Visual Facial Enhancements Can Significantly Improve Speech Perception in the Presence of Noise

Zubin Choudhary* Gerd Bruder† Gregory F. Welch‡

University of Central Florida

Abstract—Human speech perception is generally optimal in quiet environments, however it becomes more difficult and error prone in the presence of noise, such as other humans speaking nearby or ambient noise. In such situations, human speech perception is improved by *speech reading*, i.e., watching the movements of a speaker's mouth and face, either consciously as done by people with hearing loss or subconsciously by other humans. While previous work focused largely on speech perception of two-dimensional videos of faces, there is a gap in the research field focusing on facial features as seen in head-mounted displays, including the impacts of display resolution, and the effectiveness of visually enhancing a virtual human face on speech perception in the presence of noise. In this paper, we present a comparative user study ($N = 21$) in which we investigated an audio-only condition compared to two levels of head-mounted display resolution (1832×1920 or 916×960 pixels per eye) and two levels of the native or visually enhanced appearance of a virtual human, the latter consisting of an up-scaled facial representation and simulated lipstick (lip coloring) added to increase contrast. To understand effects on speech perception in noise, we measured participants' *speech reception thresholds* (SRTs) for each audio-visual stimulus condition. 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. First, we show that the display resolution significantly affected participants' ability to perceive the speech signal in noise, which has practical implications for the field, especially in social virtual environments. Second, we show that our visual enhancement method was able to compensate for limited display resolution and was generally preferred by participants. Specifically, our participants indicated that they benefited from the head scaling more than the added facial contrast from the simulated lipstick. We discuss relationships, implications, and guidelines for applications that aim to leverage such enhancements.

Index Terms—Speech perception, background noise, hearing, human faces, enhancement methods, user study.

1 Introduction

Humans often interact in noisy environments. Such ambient noise can mask important speech signals, making it difficult to perceive and understand them. In practice, humans rarely rely on a single sensory modality—they combine the information gathered from different modalities to form their speech percept. So while speech perception may seem like an auditory perception, it is usually strongly modulated by other modalities, particularly vision [28, 38]. Lip reading is well known to enhance auditory speech perception in noisy environments [16, 38, 41, 48]. People who are hard of hearing or deaf depend on lip reading and other signals to communicate with each other [61]. However, the visual modality can lead to cross-modal interaction effects in speech perception. For example, the ventriloquist effect, a speech sound can be misperceived as dislocated towards the visual source [7]. Another similar audiovisual speech phenomenon is the McGurk effect [39]. This effect occurs when the visual signal of one phoneme is dubbed onto the audio signal of another. With some audiovisual pairs, observers do not notice the intermodal mismatch, and often perceive a phoneme that is different from the audio phoneme. For example, the audiovisual pair of /ba/ and /ga/ can produce a hearing perception of /da/. In other words, lip reading can either aid or impair human speech perception depending on the characteristics of the visual component.

To study the visual modality of speech perception, researchers have traditionally used two-dimensional displays showing videos of people speaking. However, the advances of extended reality (XR) technologies, including virtual reality (VR) and augmented reality (AR), introduce new opportunities for experimenting with and even enhancing speech perception, especially with human representations [43]. In VR, particularly, we can embody and interact with a myriad variations of virtual humans. In AR, head-mounted displays (HMDs) can use their external cameras to track, segment, and manipulate other humans in real time [15]. These variations can be aimed at facilitating effective communication by leveraging different types of embodied human signals and cues [29, 37, 44]. For example, in AR and VR, over large distances, human heads can be scaled up to improve the exchange of facial interpersonal cues, such as information about facial expressions and eye gaze [12]. Despite the fact that technology creates new opportunities, there are still many challenges [33]. For instance, the limited resolution of displays can limit our ability to lip read, and the representations of virtual humans can further limit our speech perception. These circumstances can pose communication challenges, particularly in realistically noisy virtual settings such as crowded virtual public settings or virtual restaurants.

In this paper, to improve human speech perception in noise, we test the impacts of display resolution and propose one facial enhancement, which consists of adding a simulated lip stick and increasing the scale of the head. Increasing contrast should assist participants in perceiving the lip movements, while scaling heads and higher resolution should provide more details to participants about the facial and lip movements.

This led to the following research questions:

- **RQ1**: What effects does increasing *lip contrast* and *scaling heads* together have on speech perception in noise?
- **RQ2**: What effect does *display resolution* have on speech perception in noise?

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

Here we discuss related work on speech perception in the presence of noise and recent findings on scale manipulations of virtual humans,

*e-mail: zubinchoudhary@ucf.edu
†e-mail: bruder@ucf.edu
‡e-mail: welch@ucf.edu

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

motivated by the opportunity to enhance facial perception in virtual environments.

2.1 Audio-Visual Speech Perception

Over the past several decades, researchers have been extensively studying human speech perception, to understand how human listeners recognize speech and use it to understand spoken language. Until the 1970s, it had been traditionally assumed that speech perception in face-to-face contexts is a uni-modal (i.e., auditory) process and that the role of vision is independent and additive. Seminal work on the McGurk Effect [39] and by MacDonald et al. [38], provided initial evidence about the multi-modal nature of human speech perception. Similarly, in brain imaging research, it was once thought that brain regions sensitive to auditory speech (primary auditory cortex, auditory brain stem), but now known to respond to visual speech input as well [10, 42].

Nowadays, it is a known fact that visual speech or lip reading is used by all perceivers and readily integrates with auditory speech [20, 47, 55]. Virtually any time we are speaking, we take advantage of the visual information from seeing the movement of their teeth, lips, tongue, and non-mouth facial features, and we have likely been doing so all our lives. In fact, research suggests that speech perception is inherently mulitmodal [49]. Work done by Fowler et al. [18] and Sam et al. [50] shows that speech can also be *felt*, accessed either through a speaker's jaw, lips, and neck or through kinesthetic feedback from one's own speech movements.

In the realm of audiovisual speech perception, humans interact under an enormous range of auditory and visual conditions. Some conditions may aid communication, while others may be more adverse, such as background noise or poor lighting conditions. The demands on speech communication are great, especially under such circumstances, but nonetheless listeners tend to adapt well [5]. To further understand the influence of such conditions, researchers have leveraged different display technologies to simulate them. For example, Jordan et al. [25], investigated the effects of image size (100%, 10%, 5%, and 2.5%) of the talking face. They point out that the reduction in image size substantially reduces the amount of facial information available to the observer, due to limits in human visual acuity. Similarly, Alsius et al. [4] looked into how well people could perceive audiovisual speech depending on how well people could resolve fine facial details. Their findings indicate that it can be impacted by individual variance because only some users' audiovisual speech perception was negatively impacted by reduced visual details. Other studies investigated additional relevant factors such as talker-to-listener distance [9, 24], viewing angles [26], facial brightness [27], contrast and blur [32], and many more.

However, in prior research, the usage of newer display technologies, especially VR/AR displays, to represent humans was less common. Immersive environments and virtual humans have been used as a methodological tool to test several social situations [8], and make users feel *present* with other non co-located humans [52]. While this technology provides us with this unique opportunity, it comes with its own challenges, such as limited display resolution. In this paper, we prepare a VR experiment and investigated the effects of two resolutions (1832×1920 and 916×960 pixels per eye) on speech perception.

2.2 Face Perception and Enhancement

Faces provide key visual information that we use to discriminate between one person and another every single day. In a split-second glimpse of a person's face, we can learn about their identity, emotional state, and direction of attention [14, 58].

Our facial appearances can modify the way others perceive us, for example, cosmetic makeup is traditionally used to modify the visual perception of a person's facial beauty [22]. This further elicits higher ratings of physical attractiveness and influences the processing of specific facial features [56]. Lipstick is a common cosmetic product that can dramatically alter the color and lightness/contrast of the lips [22, 23]. Specifically, using red lipstick has stronger effects than other colors. Kobayashi et al. [30] demonstrated that redder lips lightened and darker lips darkened the perceived complexion. Tanaka et al. [57] used EEG results to indicate that at later stages of face-processing, the higher attractiveness of red lips is associated with slower and more careful processing. On the other hand, blue lips, which have a low attractiveness score, are processed quickly and carelessly. Another experiment by Lander et al. [35] measured users' speechreading performance when the speaker's lips were natural, concealed, or brightly colored (with lipstick or concealer). The findings demonstrated that, in comparison to real lips, lips that were concealed and colored improved speechreading.

Similar to makeup, our facial appearances can be altered dynamically and presented to us via different display technologies as well. For example, AR face filters can alter the way we look and can change the perception of ourselves [19]. Users of applications, like SnapChat or Instagram, can engage in real-time face distortion (e.g., increasing the scale of head) and feature addition (e.g., adding red lipstick). Some recent work done by Choudhary et al. [11–13, 15] demonstrated, in AR and VR, that human head scaling/magnification can significantly improve facial cue detection and recognition. Furthermore, they placed the human at distances where one could not perceive the facial cues any more. However, by up-scaling the human heads within meaningful ranges, they were able to recover the lost facial cues.

As described earlier (refer Section 2.1), the face is crucial in human audiovisual speech perception. In this paper, we investigate the effects of a facial visual enhancement on speech perception. We enhance the face by combining two facial manipulations: doubling the head scale and increasing the lip contrast. Increasing the head scale should have the effect of providing more facial details to the listener, while higher lip contrast (due to the red lipstick) should help resolve lip movements better, especially if the resolution of the display is limited. In the following section, we describe the VR simulation and human-subject study in detail.

3 Experiment

In this section we describe the speech perception experiment we conducted to investigate the research questions stated in Section 1.

3.1 Participants

We recruited 21 participants from our university community, 15 male and 7 female, ages between 19 and 37, $M = 25.7$, $SD = 5.5$. All of the participants had normal or corrected-to-normal vision, 2 wore glasses and 4 wore contact lenses during the experiment. None of the participants reported any visual or vestibular disorders, such as color or night blindness, dyschromatopsia, or a displacement of balance. 20 participants had used a VR HMD before, and 14 of them had prior experience with social VR. The participants were either students or non-student members of our university community who responded to open calls for participation, and received monetary compensation for their participation. The experiment took participants on average 50 minutes to complete.

3.2 Material

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

3.2.1 Setup

As shown in Figure 1, participants were seated within a quiet "whisper room" booth in our laboratory. They were instructed to wear a Meta Quest 2 HMD [3], which provides a field of view of up to 96 degrees, and a native resolution of 1832×1920 per eye at a refresh rate of 120 Hz. The HMD uses an inside-out tracking system, which included a tracked controller that participants held in their dominant hand and used for input during the experiment. Additionally, participants wore Logitech G Pro VR Headphones [2], characterized as full bandwidth with passive noise cancellation. All rendering was done directly on the HMD. We developed the audio-visual environment we used in this experiment in the Unity Engine version 2020.3.f1.

3.2.2 Audio Stimuli

For this study, we presented participants with audio stimuli in line with previous work by Smiths et al. [53] and Krishnamurthy et al. [34]. The audio stimuli consisted of a speech signal that was built around triplets of digits that were presented at the same time as background noise. The

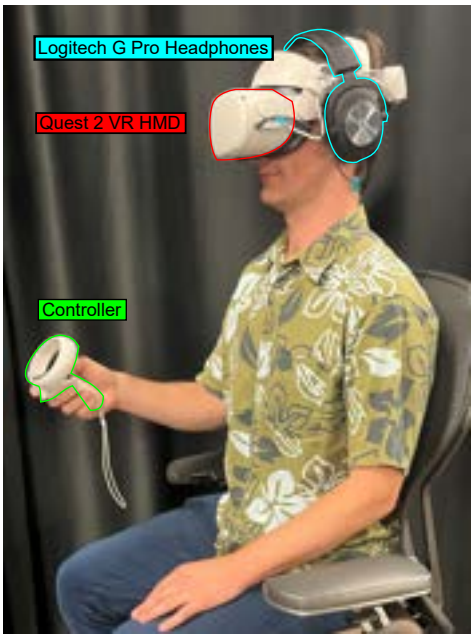


Fig. 1. Annotated photo showing a participant completing the experiment, wearing a Meta Quest 2 HMD and Logitech G Pro VR Headphones, while holding the controller in their dominant hand.

decibel levels of the speech signal were varied during the trials while that of the background noise remained fixed at 60 dBA.

Speech Signal. The speech signal consisted of a triplet of digits that were uttered in American English by a female speaker and digitally recorded by a professional Blue Yeti USB microphone[1]. For the study, only monosyllabic digits were used (i.e., 1, 2, 3, 4, 5, 6, 8, and 9). These recorded audio clips were then cleaned up using a noise reduction filter using the Audacity audio software [1]. We then generated a look-up table containing the changes required for each 8 audio clips to reproduce 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[2]. It has a measurement range of 30–130 dBA, 0.1 dB resolution, and a frequency response of 31.5 Hz–8.5 KHz.

Background Noise. As suggested by Krishnamurthy et al. [34], the background noise we used in our experiment was created using an 8-talker *babble*. We generated four male and four female speaker audio clips using Google Cloud's text-to-speech API[3], and superimposed them to create the 8-talker babble, which is a form of energetic masking [46]. All 8 clips played simultaneously, and at the end of the created babble clip, it was looped. This background noise was calibrated on the headphones to the level of 60 dBA.

3.2.3 Visual Stimuli

During this study, participants were immersed in a simulated hallway with dimensions 5 m (width) × 3 m (height) × 15 m (length). The environment replicated our lab premises, and was designed to not distract participants from the virtual human that was presented in front of them. The five visual stimuli with respect to the virtual human are shown in Figure 2.

Virtual Human. For the purpose of our study we required accurate facial features unaffected by the often serious limitations of current-state facial tracking or lip movement synthesis [31]. To accomplish this, we decided to use a hybrid approach inspired by movie productions, where we first recorded a Caucasian female actor and then applied and overlaid the captured facial details over a virtual human's face. The recordings captured life-like facial and lip movements while the end result was a 3D virtual human that provided a realistic scenario for participants to experience. We captured the actor pronouncing the different digits 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. A mask of the facial region was created using the DaVinci video editing software[4]. A video of the mask was then imported into Unity and using their projector component, we overlaid the mask over the facial region of the virtual human. The base virtual human was selected from the RocketBox Avatar Library [21], aiming for physical similarity to our actor. To further blend the skin colors of the video and the virtual human, we added point lights in the scene and made slight adjustments to the virtual human's skin texture.

We placed this virtual human at the fixed distance of 3 meters from the participants in the virtual environment. In Choudhary et al.'s work [13], their results indicated that starting at a threshold distance of 3.7 meters, participants were unable to recognize a virtual human's facial expressions, while at 3 meters they were still able to do so but at a reduced accuracy.

Display Resolution. In this study, we compared the effects of two display resolutions on participants' ability to perceive speech (see Figure 2). In the *high* display resolution conditions, we used the native resolution of 1832×1920 per eye of the Meta Quest 2 HMD, while in the *low* display resolution conditions, we reduced this to 916×960 per eye by halving the horizontal and vertical resolutions.

Enhancement Method. We further compared speech perception with the native appearance of the virtual human with that of an enhanced appearance (see Figure 2). For the latter, we combined two promising approaches. The first approach consisted of **facial scaling**, in which the head scale was doubled. Choudhary et al.'s [13] results indicate that, in similar experimental settings, participants required the head scale to be about twice the normal scale so that they read a virtual human's facial expressions optimally.

The second approach aimed to enhance the **facial contrast** by adding red lipstick to the human via the DaVinci video editing software. To produce redder lips, we first created a mask that tracked the lips. We used the Color Balance Function, and increased the red value to the maximum in the mask. We measured the CIE 1931 color values with an Urceri MT-912 light meter[5] of the skin surrounding the lips, lips without red lipstick, and lips with added red lipstick. We calculated the luminance difference or contrast between the lips and skin surrounding the lips using Michelson's equation [40]:

$$C = |I_1 - I_2| / (I_1 + I_2) \quad (1)$$

I_i: The illuminance from the observer's position of the lighting at point i, where $i = 1$ means illuminance of surrounding skin, and $i = 2$ mean illuminance of the lips (with or without simulated red lipstick). We find the contrast for the red lips to be 0.159 and for the normal lips to be 0.042, which indicates a contrast increase by 3.8 times.

3.3 Methods

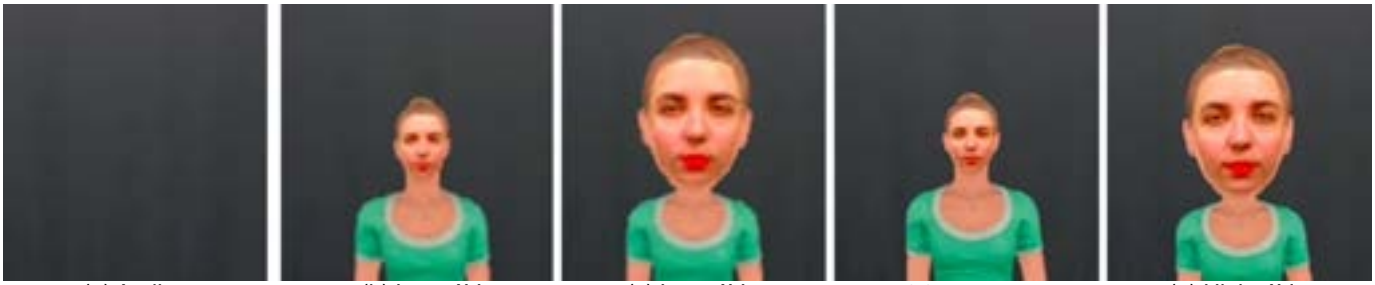
To answer our research questions (see Section 1), we decided on the following methods.

3.3.1 Study Design

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

- **Non-Visual Control Condition (1 level):**
  1. **Audio:** In this condition, the virtual human was not visible to participants. They only heard the speech.
- **Display Resolution (2 levels):**

[1] https://www.bluemic.com/en-us/products/yeti/
[2] https://www.tekcoplus.com/products/slm25tk
[3] https://cloud.google.com/text-to-speech/
[4] https://www.blackmagicdesign.com/products/davinciresolve/
[5] https://www.urceri.com/mt-912-light-meter.html



(a) Audio (b) Low_AV (c) Low_AV$_E$ (d) High_AV (e) High_AV$_E$

Fig. 2. Visual stimuli used in the five conditions of the experiment: (a) audio-only, (b) 916×960 resolution with no facial enhancement, (c) 916×960 resolution with facial enhancement, (d) 1832×1920 resolution with no facial enhancement, and (e) 1832×1920 resolution with facial enhancement.

1. **High Resolution:** The virtual environment was presented with a resolution of 1832×1920 pixels per eye, which is the full resolution of the Meta Quest 2 HMD we used.
2. **Low Resolution:** The display resolution was reduced to 916×960 pixels per eye, which is half the maximal vertical and horizontal resolution of the HMD.

- **Facial Enhancement (2 levels):**

1. **Audio + Visuals (*AV*)**: Participants received the audio stimuli and additionally saw the virtual human's speech. The virtual human's face was presented without any enhancement.
2. **Audio + Enhanced Visuals (*AV$_E$*)**: In this condition, the virtual human was presented with a facial enhancement consisting of two changes: first, the head scale was doubled, and second, we increased the contrast of the virtual human's lips compared to the rest of the face.

Hence, in total, the participants experienced five conditions (see Table 1). All five conditions were presented and randomized based on a Latin Square table.

Table 1. Table illustrating the five conditions we tested in this experiment, including the Audio-only condition, the two levels of display resolution, and the two levels of facial enhancement.

		Facial Enhancement	
Resolution	*Audio-Only*	*Audio + Visual (AV)*	*Audio + Enhanced Visuals (AV$_E$)*
Low	X	Low_AV	Low_AV$_E$
High	X	High_AV	High_AV$_E$

3.3.2 Procedure

Upon arrival, participants read through a consent form, and were asked to give their verbal consent to participate in the experiment.

The experimenter then described the task protocol to the participants, in which they would be see and hear a 3-digit number by a virtual human The participants were required to repeat that number on a keypad and answer how confident they felt if their answer was correct (see Figure 3). The experimenter explained all the conditions that the participants would experience with respect to the factors, which we described in Section 3.3.1. Participants then donned the Meta Quest 2 with the right controller and the Logitech headphones, and began the application in the headset. The application begins by welcoming the participant and then asking for their unique participant ID. Based on their ID, the order of the conditions were decided. Shortly after, there was a practice session in which they were exposed to all the conditions, and practiced the protocol a few times. This was done so that the participants were familiar with the experiment task and conditions.

Once they completed the practice session, the experimental trials begins. They ran 5 conditions, and each condition had at least 24 trials. Each trial consisted of the virtual human speaking a 3-digit number, and the participant responding to the keypad and confidence user interface. When the participant completed the first three conditions, they were provided a 2 minute break to minimize the effects of the HMD on participants' eyestrain or potential simulator sickness. After completing all conditions, they proceeded to complete a post-questionnaire, assessing their demographics and prior VR experience, and we asked their general perception and preference of the virtual human conditions as well as the reasoning behind their answers. Finally, the experiment ended with a monetary compensation.

3.4 Measures

In this section, we describe the measures of the experiment to understand the impact of facial enhancement on a user's speech perception.

3.4.1 Speech Reception Thresholds in Noise (SRT$_N$) and Confidence Levels:

To measure SRT$_N$, we followed the adaptive protocol as described by Plomp & Mimpen [45] with minor adjustments. We present a random triplet of non-repeating monosyllabic English digits (1, 2, 3, 4, 5, 6, 8 and 9), and the participants attempt to repeat the triplet. The original triplet digit test by telephone [53] consists of 23 trials, and later versions have presented between 23 and 30 trails [59]. 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 Speech Reception Threshold in Noise Protocol:

1. The first triplet is presented repeatedly, 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$_n$ is calculated as the average signal-to-noise ratio of last 10 triplets.

Additionally after every triplet presentation, we ask participants about how confident they were with their response being correct (see Figure 3(b)). They responded with low or high confidence.

3.4.2 Post Experiment Questions:

After the experiment, we asked the participants' the following questions and asked their reasoning behind their answers:

Q *In terms of speech understanding, how difficult was each condition on a likert scale from 1 (very hard) to 7 (very easy).*

Q *In the conditions with the facial enhancement, which manipulation helped you the most, head scaling or lip contrast or both equally?*

3.5 Hypotheses

The general hypotheses we established were (better speech perception corresponds to lower SRT_n's):

- **H1** Better speech perception and higher confidence levels for AV_E conditions, then *AV* conditions, and worst for *A* conditions.
- **H2** Better speech perception and higher confidence levels for higher resolution conditions than lower resolution conditions.
- **H3** Visual enhancements can compensate lower resolutions with respect to speech perception and confidence.

Regarding the visual enhancement, Hypothesis **H1**, we expect the contrast enhancement to help emphasize the color differences of the lips and the skin, while increasing the head scale should provide more details to the lips and its movements. With respect to Hypothesis **H2**, higher resolution provides more fine grain details of the lips and its movement. Additionally, from Hypothesis **H3**, we expect the speech perception loss due to low resolution, to be compensated by the visual enhancements. We believe so because despite having a larger head in low resolution, the number of pixels dedicated to the head should be similar to a higher resolution display.

(a) Keypad UI (b) Confidence UI

Fig. 3. Two user interfaces (UIs) in the experiment for participants to indicate their responses after hearing each speech: (a) a keypad used to input the three digits they just heard, and (b) an interface where participants indicate how confident they feel that the three digits they just entered on the keypad are correct.

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.

Sanity Checks. We performed two sanity checks to confirm the validity of our data. First, we ran a Pearson product-moment correlation to determine the relationship between participants' objective SRT scores and their subjective confidence in their performance. We found a strong, statistically significant correlation ($r = -0.623$, $n = 105$, $\boldsymbol{p<0.001}$). Second, we compared individual audio-only SRT scores to normal hearing thresholds (ranging between -10 to 25dB), which showed that all participants were within normal human hearing ranges .

4.1 Effects Between All Conditions

In this section, we present our comparative analysis between all five conditions in this experiment. The measures we considered are the SRTs, confidence ratings, and performance scores.

Figure 4 shows participants' measured (a) SRTs, (b) confidence ratings, and (c) performance scores. We analyzed the results for effects of the five conditions on the three measures. The statistical test results of the one-way RM-ANOVAs and pairwise comparisons are shown in Table 2. We discuss these findings in Section 5.

SRTs. Our results show that the *Audio* condition was significantly worse than all other conditions in terms of participants' SRTs. Moreover, our results show that the three conditions Low_AV_E, *High_AV*, and $High\_AV_E$ were significantly better than the *Low_AV* condition. We found no significant differences between the Low_AV_E, *High_AV*, and $High\_AV_E$ conditions.

Confidence. Our results show that participants' confidence in their responses in the *Audio* condition were significantly worse than all other conditions. Moreover, our results show that confidence ratings in the Low_AV_E and $High\_AV_E$ conditions were significantly higher than in the *Low_AV* condition.

Self-Assessed Performance. Our results show that participants subjectively estimated their performance to be highest in the $High\_AV_E$ condition, followed by *High_AV*, Low_AV_E, *Low_AV*, and finally *Audio*, which was the subjectively rated worst condition.

4.2 Effects of Display Resolution

To make comparisons between the two levels of display resolution, we analyzed the responses with a 2 (resolution) $\times$ 2 (enhancement) two-way RM-ANOVA. The results are shown in Table 2.

SRTs. Our results show a significant main effect of display resolution on the SRTs, indicating that the higher resolution resulted in significantly better SRTs than the lower resolution.

Confidence. Our results show a significant main effect of display resolution on participants' confidence in their responses, indicating that the higher resolution resulted in significantly higher confidence than the lower resolution.

Self-Assessed Performance. Our results show a show a significant main effect of display resolution on participants' self-assessed performance scores, which indicates that they were aware they performed better under higher resolution conditions.

4.3 Effects of Facial Enhancement

We further compared the two levels of facial enhancement with a 2 (resolution) $\times$ 2 (enhancement) two-way RM-ANOVA. The results are shown in Table 2.

SRTs. Our results show a significant main effect of facial enhancement on the SRTs, indicating that the facial enhancement resulted in significantly better SRTs than the native facial appearance.

Confidence. Our results show a significant main effect of facial enhancement on participants' confidence in their responses, indicating that the facial enhancement resulted in significantly higher confidence than the native facial appearance.

Self-Assessed Performance. Our results show a significant main effect of facial enhancement on participants' self-assessed performance scores, which indicates that they were aware they performed better with the facial enhancement.

4.4 Interaction Effects between Display Resolution and Facial Enhancement

We present the interaction effects from the 2 (resolution) $\times$ 2 (enhancement) two-way RM-ANOVA. The results are shown in Table 2.

SRTs. Our results show a significant interaction effect between resolution and facial enhancement on the SRTs. We found the same significant pairs from the one-way RM ANOVA (see Section 4.1).

Confidence. Our results show a significant interaction effect between resolution and facial enhancement on their confidence in their responses. We found the same significant pairs from the one-way RM ANOVA (see Section 4.1).

Self-Assessed Performance. Our results did not show a significant interaction effect between resolution and facial enhancement on the participants' self-assessed performance scores..

Table 2. Statistical test results for the SRT measure, confidence ratings, and performance scores.

Measure	RM-ANOVA	Factor	df_G	df_E	F	p	η_p^2	Pairwise Comparisons
SRTs	One-way	Condition	2.5	49.1	50.0	$<$**0.001**	0.71	All $\boldsymbol{p}<$**0.05**, except (Low_AV$_E$, High_AV), (Low_AV$_E$, High_AV$_E$), (High_AV, High_AV$_E$)
	Two-way	Resolution	1	20	61.7	$<$**0.001**	0.76	N/A
		Enhancement	1	20	23.1	$<$**0.001**	0.54	N/A
		Resolution * Enhancement	1	20	11.1	**0.003**	0.35	Following are $\boldsymbol{p}<$**0.05**, (Low_AV, Low_AV$_E$), (Low_AV, High_AV)
Confidence	One-way	Condition	2.3	45.8	23.7	$<$**0.001**	0.54	All $\boldsymbol{p}<$**0.05**, except (Low_AV, High_AV), (Low_AV$_E$, High_AV), (Low_AV$_E$, High_AV$_E$), (High_AV, High_AV$_E$)
	Two-way	Resolution	1	20	4.8	**0.039**	0.19	N/A
		Enhancement	1	20	11.1	**0.003**	0.36	N/A
		Resolution * Enhancement	1	20	6.9	**0.016**	0.26	Following are $\boldsymbol{p}<$**0.05**, (Low_AV, Low_AV$_E$), (Low_AV, High_AV)
Performance	One-way	Condition	1.8	35.7	66.9	$<$**0.001**	0.77	All $\boldsymbol{p}<$**0.001**
	Two-way	Resolution	1	20	41	$<$**0.001**	0.67	N/A
		Enhancement	1	20	52	$<$**0.001**	0.72	N/A
		Resolution * Enhancement	1	20	4.09	0.056	0.14	N/A

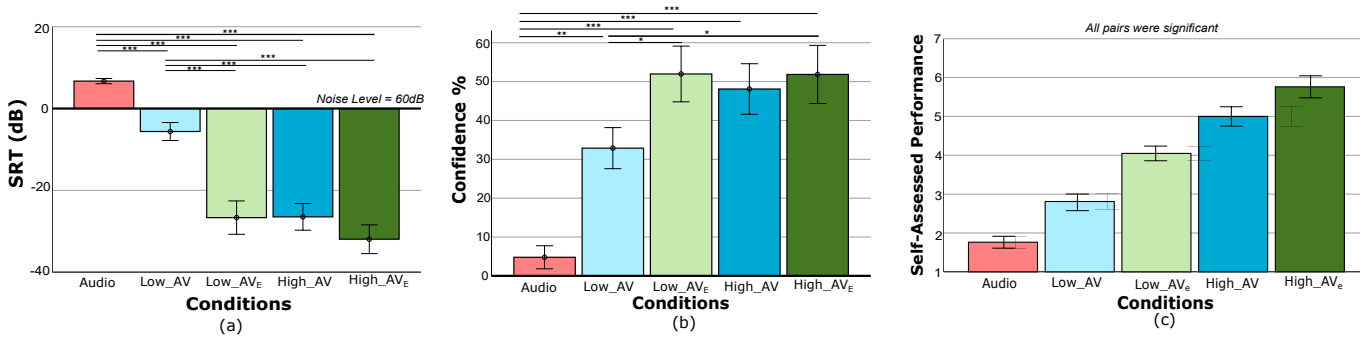


Fig. 4. Bar charts showing our experimental results. The x-axes show our five conditions. The y-axes show the results for our three measures: (a) Speech level relative to the noise level (fixed at 60 dB) for participants to understand 50% of the speech stimuli (lower is better); (b) Confidence %; (c) Self-assessed performance scores on a scale from 1=worst performance to 7=best performance. The vertical error bars indicate the standard error. The horizontal bars and asterisks indicate statistical significance ($*$ $p<0.05$; $**$ $p<0.01$; $***$ $p<0.001$); in (c) all pairs were significant, so we did not explicitly add such horizontal bars.

4.5 Differences Between Participant Groups

From looking at participants' corresponding SRTs, we observed that our participant sample was roughly split in half in terms of their perception of the speech stimuli. These groupings are visually most evident for the Low_AV$_E$ condition, and we labeled them into two groups, A and B, which are colored yellow and brown, respectively, in the scatter plot shown in Figure 6. To better understand the differences between these two apparent participant groups, we made the decision to follow an exploratory analysis approach. Group A consisted of 10 participants (8 male, 2 female; ages 18 to 31), while Group B consisted of 11 participants (7 male, 5 female; ages 19 to 39). While both groups seemed to benefit from higher resolution and the facial enhancement, it surprised us to see that participant Group B apparently gained significantly more from these factors compared to Group A.

To understand the differences between these participant groups, we performed a mixed ANOVA, where we modelled the participant groups as a between-subjects variable additionally to the five conditions as a within-subjects variable. The results are shown in Figure 6.

SRTs. We found a significant main effect of the participant groups on the *SRTs*, $F(1,19) = 51.5$, $\boldsymbol{p < 0.001}$, $\eta_p^2 = 0.73$.

Post-hoc tests with Bonferroni correction for Group A showed that the following pairs were significant: (*Audio*, *Low_AV$_E$*), (*Audio*, *High_AV*), (*Audio*, *High_AV$_E$*), and (*Low_AV$_E$*, *High_AV$_E$*). For Group B, all pairs were significant, except (*Low_AV*, *Low_AV$_E$*), (*Low_AV$_E$*, *High_AV*), and (*High_AV*, *High_AV$_E$*).

Confidence. We found a significant main effect of the participant groups on the *confidence*, $F(1,19) = 7.38$, $\boldsymbol{p = 0.014}$, $\eta_p^2 = 0.28$. Post-hoc tests with Bonferroni correction for Group A showed that none of the pairs were significant. For Group B, all pairs were significant, except (*Low_AV$_E$*, *High_AV*), (*Low_AV$_E$*, *High_AV$_E$*), and (*High_AV*, *High_AV$_E$*).

Self-Assessed Performance. We did not find a significant main effect of the participant groups on self-assessed *performance*.

5 Discussion

Our experiment further reinforces prior research about the benefits of lip reading on speech perception. From our results in Section 4.1, we see clear benefits of visuals on participants' speech perception, confidence and self-assessed performance. Especially in the realm of social communication through displays, the strong effect size ($\eta_p^2 = 0.71$) revealed how meaningful the presence of humans and their faces were.

In Section 5.1, based on our results, we discuss the research questions and hypotheses, which we posited in Section 1 and Section 3.5. We go into further details into the interesting effects from the experiment, and specifically the effects of display resolution, enhancing the human face. Additionally, in Section 5.2, to understand the groupings, we discuss the interpersonal differences in users' audiovisual speech perception.

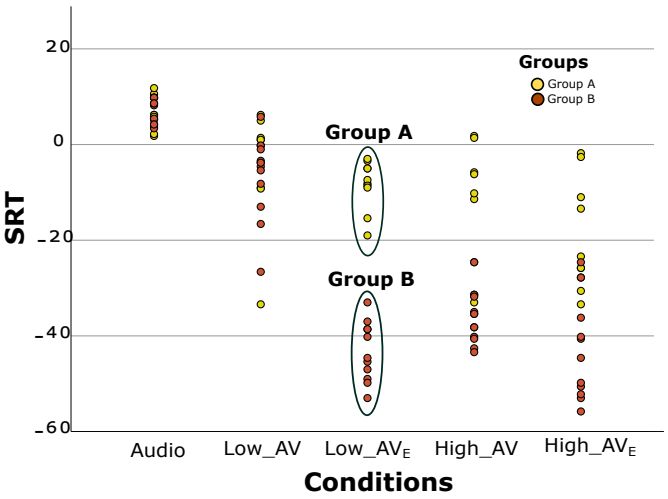


Fig. 5. Scatter plot showing the color-coded differences in SRTs for the two participant groups we observed in our study. Group A is colored in yellow, while Group B is colored in red. It appears that Group B benefited clearly more from the higher resolution and facial enhancement than Group A.

5.1 Display Resolution and Facial Enhancement on Audiovisual Speech Perception

To address **RQ1** and **RQ2**, we hypothesized better speech perception and higher confidence when – **H1** the face was enhanced than not, and **H2** the display had higher resolution than lower. In line with our Hypothesis **H1** and **H2**, we see a clear benefit with face enhancement and higher display resolution on speech perception and confidence levels (refer to Sections 4.2 and 4.3). The toughest condition to perceive speech was the Low_AV condition (lowest resolution and no enhancement), which makes sense because it provided the fewest details to the users. As more details were provided via higher resolution or the facial enhancement, we see a clear improvement in participants' speech perception, confidence, and self-assessed performance scores. This would suggest that such visual changes are viable factors to effect humans' audiovisual speech perception.

Partially in line with our Hypothesis **H3**, when we compare conditions involving higher resolution and/or the visual enhancement, even with a strong main effect, we did not find all pairs to be significant, especially between Low_AV$_E$, High_AV and High_AV$_E$ conditions. Amongst the visual details users' may take advantage off, we believe an understandable factor to assume is the pixels dedicated to the face/mouth. More pixels-to-face/mouth should help users' resolve lip movements better. To further understand the non-significance of the three conditions (Low_AV$_E$, High_AV and High_AV$_E$), we calculated the number of pixels-to-face/mouth for each of them. Our human had the following spatial properties: standing at 3 meters from the user, and head height and breadth of 24.1 cm and 14.5 cm respectively [36]. We calculated the horizontal and vertical angle to the head to be 2.76 ° and 4.46 ° for the non-enhanced conditions, and 5.53 ° and 9.15 ° for the enhanced condition. Please note that the enhanced condition additionally had higher contrast which users would benefit from which cannot be determined by pixels-to-face/mouth alone. To calculate pixels-to-face/head, we then used the angular pixel density[6] of the Meta Quest 2. For high resolution, pixels per degree horizontally and vertically were 18.88 and 18.69 respectively, and for lower resolution, it was 9.44 and 9.34 respectively. Finally, we calculated the following horizontal and vertical pixels-to-face: Low_AV$_E$ (52.2 px and 85.5 px), High_AV (52.1 px and 83.36 px) and High_AV$_E$ (104.4 px and 171.1 px). Despite having different visuals, we can evidently see Low_AV$_E$ (52.2 px and 85.5 px) and High_AV (52.1 px and 83.36 px) dedicate almost the same number of pixels to the human face. This could possibly explain the non-significance the pair, and almost equal speech perception and confidence responses.

If we keep up with this train of thought, for condition High_AV$_E$, along with higher contrast, we see almost double pixels-to-face of 104.4 px and 171.1 px, but we still do not see any significance. We expected this condition to be the best condition for audiovisual speech perception. We think that in this scenario, the amount of details provided to the user was reaching or had already reached a saturation threshold. Beyond this threshold, visual enhancements may not have as strong an effect on speech perception on our audiovisual speech perception.

In our enhancement condition, we investigate the effects of higher contrast and head scaling collectively, not individually. Since we cannot claim statistically about which change benefited more, we asked participants after the experiment, amongst the two changes (contrast and head scaling), which one they believed helped them the most. 10 voted for head scaling, 9 for both equally and 2 for the contrast. The participants who preferred head scaling stated that it was easy:

> P2: *"The big head also enlarged the lips and that made reading the lips and even focusing better."*
>
> P3: *"The big head helped for low res conditions because it increased the amount of signal I got out of a low-signal condition. I could better tell the way the lips were moving in that case."*
>
> P19: *"Increases the effective resolution of the mouth animation cues"*

Participants who found contrast to be helpful stated that:

> P6: *"Red lips helped to understand the interpretation and sounds that the person is trying to make."*
>
> P11: *"It showed more contrast in the face to see clearly, especially at low resolution."*

Participants believed they benefited from head scaling more than the contrast change, nevertheless, with our results and participant anecdotes, we believe both enhancements were helpful and can be used in tandem to improve human audiovisual speech perception.

In our experiment, with noise level of 60 dB, participants required speech levels improved from 6.7 dB above the noise level to 31 dB below the noise level. This is a rather large improvement and could make communication more efficient and accessible. We see two main applications where this could possibly be beneficial, collaborative environments and people who are hard-of-hearing. In collaborative environments, there will be situations where there is intended or unintended noise that can negatively effect users to communicate, in such situations, our enhancement can be used to help people hear clearer. Similarly, people who are hard-of-hearing can also take advantage of this enhancement.

5.2 Interpersonal Differences

To try to understand the grouping of participants discussed in Section 4.5 we did additional analysis to test its significance with our two measures, and found a significant main effect between the groups for both measures. Additionally, out of 10 total conditions pairs, only 4 pairs were significant in Group A, whereas there were 7 significant pairs in Group B. This would suggest that our strong overall effect, could be mainly contributed by participants from Group B. This further implies that everybody benefited from higher display resolution and/or the facial enhancement on their speech perception, however some more than others. Since human speech perception is multi-modal and can be effected by multiple factors, we believe this grouping could due to interpersonal differences between the participants. Because we were not expecting such a strong effect due to interpersonal differences, we collected only limited demographic information from the subjects at the time of the experiment. Using that information we ran the following correlation tests: participant age, age group (young and old adults),

[6] https://vr-compare.com/headset/oculusquest2

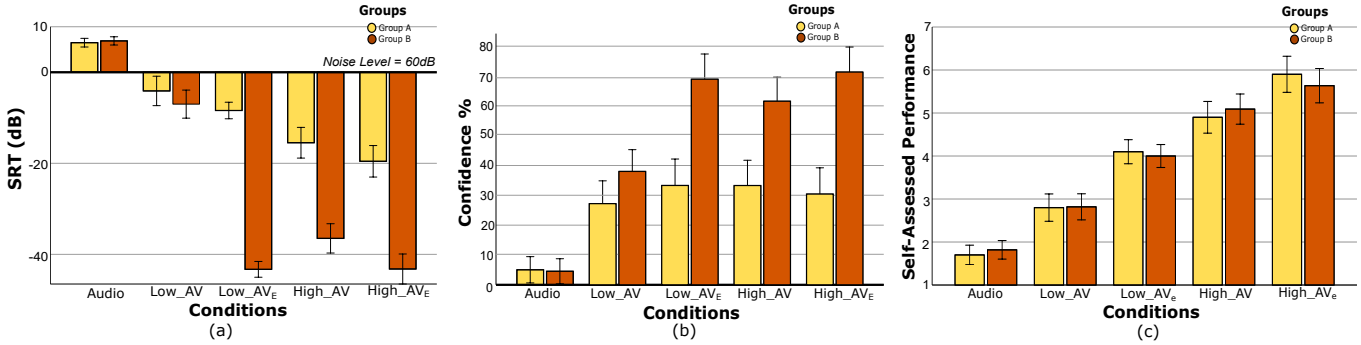


Fig. 6. Exploratory analysis between two participant groups, A and B, colored yellow and brown, respectively. The x-axes show our five conditions. The y-axes show the results for our three measures: (a) Speech level relative to the noise level (fixed at 60 dB) for participants to understand 50% of the speech stimuli (lower is better); (b) Confidence %; (c) Performance scores on a scale from 1=worst performance to 7=best performance. The error bars show standard error.

preferred gender, and familiarity with VR. We did not find any effect with the groups. We also found no correlation with the dates or times of the trials of the group members. Having ruled out all of these factors, we began to wonder about other individual traits that could affect the measures. A review done by Belojevic et al. [6] investigated individual personality differences on mental performance in the presence of noise. They focused on the roles of neuroticism, extroversion and subjective noise sensitivity. Their results suggests that personalities with extroversive tendencies better adapt to noisy environments than opposite personality traits. Cultural and language differences can also play a role in the use of visual facial cues. For example, work done by Sekiyama et al. [51] shows that Japanese listeners were less influenced by visual cues than American listeners. While we did not record the participants' race or cultural upbringing, it is possible that cultural differences are playing a role in the groupings. So while we do not have a definitive explanation for the groupings, the strong effect on speech perception could be explained by a combination different theories about interpersonal differences. This unexpected grouping warrants further investigation.

6 Limitations & Future Work

Our study showed interesting effects of display resolution and a visual facial enhancement on speech perception in noise. We see clearly the improvement due to higher resolution and enhancement; however, there are also a few limitations to the current work, which can lead us to interesting research vistas that may be investigated in the future.

Firstly, this study used a female human as the speaker, and the type of speaker has been shown to affect our speech perception [54]. Future work may investigate different speaker types, such as race, skin color, age, and gender, as well as rendering styles (for displays), such as photorealistic or abstracted cartoonish characters.

Second, our target speech material was limited to 3-digit numbers. There are alternative established speech materials , such as BKB words and IEEE sentences. Based on the material we select, different measuring protocols can also be employed, such as BKB-SIN, HINT, QuickSIN, or WIN [60].

Thirdly, despite making efforts to replicate life-like scenarios, we could attempt this experiment with a AR display instead of VR. This would require a computer vision solution to identify the speakers head and apply the enhancement, similar to Choudhary et al.'s prototype [15]. One could further test different environmental lighting conditions on the AR solution, such as in a bright outdoor environment, AR displays perform poorly [17]. Such a solution could be especially beneficial for people who are hard-of-hearing.

Lastly, while our facial enhancement was limited to lip contrast and head scaling. One could investigate other facial features to manipulate, such as facial complexion or the eye details. Similarly, one could attempt to amplify the lip motion and study its influence on speech perception.

7 Conclusion

In this paper, we presented a comparative user study in which we investigated an audio-only condition compared to two levels of head-mounted display resolution and two levels of the native or visually enhanced appearance of a virtual human, the latter consisting of an up-scaled facial representation and simulated lipstick (lip coloring) added to increase contrast. First, we our results show that display resolution affected participants' speech perception in noise. Second, we found that our visual enhancement method was able to improve participants' speech perception and could also compensate for lower display resolutions. Third, among our facial enhancements, participants generally preferred head scaling over increasing facial contrast. Our results indicate that similar facial enhancements may be leveraged by practitioners with a range of VR/AR technologies to improve human speech perception in the presence of noise. We discussed potential explanations, implications, and applications to guide practitioners aiming to leverage these techniques, and we discussed the limitations of our experiment as well as future avenues for research.

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] Audacity: Free, open source, cross-platform audio software. `https://www.audacityteam.org/`, 2022.

[2] 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`, 2022.

[3] Meta Quest 2: Virtual reality headset by meta. `https://www.meta.com/quest/products/quest-2/`, 2022.

[4] A. Alsius, R. V. Wayne, M. Paré, and K. G. Munhall. High visual resolution matters in audiovisual speech perception, but only for some. *Attention, Perception, & Psychophysics*, 78:1472–1487, 2016.

[5] P. Assmann and Q. Summerfield. The perception of speech under adverse conditions. *Speech processing in the auditory system*, pp. 231–308, 2004.

[6] G. Belojevic, B. Jakovljevic, V. Slepcevic, et al. Noise and mental performance: Personality attributes and noise sensitivity. *Noise and Health*, 6(21):77, 2003.

[7] P. Bertelson and M. Radeau. Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. *Perception & psychophysics*, 29(6):578–584, 1981.

[8] J. Blascovich, J. Loomis, A. C. Beall, K. R. Swinth, C. L. Hoyt, and J. N. Bailenson. Immersive virtual environment technology as a methodological tool for social psychology. *Psychological inquiry*, 13(2):103–124, 2002.

[9] R. E. Bouserhal, A. Bockstael, E. MacDonald, T. H. Falk, and J. Voix. Modeling speech level as a function of background noise level and talker-to-listener distance for talkers wearing hearing protection devices. *Journal of Speech, Language, and Hearing Research*, 60(12):3393–3403, 2017.

[10] G. Calvert, E. Bullmore, M. Brammer, R. Campbell, S. Iversen, P. Woodruff, P. McGuire, S. Williams, and A. David. Silent lipreading activates the auditory cortex. *Science*, 276:593–596, 1997.

[11] Z. Choudhary, G. Bruder, and G. F. Welch. Scaled user embodied representations in virtual and augmented reality. In *Workshop on User-Embodied Interaction in Virtual Reality (UIVR) 2021*, 2021.

[12] Z. Choudhary, A. Erickson, N. Norouzi, K. Kim, G. Bruder, and G. Welch. Virtual big heads in extended reality: Estimation of ideal head scales and perceptual thresholds for comfort and facial cues. *ACM Transactions on Applied Perceptions*, 20(1):1–31, 2023.

[13] Z. Choudhary, K. Kim, R. Schubert, G. Bruder, and G. F. Welch. Virtual big heads: Analysis of human perception and comfort of head scales in social virtual reality. In *2020 IEEE conference on virtual reality and 3D user interfaces (VR)*, pp. 425–433. IEEE, 2020.

[14] Z. Choudhary, N. Norouzi, A. Erickson, R. Schubert, G. Bruder, and G. F. Welch. Exploring the social influence of virtual humans unintentionally conveying conflicting emotions. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 571–580. IEEE, 2023.

[15] Z. Choudhary, J. Ugarte, G. Bruder, and G. Welch. Real-time magnification in augmented reality. In *Proceedings of the 2021 ACM Symposium on Spatial User Interaction*, pp. 1–2, 2021.

[16] B. E. Dodd and R. E. Campbell. *Hearing by eye: The psychology of lip-reading.* Lawrence Erlbaum Associates, Inc, 1987.

[17] A. Erickson, K. Kim, G. Bruder, and G. F. Welch. Exploring the limitations of environment lighting on optical see-through head-mounted displays. In *Proceedings of the 2020 ACM Symposium on Spatial User Interaction*, pp. 1–8, 2020.

[18] C. A. Fowler and D. J. Dekle. Listening with eye and hand: cross-modal contributions to speech perception. *Journal of experimental psychology: Human perception and performance*, 17(3):816, 1991.

[19] R. Fribourg, E. Peillard, and R. Mcdonnell. Mirror, mirror on my phone: Investigating dimensions of self-face perception induced by augmented reality filters. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 470–478. IEEE, 2021.

[20] M. Gonzalez-Franco, A. Maselli, D. Florencio, N. Smolyanskiy, and Z. Zhang. Concurrent talking in immersive virtual reality: on the dominance of visual speech cues. *Scientific reports*, 7(1):3817, 2017.

[21] M. Gonzalez-Franco, E. Ofek, Y. Pan, A. Antley, A. Steed, B. Spanlang, A. Maselli, D. Banakou, N. Pelechano, S. Orts-Escolano, et al. The rocketbox library and the utility of freely available rigged avatars. *Frontiers in virtual reality*, p. 20, 2020.

[22] A. L. Jones and R. S. Kramer. Facial cosmetics have little effect on attractiveness judgments compared with identity, 2015.

[23] A. L. Jones and R. S. Kramer. Facial cosmetics and attractiveness: Comparing the effect sizes of professionally-applied cosmetics and identity. *PloS one*, 11(10):e0164218, 2016.

[24] T. R. Jordan and P. Sergeant. Effects of distance on visual and audiovisual speech recognition. *Language and Speech*, 43(1):107–124, 2000.

[25] T. R. Jordan and P. C. Sergeant. Effects of facial image size on visual and audio–visual speech recognition. 1998.

[26] T. R. Jordan and S. M. Thomas. Effects of horizontal viewing angle on visual and audiovisual speech recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 27(6):1386, 2001.

[27] R. Kanzaki and R. Campbell. Effect of facial brightness reversal on visual and audiovisual speech perception. In *AVSP'99-International Conference on Auditory-Visual Speech Processing*, 1999.

[28] N. Kitagawa and S. Ichihara. Hearing visual motion in depth. *Nature*, 416(6877):172–174, 2002.

[29] K. Kiyokawa, H. Takemura, and N. Yokoya. A collaboration support technique by integrating a shared virtual reality and a shared augmented reality. In *IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 99CH37028)*, vol. 6, pp. 48–53. IEEE, 1999.

[30] Y. Kobayashi, S. Matsushita, and K. Morikawa. Effects of lip color on perceived lightness of human facial skin. *i-Perception*, 8(4):2041669517717500, 2017.

[31] S. Kopp, J. Allwood, K. Grammer, E. Ahlsén, and T. Stocksmeier. Modeling embodied feedback with virtual humans. *Lecture Notes in Computer Science*, 4930:18, 2008.

[32] J. Krauskopf and B. Farell. Vernier acuity: effects of chromatic content, blur and contrast. *Vision Research*, 31(4):735–749, 1991.

[33] V. Krauß, A. Boden, L. Oppermann, and R. Reiners. Current practices, challenges, and design implications for collaborative ar/vr application development. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2021.

[34] N. Krishnamurthy and J. H. Hansen. Babble noise: modeling, analysis, and applications. *IEEE transactions on audio, speech, and language processing*, 17(7):1394–1407, 2009.

[35] K. Lander and C. Capek. Investigating the impact of lip visibility and talking style on speechreading performance. *Speech Communication*, 55(5):600–605, 2013.

[36] J.-h. Lee, S.-J. Hwang Shin, and C. L. Istook. Analysis of human head shapes in the united states. *International journal of human ecology*, 7(1):77–83, 2006.

[37] M. Lee, N. Norouzi, G. Bruder, P. J. Wisniewski, and G. F. Welch. Mixed reality tabletop gameplay: Social interaction with a virtual human capable of physical influence. *IEEE transactions on visualization and computer graphics*, 27(8):3534–3545, 2019.

[38] J. MacDonald and H. McGurk. Visual influences on speech perception processes. *Perception & psychophysics*, 24(3):253–257, 1978.

[39] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, 1976.

[40] A. A. Michelson. *Studies in optics.* Courier Corporation, 1995.

[41] M. Middelweerd and R. Plomp. The effect of speechreading on the speech-reception threshold of sentences in noise. *The Journal of the Acoustical Society of America*, 82(6):2145–2147, 1987.

[42] G. Musacchia, M. Sams, T. Nicol, and N. Kraus. Seeing speech affects acoustic information processing in the human brainstem. *Experimental Brain Research*, 168:1–10, 2006.

[43] N. Norouzi, K. Kim, G. Bruder, A. Erickson, Z. Choudhary, Y. Li, and G. Welch. A systematic literature review of embodied augmented reality agents in head-mounted display environments. In *Proceedings of the International Conference on Artificial Reality and Telexistence & Eurographics Symposium on Virtual Environments*, 2020.

[44] T. Piumsomboon, A. Day, B. Ens, Y. Lee, G. Lee, and M. Billinghurst. Exploring enhancements for remote mixed reality collaboration. In *SIGGRAPH Asia 2017 Mobile Graphics & Interactive Applications*, pp. 1–5. 2017.

[45] R. Plomp and A. Mimpen. Improving the reliability of testing the speech reception threshold for sentences. *Audiology*, 18(1):43–52, 1979.

[46] I. Pollack. Auditory informational masking. *The Journal of the Acoustical Society of America*, 57(S1):S5–S5, 1975.

[47] L. Rosenblum. Audiovisual speech perception and the mcgurk effect. *Oxford Research Encyclopedia, Linguistics*, 2019.

[48] L. D. Rosenblum. Speech perception as a multimodal phenomenon. *Current directions in psychological science*, 17(6):405–409, 2008.

[49] L. D. Rosenblum, D. A. Yakel, and K. P. Green. Face and mouth inversion effects on visual and audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 26(2):806, 2000.

[50] M. Sams, R. Möttönen, and T. Sihvonen. Seeing and hearing others and oneself talk. *Cognitive Brain Research*, 23(2-3):429–435, 2005.

[51] K. Sekiyama and Y. Tohkura. Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics*, 21(4):427–444, 1993.

[52] M. Slater and M. Usoh. Presence in immersive virtual environments. In *Proceedings of IEEE virtual reality annual international symposium*, pp. 90–96. IEEE, 1993.

[53] C. Smits, T. S. Kapteyn, and T. Houtgast. Development and validation of an automatic speech-in-noise screening test by telephone. *International journal of audiology*, 43(1):15–28, 2004.

[54] E. A. Strand. Uncovering the role of gender stereotypes in speech perception. *Journal of language and social psychology*, 18(1):86–100, 1999.

[55] Q. Summerfield. Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 335(1273):71–78, 1992.

[56] H. Tanaka. Facial cosmetics exert a greater influence on processing of the mouth relative to the eyes: Evidence from the n170 event-related potential component. *Frontiers in psychology*, 7:1359, 2016.

[57] H. Tanaka. Lip color affects erp components in temporal face perception processing. *Journal of Integrative Neuroscience*, 20(4):1029–1038, 2021.

[58] D. Y. Tsao and M. S. Livingstone. Mechanisms of face perception. *Annu. Rev. Neurosci.*, 31:411–437, 2008.

[59] E. Van den Borre, S. Denys, A. van Wieringen, and J. Wouters. The digit triplet test: a scoping review. *International journal of audiology*, 60(12):946–963, 2021.

[60] R. H. Wilson, R. A. McArdle, and S. L. Smith. An evaluation of the bkb-sin, hint, quicksin, and win materials on listeners with normal hearing and listeners with hearing loss. 2007.

[61] L. Woodhouse, L. Hickson, and B. Dodd. Review of visual speech perception by hearing and hearing-impaired people: Clinical implications. *International Journal of Language & Communication Disorders*, 44(3):253–270, 2009.