amp}	VH_{Vis}

(2) **Real Human:** Along with the speech, participants would see a video recording of the speaker.

In total, participants experienced 8 conditions, including 2 control conditions and 3×2 (*Visualization* × *Modulation Type*) experimental conditions. All 8 conditions were presented and randomized based on a Latin Square table.

3.4 Procedure

To participate in the experiment, the participants read through a consent form, and were asked to give their verbal consent to participate. The experimenter then described the task protocol and overall flow of the experiment to the participants. The experimenter explained to the participants that they would hear a 3-digit number and see one of the different visual stimuli (except in the audio-only condition). After hearing the 3-digit number, the participants would then be asked to repeat that number on a keypad (see Figure 1b). The experimenter explained all the visual stimuli that the participants would experience with respect to the factors that we described in Section 3.2.3. Participants then put on the Logitech headphones and started the application on the phone. The application began by welcoming the participant and asked for their unique participant ID. After this, the participants were familiarized with the task and all the visual stimuli in a practice session.

Once they completed the practice session, the experimental trails began. In each trial, they were exposed to the visual stimuli (except in the audio-only condition, in which the display was black) and the audio of a 3-digit number. After this, a keypad appeared, participants enter the 3-digit number, pressed the *"save"* button, and then the keypad disappeared. This was one trial, while a condition contained at least 24 trials. There were in total 8 conditions. When participants completed a condition, they were asked to answer the SUS (System Usability Scale) and UEQ-S (User Experience) questionnaires (see Section 3.5) on a laptop. To minimize participants' eyestrain and to help them regain focus, answering the questionnaire after every condition provided a short break away from the device. Additionally, after four conditions, we provided a 10-minute break. At the end of all 8 conditions, participants proceeded to complete a post-questionnaire, which assessed their demographics, prior VR experience, and general perception and preference of the different visual stimuli they experienced, along with their reasoning behind their answers.

3.5 Measures

We collected participants' speech perception in noise by measuring their Speech Reception Thresholds (SRT), and obtained subjective measures via questionnaires.

3.5.1 Speech Reception Threshold in Noise. To measure SRT, we followed the adaptive protocol as described by Plomp & Mimpen [21] with minor adjustments. We presented a random triplet of non-repeating monosyllabic English digits (1, 2, 3, 4, 5, 6, 8 and 9), and the participants attempted to repeat the triplet. The original triplet digit test by telephone [29] consisted of 23 trials, and later versions presented between 23 and 30 trails [35]. We presented a minimum of 25 presentations. In the test, a constant 8-talker babble noise was fixed at 60 dB and the speech level was varied. The response triplet was judged to be correct only when all digits were correctly replied.

Adaptive SRT Protocol:

(1) The first triplet is presented repeatedly, each time increasing the speech level (step size 4 dB) until the triplet is entered correctly.
(2) The speech level is decreased by 2 dB, and the second triplet is presented.
(3) Based on the user's response, the subsequent triplets are presented at a 2 dB higher level (incorrect response) or a 2 dB lower level (correct response).
(4) The SRT is calculated as the average signal-to-noise ratio of the last 10 triplets.

3.5.2 estionnaires. We used the following questionnaires to gain insights about participants' sense of the usability of the system, user experience, and subjective preferences for all the conditions.

- **System Usability Scale (SUS):** We employed Brooke et al.'s [5] system usability scale questionnaire to rate each condition's usability. Answers were given on a 1-to-5 scale to express agreement or disagreement with the questionnaire's statements, where 1 denotes strongly disagree and 5 denotes strongly agree. The sum of the contributions from each item determines the final SUS scores, which ranges from 0 to 100.
- **User Experience Questionnaire (UEQ-S):** We used the short version of the user experience questionnaire developed by Schrepp et al. [26] to assess participants' user experience with each condition. The questionnaire consists of 8 items through which scores are calculated for two dimensions of their experience, *pragmatic* and *hedonic quality*. *Pragmatic qualities* describe qualities that relate to the tasks or goals the user aims to reach when using the product, while *hedonic qualities* do not relate to tasks and goals, but describe aspects related to pleasure or fun while using the product. After every condition, participants rate each item in the questionnaire on a 7-point Likert scale (from -3 to +3). A weighted sum of their item ratings determine the *pragmatic* and *hedonic quality* scores.
- **Preferences:** We asked participants to indicate their subjective preferences among all five visual conditions (3 *visualizations* and 2 *control*) in terms of speech understanding by ranking the five visualizations from most preferred (rank of 1) to least preferred (rank of 5).

Table 2: Statistical test results for the 3 × 2 (*visualization* × *modulation type*) experimental conditions for the *SRT* measure, *Usability (SUS) ratings*, and *User Experience (UEQ-S)* ratings for *pragmatic* and *hedonic qualities*.

Measure	RM-ANOVA	Factor	df_G	df_E	F	p	η_p^2	Pairwise-Comparisons
SRT	Two-way *Visualization (3 levels)* *Modulation (2 levels)*	Visualization	1.77	28.41	9.4	0.24	0.09	N/A
		Modulation	1	16	6.6	**0.02**	0.29	N/A
		Visualization * Modulations	1.93	31.01	3.76	**0.03**	0.19	p Ÿ **0.02**: (Oval$_{Amp}$ Ÿ *Oval*$_{Vis}$), (VHLips$_{Vis}$ Ÿ *Oval*$_{Vis}$)
Usability (SUS)	Two-way *Visualization (3 levels)* *Modulation (2 levels)*	Visualization	1.88	30	0.74	0.48	0.04	N/A
		Modulation	1	16	0.29	0.59	0.01	N/A
		Visualization * Modulations	1.87	30	1.27	0.30	0.07	N/A
User Experience (UEQ-S) *Pragmatic Quality*	Two-way *Visualization (3 levels)* *Modulation (2 levels)*	Visualization	1.59	25.37	1.23	**0.04**	0.19	p Ÿ **0.05**: (*Oval*, *VH*)
		Modulation	1	16	1.23	0.28	0.007	N/A
		Visualization * Modulations	1.35	21.53	3.16	0.08	0.16	N/A
User Experience (UEQ-S) *Hedonic Quality*	Two-way *Visualization (3 levels)* *Modulation (2 levels)*	Visualization	1.63	26.05	3.33	0.06	0.17	N/A
		Modulation	1	16	0.72	0.40	0.04	N/A
		Visualization * Modulations	1.44	22.96	1.66	0.21	0.09	N/A

3.6 Hypotheses

We developed the following hypotheses based on our motivation and the research questions we posed in Section 1:

- **H1** Better speech perception for our six experimental conditions than the audio-only control condition.
- **H2** Worse speech perception for our six experimental conditions than the real human control condition.
- **H3** Among our experimental conditions, we expect the Oval$_{Amp}$, VHLips$_{Vis}$, and VH$_{Vis}$ conditions to result in better speech perception than the Oval$_{Vis}$, VHLips$_{Amp}$, and VH$_{Amp}$ conditions.

4 RESULTS

We analyzed the responses with repeated-measures analyses of variance (RM-ANOVAs) and Tukey multiple comparisons with Bonferroni correction at the 5% significance level. We confirmed the normality with Shapiro-Wilk tests at the 5% level and QQ plots. Degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity when Mauchly's test indicated that the assumption of sphericity was not supported.

4.1 Speech Reception Thresholds in Noise

4.1.1 ***Effects with Control Conditions.*** Here, we present our comparative analysis between all six artificial visual conditions and both control conditions, one at a time. Therefore, we started by analyzing the responses with one-way RM-ANOVAs with seven conditions (6 artificial and 1 control), followed by pairwise comparisons with Bonferroni correction, as shown in Figure 3.

With Audio Condition. We found a significant main effect of the conditions on the *SRTs*, $F(4.12, 65.88) = 5.60$, p Ÿ *0.001*, $\eta_p^2 = 0.26$. Post-hoc tests with Bonferroni correction showed that the following conditions paired with the *Audio-only* condition were significant: *Oval*$_{Amp}$, *VHLips*$_{Amp}$, and *VH*$_{Amp}$.

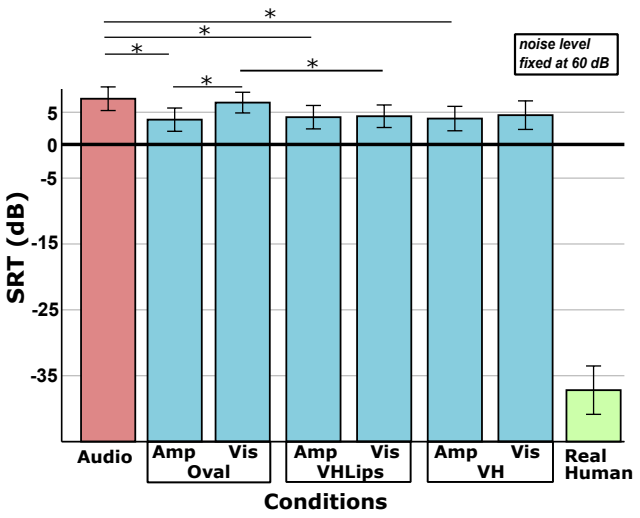


Figure 3: SRT results are shown in a bar graph with our eight experimental conditions on the x-axis. The SRT denotes the required speech level relative to the noise level (fixed at 60 dB) for participants to understand 50% of the speech stimuli (lower is better). The vertical error bars indicate the standard error. The horizontal bars and asterisks indicate statistical significance (∗ pŸ*0.05*). All pairs with the real human are significant.

Our results show that the *Audio-only* condition was significantly worse than all conditions with the *Amplitude* modulation type in terms of participants' SRTs.

With Real Human Condition. We found a significant main effect of the conditions on the *SRTs*, $F(6, 96) = 333.52$, $p \ddot{Y} 0.001$, $\eta_p^2 = 0.95$. Post-hoc tests with Bonferroni correction showed that all conditions paired with the *Real Human* condition were significant.

Our results show that the *Real Human* condition was significantly better than all conditions in terms of participants' SRTs.

4.1.2 ***Effects between Experimental Conditions.*** In this section, we compare all six of our experimental conditions (excluding the control conditions). The statistical test results of the RM-ANOVAs and pairwise comparisons are shown in Table 2 and Figure 3. We analyzed the responses with a two-way RM-ANOVA with 3 visualization levels × 2 modulation types.

We did not find a significant main effect of *visualization* on SRT. However, we did find a significant main effect of *modulation type* on SRT. We further found a significant interaction effect between *visualization* and *modulation type* on SRT.

Our results show that the abstract condition $Oval_{Amp}$ performed better than an abstract visualization with a more human-like modulation, $Oval_{Vis}$. Furthermore, participants performed better when we used a more human-like pair, $VHLips_{Vis}$, compared to when we replaced the VH with an abstract oval, $Oval_{Vis}$.

4.2 Questionnaires

4.2.1 SUS. Among the six experimental conditions, we did not find a significant main effect with *visualization* or *modulation* or an interaction effect on SUS scores, as shown in Figure 4.

4.2.2 UEQ-S. Among the six experimental conditions, we did not find a significant main effect with *visualization* or *modulation* or an interaction effect on UEQ-S scores, as shown in Figure 4.

4.2.3 Preferences. Participants ranked the *Real Human* condition as the best and the *Audio-only* condition as the worst. Among the experimental conditions, *VH Full Face* was preferred over *VHLips*, and lastly *Oval*, as shown in Figure 5.

5 DISCUSSION

Our experiment demonstrated that in the presence of noise and the absence of the speaker's face, some artificial visual speech stimuli can improve users' speech perception. In this section, we discuss the research questions and hypotheses we posited in Section 1 and Section 3.6, respectively. We go into further detail on comparisons with no visuals, with the real human, and among the artificial visual speech conditions.

Arti cial Visual Speech compared to Audio-Only. In this study, we presented three visualizations: an oval shape, virtual human lips, and a virtual human face. These representations were modulated in two different ways dependent on the speech: amplitude-based and viseme-based. Our first research question, **RQ1**, enquired if our experimental conditions improved speech perception in comparison to an empty screen, which simulates an audio call. Due to the lack of any visuals or any speech reading cues, we considered an empty screen as the worst case situation. Hence, our **Hypothesis H1** stated that our experimental conditions would offer sufficient speech reading clues to be significantly better than the audio-only condition with no visuals. However, we can only partially accept **H1** based on the results in Section 4.1.1. We found that only the experimental conditions with the *amplitude-based* modulation type ($Oval_{Amp}$, $VHLips_{Amp}$, and VH_{Amp}) improved participants' speech perception over the audio-only condition, while *viseme-based* conditions did not. This is an interesting finding, and we think there are at least two reasons for it, or a combination of them.

First, the *amplitude-based* modulation is more basic in nature and relies only on the signal's loudness, whereas the *viseme-based* modulation is more complex and naturally pairs visually with a human face. Three abstract visualizations – an oval, a virtual human's lips, and a virtual human's face – were paired with these modulation types; the oval was simple and non-human like, and the virtual human was almost cartoonish in appearance. We think the *amplitude-based* conditions performed better than the *viseme-based* conditions because they paired better with our non-human-like visualizations. Hence, when *viseme-based* modulation is paired with a more realistic virtual human, like Unreal Engine's MetaHuman [8], we can expect significantly more speech reading cues.

Second, we used a plug-and-play lip sync solution for the *viseme-based* modulations. We believe that more precise and accurate lip movements, similar to how animation studios do them, can be produced. Furthermore, other facial movements, such as tongue movements or breathing, can provide additional speech reading cues when we speak. *Viseme-based* modulations could be made more accurate and precise by combining lip movements with other facial movements. Here are examples of comments we received from two participants that illustrate how challenging it was to read the virtual human's lips.

> P16: *"The virtual human was helpful with showing the whole virtual face or a person, but the lips did not move like a real human so it was not the most beneficial."*
> P17: *"The mouth movements were difficult to memorize and map to certain numbers. Unnatural movement of the mouth made it hard."*

Arti cial Visual Speech compared to a Real Human. Our second research question, **RQ2**, compared our experimental conditions to a real human speaking in terms of audio-visual speech perception. All the speech reading cues we need come from a real human speaking, which led to our **Hypothesis H2** that all of our experimental conditions would perform poorly in comparison. We can accept **H2** based on our results in Section 4.1.1 and participants' ranking in Section 4.2.3. Participants ranked the real human the best, and their required speech levels improved from 4 dB above the noise level for experimental conditions to 36 dB below the noise level with a real human, at a noise level of 60 dB. If we consider a real human the standard for speech reading, we are seeing much room for improvement.

Arti cial Visual Speech compared to each other. Since our conditions *amplitude-based oval* and *viseme-based virtual humans* are natural pairs, our **Hypothesis H3** stated that they would be the best when comparing the conditions to each other. Based on our findings this can only be partially accepted. $Oval_{Amp}$ was better than $Oval_{Vis}$ and $VHLips_{Vis}$ was better than $Oval_{Vis}$, but nothing

[8] https://www.unrealengine.com/en-US/metahuman

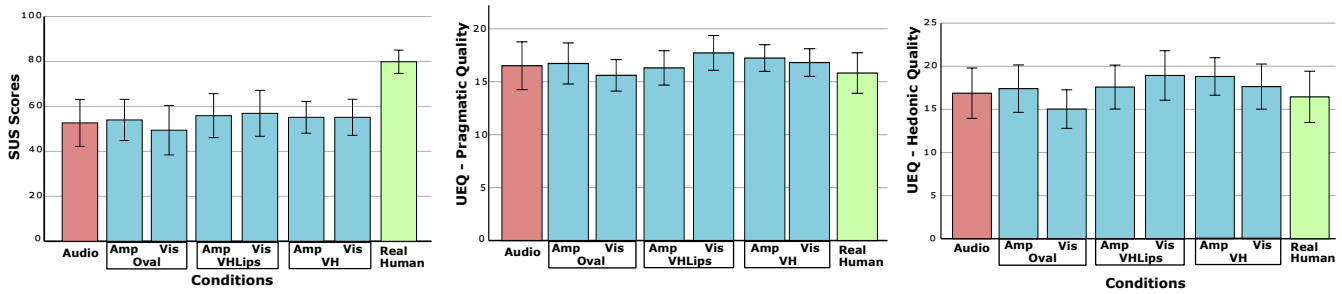


Figure 4: Bar charts showing our usability and user experience results with the x-axis showing our eight experimental conditions. The y-axes show the results for our *SUS Scores*, *UEQ-Pragmatic Quality*, and *UEQ-Hedonic Quality* (higher is better). The vertical error bars indicate the standard error.

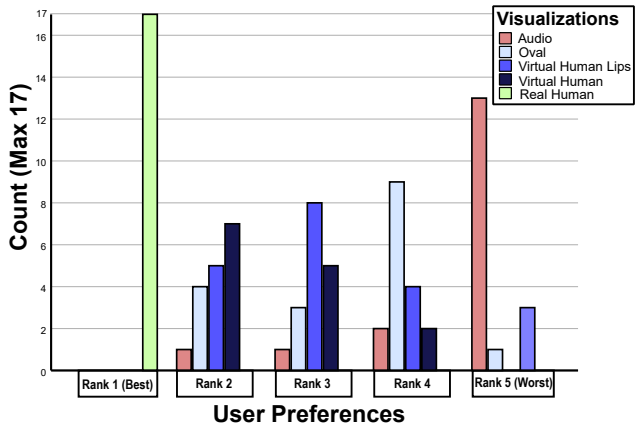


Figure 5: User Preferences, among five visualizations, are shown in a bar graph with 5 ranks on the x-axis with respect to speech understanding: Rank 1 (Best) to Rank 5 (Worst).

else. This suggests that using natural pairs results in some improvement in SRT scores, whereas using non-natural pairs, according to some participants' anecdotes, causes more confusion. For example, for non-natural pairs $Oval_{Vis}$ and VH_{Amp}, two participants stated the following:

> P2: *"Sometimes it seemed as if the lips were moving but nothing was coming out."*
> P10: *"Certain numbers looked almost identical so when I focused on the oval I got more confused."*

For the *virtual human* conditions, we believe the level of anthropomorphism plays an important role. Higher levels of virtual human anthropomorphism would provide more familiar speech reading cues, improving social presence as well [10]. It is important to note that, while visual realism and behaviors are important for social presence, for speech perception it is for necessary facial behaviors such as lip movements, eye saccades, and facial expressions to be in synchrony. Our virtual human, for example, had natural eye saccades and facial movements, which were intended to make them more believable and present, but they distracted participants from speech reading:

> P1: *"Having the whole face was nice, it felt more like actual lip reading. I think I was able to pick it up better in this condition. Maybe it was actually harder than another condition but it was the most familiar as it most closely resembled an actual person's face."*
> P3: *"The lips helped me understand numbers but still it wasn't as easy a real human. Maybe due to facial expressions not being as exact as real human."*
> P7: *"Her sporadic eye movements were distracting making it harder to focus on her lips when I couldn't hear."*

When we compare the *oval* visualization to the virtual humans, we find that they have similar SRT scores, but they are ranked the worst in terms of user preference. Reinforcing Vroomen et al.'s [36] results, we believe that the oval provides important time cues about the speech signal, which aids users in anticipating the speech signal. However, it lacks the various speech reading cues that a real human face provides, which limits its ability to be as good as a real human speaking.

> P2: *"It only helped with when to expect to hear a number (when the shape started moving). But it did not help with what the number was as its motion was very inconsistent with the digit."*
> P10: *"Difficult to lip read but simple graphic, so not overwhelming"*

Overall, our findings show that when there is a noisy audio signal but no visuals, we can use some artificially generated visual stimuli to improve speech perception. However, the techniques we currently have available still perform poorly when compared to a real human speaking. Nonetheless, in the absence of any visuals, such as an audio call these findings are useful. In situations with partial facial visibility, such as a user wearing a VR HMD or wearing a face mask, parts that are occluded (such as the lips or eyes) could be replaced by a similar looking VH. Furthermore, larger (and more space-consuming) visualizations, such as a full virtual human, have the best chance of being as good as a real human speaking. That said, when we have limited screen space available, such as on a smartphone or with a low field of view VR/AR HMD, visualizations like the oval or virtual human lips may still be helpful in improving speech perception.

6 LIMITATIONS AND FUTURE WORK

Our study showed that artificial visual stimuli can improve speech perception in the presence of noise and the absence of the speaker's face. However, there are a few limitations to the current work that can lead us to interesting research vistas that may be investigated in the future.

To begin with, we used a 2D display to simulate an audio call situation. An audio call may be the most typical scenario with respect to our daily lives where the listener does not have access to visuals of the speaker's face. Another scenario in which we have poor or no visuals of the speaker is when they are wearing a face mask or a head-mounted display, be that when multiple users try to communicate in the real world while wearing augmented reality head-mounted displays, or be it that they are meeting in a virtual environment while they are only able to see each other's virtual avatars. Because the listener does not have to use a 2D display in these scenarios, another line of research could look into 3D visualizations in a virtual or augmented reality setting to improve speech perception.

Secondly, our study used a female speaker and a similar looking virtual human. The type of speaker has been shown to affect our speech perception [30], so future work could look into different speaker types, such as race, skin color, age, and gender, as well as different levels of anthropomorphism (for virtual humans), such as photorealistic or abstracted cartoonish characters.

Last but not least, in our study we tested only one popular plug-and-play lip sync software (UMA OneClick preset from the SALSA LipSync Suite). There are alternatives, such as Oculus Lipsync [9] or RogoDigital Lipsync [10], and other lip sync approaches, such as based on deep learning [22] or hand drawings. We propose that future research should explore these lip sync solutions with respect to SRTs to understand which of them is best suited for enhancing speech perception.

7 CONCLUSION

In this paper, we presented a human-subject experiment in which we tested different artificial visual stimuli to aid human speech perception in the presence of noise and the absence of the speaker's facial visuals. We tested three levels of visualizations, *oval*, *virtual humans lips*, and *virtual humans face*, and two levels of modulation, *amplitude-based* and *viseme-based*. First, our results showed that some of our tested artificial visual stimuli proved better than an audio-only condition with no visuals. Second, while there were benefits in terms of speech perception, participants' results were considerably worse compared to seeing a real person speaking, which indicates that there is still much room for improvement. Overall, our results show that using artificial visual speech stimuli is a viable option in settings with poor speech signals and no visuals. We discussed potential explanations and implications to help practitioners who want to use these techniques, as well as the limitations of our experiment and future research directions.

[9]https://developer.oculus.com/documentation/unity/audio-ovrlipsync-unity/
[10]https://lipsync.rogodigital.com/

ACKNOWLEDGMENTS

This material includes work supported in part by the National Science Foundation under Award Numbers 2235066 and 1800961 (Dr. Ephraim P. Glinert, IIS); the Office of Naval Research under Award Numbers N00014-21-1-2578 and N00014-21-1-2882 (Dr. Peter Squire, Code 34); and the AdventHealth Endowed Chair in Healthcare Simulation (Prof. Welch).

REFERENCES

[1] 2023. Display Dynamics : Average smartphone display size stays at 6.3 inches. https://omdia.tech.informa.com/OM022757/Display-Dynamics--January-2022-Average-smartphone-display-size-stays-at-63-inches-while-the-resolution-can-be-potentially-enhanced.
[2] 2023. Logitech G-PRO VR: Headphones for Meta Quest 2. https://www.logitechg.com/en-us/products/gaming-audio/pro-gaming-headset-oculus.981-001003.html.
[3] Peter Assmann and Quentin Summerfield. 2004. The perception of speech under adverse conditions. *Speech processing in the auditory system* (2004), 231–308.
[4] Lynne E. Bernstein, Edward T. Auer, and Sumiko Takayanagi. 2004. Auditory speech detection in noise enhanced by Lipreading. *Speech Communication* 44, 1-4 (2004), 5–18. https://doi.org/10.1016/j.specom.2004.10.011
[5] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.
[6] Zubin Choudhary, Gerd Bruder, and Gregory F. Welch. 2023. Visual Facial Enhancements Can Significantly Improve Speech Perception in the Presence of Noise. *IEEE Transactions on Visualization and Computer Graphics, Special Issue on the IEEE International Symposium on Mixed and Augmented Reality (ISMAR) 2023.* (2023).
[7] Joon Hao Chuah, Andrew Robb, Casey White, Adam Wendling, Samsun Lampotang, Regis Kopper, and Benjamin Lok. 2013. Exploring agent physicality and social presence for medical team training. *Presence: Teleoperators and Virtual Environments* 22, 2 (2013), 141–170.
[8] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (2007), 175–191.
[9] Carol A Fowler and Dawn J Dekle. 1991. Listening with eye and hand: cross-modal contributions to speech perception. *Journal of experimental psychology: Human perception and performance* 17, 3 (1991), 816.
[10] Maia Garau, Mel Slater, David-Paul Pertaub, and Sharif Razzaque. 2005. The responses of people to virtual humans in an immersive virtual environment. *Presence: Teleoperators & Virtual Environments* 14, 1 (2005), 104–116.
[11] Mar Gonzalez-Franco, Antonella Maselli, Dinei Florencio, Nikolai Smolyanskiy, and Zhengyou Zhang. 2017. Concurrent talking in immersive virtual reality: on the dominance of visual speech cues. *Scientific reports* 7, 1 (2017), 3817.
[12] Jonathan Gratch, Jeff Rickel, Elisabeth André, Justine Cassell, Eric Petajan, and Norman Badler. 2002. Creating interactive virtual humans: Some assembly required. *IEEE Intelligent systems* 17, 4 (2002), 54–63.
[13] Sabine Hochmuth, Birger Kollmeier, Thomas Brand, and Tim Jürgens. 2015. Influence of noise type on speech reception thresholds across four languages measured with matrix sentence tests. *International journal of audiology* 54, sup2 (2015), 62–70.
[14] Timothy R Jordan and Sharon M Thomas. 2011. When half a face is as good as a whole: Effects of simple substantial occlusion on visual and audiovisual speech perception. *Attention, Perception, & Psychophysics* 73 (2011), 2270–2285.
[15] Nitish Krishnamurthy and John HL Hansen. 2009. Babble noise: modeling, analysis, and applications. *IEEE transactions on audio, speech, and language processing* 17, 7 (2009), 1394–1407.
[16] John MacDonald and Harry McGurk. 1978. Visual influences on speech perception processes. *Perception & psychophysics* 24, 3 (1978), 253–257.
[17] Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264, 5588 (1976), 746–748.
[18] Michael Meehan, Brent Insko, Mary Whitton, and Frederick P Brooks Jr. 2002. Physiological measures of presence in stressful virtual environments. *Acm transactions on graphics (tog)* 21, 3 (2002), 645–652.
[19] Kristine L Nowak and Frank Biocca. 2003. The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators & Virtual Environments* 12, 5 (2003), 481–494.
[20] Rick Parent, Scott King, and Osamu Fujimura. 2002. Issues with lip sync animation: can you read my lips?. In *Proceedings of Computer Animation 2002 (CA 2002).* IEEE, 3–10.
[21] Reinier Plomp and AM Mimpen. 1979. Improving the reliability of testing the speech reception threshold for sentences. *Audiology* 18, 1 (1979), 43–52.

[22] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 2020. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*. 484–492.
[23] Lawrence Rosenblum. 2019. Audiovisual speech perception and the McGurk effect. *Oxford Research Encyclopedia, Linguistics* (2019).
[24] Lawrence D Rosenblum, Deborah A Yakel, and Kerry P Green. 2000. Face and mouth inversion effects on visual and audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance* 26, 2 (2000), 806.
[25] Mikko Sams, Riikka Möttönen, and Toni Sihvonen. 2005. Seeing and hearing others and oneself talk. *Cognitive Brain Research* 23, 2-3 (2005), 429–435.
[26] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2017. Design and evaluation of a short version of the user experience questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence, 4 (6), 103-108.* (2017).
[27] Ralph Schroeder. 2002. Copresence and interaction in virtual environments: An overview of the range of issues. In *Presence 2002: Fifth international workshop*. Citeseer, 274–295.
[28] Mel Slater. 2009. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 1535 (2009), 3549–3557.
[29] Cas Smits, Theo S Kapteyn, and Tammo Houtgast. 2004. Development and validation of an automatic speech-in-noise screening test by telephone. *International journal of audiology* 43, 1 (2004), 15–28.
[30] Elizabeth A Strand. 1999. Uncovering the role of gender stereotypes in speech perception. *Journal of language and social psychology* 18, 1 (1999), 86–100.
[31] William H Sumby and Irwin Pollack. 1954. Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america* 26, 2 (1954), 212–215.
[32] Quentin Summerfield. 1992. Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 335, 1273 (1992), 71–78.
[33] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)* 36, 4 (2017), 1–13.
[34] Sharon M Thomas and Timothy R Jordan. 2004. Contributions of oral and extraoral facial movement to visual and audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance* 30, 5 (2004), 873.
[35] Elien Van den Borre, Sam Denys, Astrid van Wieringen, and Jan Wouters. 2021. The digit triplet test: a scoping review. *International journal of audiology* 60, 12 (2021), 946–963.
[36] Jean Vroomen and Jeroen J Stekelenburg. 2010. Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *Journal of cognitive neuroscience* 22, 7 (2010), 1583–1596.
[37] Yi Yuan, Yasneli Lleo, Rebecca Daniel, Alexandra White, and Yonghee Oh. 2021. The impact of temporally coherent visual cues on speech perception in Complex auditory environments. *Frontiers in neuroscience* 15 (2021), 678029.

Received 20 June 2023; revised XX March 2023; accepted XX June 2023