CollarMic: Interacting with Voice Assistant on HMDs in Social Leveraging Speak-to-Shoulder Gesture

Category: Technical

ABSTRACT

Voice assistants (VAs) are becoming increasingly popular in VR and AR. However, speaking to the VA in social (e.g., during conversation with others) may lead to misunderstanding and embarrassment. In this paper, we proposed CollarMic, a technique that allowed the users to use speak-to-shoulder gesture to indicate whether to talk to the VA. Through a brainstorming with 10 experts, we first collected a total of 62 voice input triggering gestures in social that leveraged different body parts (e.g., head and hand). We elicitated 6 candidate gestures based on voting, and asked another 26 participants to try and rate them in terms of different subjective dimensions. According to the result, the speak-to-shoulder gesture was selected for CollarMic. We then developed the algorithm of CollarMic, which used Gaussian Hidden Markov Model (GHMM) to detect the triggering gesture. Based on the collected data from 15 real users, this algorithm reached an accuracy of 98.0%, and a false triggering rate of 0.13 times per minute, which was higher than the SVM-based algorithm. In the usability evaluation study with two different social scenarios (business and chatting), CollarMic was significantly more preferred by the users than keyword spotting, in terms of interaction speed and social acceptance.

Index Terms: Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Mixed / augmented reality; Human-centered computing—Human computer interaction (HCI)—Interaction devices—Sound-based input / output

1 INTRODUCTION

With the rapid development of speech recognition algorithms, voice assistants (VAs) are becoming increasingly popular in our daily lives. The ability of eyes-free and hands-free interaction makes interacting with VAs superior to typical touch and gesture interaction in various scenarios, including smart home (e.g., Amazon Alexa[1]), driving (e.g., Tesla[2]), and on-device agent (e.g., Apple Siri[3]). In recent years, voice interaction has also become an attractive solution for virtual reality (e.g., Meta Quest 2[4]) and augmented reality (e.g., Apple Vision Pro[5]).

Noticeably, nowadays, VAs are mostly used for personal tasks in single-user scenarios (e.g., home appliance controlling and information retrieval). However, with the development of artificial intelligence (e.g., large language models), it is expected that VAs be used for more complex tasks, and in more various scenarios. For example, during a working meeting, the user could ask the VA to provide additional data. During the conversation with friends, the user could ask the VA to quickly respond to SMS.

However, interacting with VAs on social may lead to misunderstanding, as other people around the user (either physically or virtually) may not be able to distinguish the object of the voice command. Most of today's VAs used keyword spotting to detect voice commands (e.g., "Hey Siri") and avoid false triggering. However, this would significantly reduce the interaction efficiency and may lead to embarrassment when used in social. Although researchers have proposed various keyword-free triggering methods (e.g., ProxiMic [24]), none of them have been tested in social, in which both the experience of the speaker and the listener should be considered.

In this paper, we explored how to trigger voice interaction with VAs in social, with the goal of achieving a balance between interaction performance and social experience. Specifically, we chose head-mounted-displays (HMDs) as the apparatus, as we envisioned the pervasiveness of social VR [17] and mobile AR [8]. Furthermore, we emphasized keyword-free triggering gestures, as they have potentially higher interaction speed, and are expected to be less interrupting.

To elicitate the possible gestures, we conducted a brainstorming with 10 experts and generated a total of 62 gestures that leveraged body-motions (foot, hand, mouth, and head) or acoustic features. Then, we asked another 26 participants to rate the 6 most voted gestures in terms 6 dimensions, from the perspective of both the speaker and the listener. Results showed that the "tilt head to the side and lower it" gesture was the most preferred.

Based on the findings, we proposed CollarMic, a technique for keyword-free voice interaction with VAs in social. The user could perform the gesture above to indicate the intention of talking to the VA, which is also easy for the listener to understand. After the interaction, the user can turn his/her head back to neutral orientation, and continue talking to the listener.

We designed and validated the algorithm of CollarMic based on the collected data from 15 users, including both positive samples when performing the gesture, and negative samples during wandering and chatting. Simulation results suggested a GHMM to detect the triggering gesture, which achieved an accuracy of 98.0%, and a false triggering rate of 0.13 times per minute.

We finally evaluated the usability of CollarMic in two realistic social scenarios (business and chatting). Compared with the keyword spotting (KWS) baseline technique, CollarMic was superior in terms of confusion and was significantly more preferred by the participants in terms of all other dimensions.

The contributions of this paper are three-folded:

- Through brainstorming and questionnaire, we first investigated the design space of triggering gestures with HMD VAs in social. The results also highlighted the gesture that was the most preferred by the participants.
- We proposed CollarMic, a novel technique for interacting with VAs on HMD. CollarMic leveraged head movement for triggering, and used GHMM for accurate and robust real-time motion detection.
- We evaluated the usability of CollarMic in realistic business and chatting scenarios. Results showed that CollarMic outperformed the keyword spotting baseline technique in terms of all the subjective rating dimensions.

2 RELATED WORK

2.1 Input Triggering Approaches in Voice Interaction

Voice interaction has seen significant advancements through various input triggering approaches. A notable method has been the use of sound for real-time event detection and classification. For instance, Voice Activity Detection (VAD) algorithms, as referenced in [26,36],

[1] https://alexa.amazon.com/

[2] https://www.tesla.com/support/voice-commands

[3] https://www.apple.com/siri/

[4] https://www.meta.com/quest/products/quest-2/

[5] https://www.apple.com/apple-vision-pro/

have been specifically designed to discern the presence of human speech. Traditionally, the robustness of Key Word Spotting (KWS) technology has been harnessed, with wake-up phrases serving as the primary activation mechanism for voice assistants, as documented in [7, 13, 34]. However, this approach isn't without its challenges, such as cumbersome interactions and concerns related to privacy and security. This has led to a shift in focus among researchers towards exploring alternative activation methods. For instance, gaze wake-up has been introduced for devices with static positions like smart speakers and in-vehicle systems, where a user can activate voice input merely by directing their gaze at the device [19, 29]. In augmented reality settings, FaceSight [39] employs a camera on glasses to detect the cover-mouth gesture, subsequently activating voice input. Apple, in its innovative stride, introduced the "Raise to Speak" feature for smartwatches [47], necessitating a vertical wrist movement of approximately 4 inches for activation.

Instead of modeling events by extracting features from a single input, PrivateTalk [43] uniquely identifies the Hand-On-Mouth gesture by comparing the differences between two distinct audio inputs. ProxiMic [25] incorporates hand movements and proximity to the microphone. ProxiTalk [45] stands out with its multifaceted approach, integrating signals from cameras, IMU, and dual microphones to realize a resilient voice activation system tailored for smartphones. However, these methods require the coordination of the head or hands, thus can lead to user fatigue and may not be optimal for VR/MR environments.

Our CollarMic technique amalgamates voice interaction with a straightforward speak-to-shoulder gesture. This approach amplifies usability in VR/MR contexts, curtails user fatigue, and maintains privacy.

2.2 Voice Interaction with Other Modalities

Many past researchers have hoped to fuse voice with other modalities to enrich the interaction. Some examined mid-air hand gestures [3, 9, 11, 15, 21], others only looked at a subset of gesturing such as pointing gestures using hand [2, 28], paddling gestures using hand [10], or two dimensional (2D) hand gestures [20, 28]. These works expanded the expressiveness of voice interaction in complex tasks but at the cost of higher learning effort and more fatigue.

Irawati et al. [10] presented an augmented reality system that harnesses multi-modal input by integrating paddle gestures with voice commands. Notably, this system doesn't accommodate free-hand gestures. A comparative study [16] assessing voice-only, free hand gesture only, and multimodal inputs determined that the multimodal approach was superior in usability and user satisfaction for object selection tasks in augmented reality. However, they only conducted the study on a 2D interface. Thammathip et al. [23] developed G-Speech, a multimodal interaction method that combined voice input and free-hand gesture for 3D object manipulation. Complementing voice channels through gestures was proved to largely increase efficiency, whereas it may result in a higher mental load.

Building upon these foundations, we introduced a distinct speak-to-shoulder gesture for triggering the voice interaction, which enhanced usability in social contexts and ensured precise voice input activation.

2.3 Communication Intention Interpretation in Social Virtual Reality

Among interaction between users and agents in social virtual reality, the interpretation of users' intentions was important. To interpret users' communication intention, most of the mentioned platforms apply Broadcast method to convey users' voice communication that the speech of each user will be delivered to everyone else. However, it leads to turn-taking conflicts [4, 32] and does not support simultaneous conversations. VRChat provided 3D spatialized audio [37] to help users distinguish important conversations heard from different directions. Yan et al. [42] proposed a technique, ConeSpeech, using directional speech delivery in line with the user's head orientation to conduct voice communication in virtual reality. Researchers have also explored non-verbal communication signals in virtual reality including nodding [1], waving [17], applause [17], etc. They used different ways (e.g., spatial directions) to differentiate the communication target. But the experience of the listener was not formally examined, nor the privacy and the social effect of these communication manners.

In contrast, attitudes, presence [5, 30, 31] and privacy were also proved important by researchers in social VR scenarios. Julian et al. [27] and Tarr et al. [38] underlined the positive impact of presence on user cohesion, attitudes, and overall behavior during intent communication in multi-user VR environments. [22] found the users' inclination towards silent speech input to bolster privacy and security. Drawing from these insights, CollarMic introduced a socially considerate gesture for voice interaction in VR and AR contexts to amplify user immersion, deepen emotional engagement, and safeguard privacy.

3 Study 1: Brainstorming the Voice Input Triggering Gesture in Social

We aimed to explore suitable user-defined gestures to activate voice assistants. To achieve this, we conducted a brainstorming session following the guidelines established in [35]. The session is designed to deepen understanding of interaction spaces by enabling participants to design and discuss collaboratively, guiding consensus toward high-quality results. During the session, experienced experts in VR were assembled to share and evaluate their original ideas. Finally, the experts voted on the ideas that closely aligned with our design goals, leading to the selection of the highest-quality designs. We applied filtering rules based on agreement ratios to maintain a balance between the quantity and quality of the selected ideas.

3.1 Participants

This study engaged 10 knowledgeable participants from our campus specializing in VR interaction. The group comprised 8 males and 2 females, aged between 19 and 26. All participants were educational or academic professionals; 8 had experience in VR/MR research, and 2 had experience using VR/MR devices. Each participant had prior experience in conducting interactive research on smart devices.

3.2 Study Design

We sought to explore the ability of users to seamlessly switch interaction targets between third parties and smart agents in VR/MR environments. Prior methods have concentrated on language expression, gaze and hand gestures. However, our objective is to examine the design space for intention switching, without confining users to specific gestures or considering recognition difficulty.

To maintain gesture practicality, we impose one limitation: users should not directly utilize facial expressions, as their excessive use may interfere with social interactions and must be employed with discretion.

We outlined several target scenarios, including: 1) educational settings where teachers alternate between students and their smart agents; 2) workplace situations where employees transition from conversations with superiors to their smart agents; and 3) social interactions where users shift focus from peers to their smart agents. The application scenarios are semi-open, excluding situations where attention cannot be freely diverted.

We encouraged participants to think creatively, providing the following guidance:

- Please explore a range of possibilities and ideas. We welcome innovative and unique thoughts.

- To avoid constraining the emergence of new ideas, participants were instructed not to judge others' ideas when they were presenting.

3.3 Procedure

We commenced by assembling participants in a tranquil meeting room, allowing them some time to familiarize themselves with each other. A large blackboard and differently colored stickers were used to document the suggested ideas. Participants were tasked with designing gestures capable of expressing users' intentions to shift interaction targets between other subjects and smart agents coexisting in VR/MR environments.

Initially, each participant was given 15 minutes in a quiet setting to independently formulate ideas, encouraged to jot down 30 distinct thoughts individually. Subsequently, participants sequentially presented their ideas, affixing them to the board. However, if any proposed idea was a repetition, they were prompted to substitute it with a new one. This cycle continued until no fresh ideas emerged, with everyone permitted to note down additional thoughts in the meantime. Following the unified presentation, a group discussion was held to address any overlooked ideas, with participants free to explore inspiration sources or pinpoint design features. The discussion ranged between 15 to 30 minutes, concluding once all topics were covered.

Ultimately, participants were requested to cast their votes for up to 20 of the most satisfactory ideas from the pool. Ideas garnering votes were considered of superior quality. Voting was guided by the following criteria: confusion, simplicity, fatigue, and embarrassment.

3.4 Results

3.4.1 Collected Gestures

We initially analyzed the ideas generated from the brainstorming session and their distributions. Users proposed a total of 68 different ideas, but 6 were disqualified as they infringed on the restriction against facial expressions, leaving 62 valid ideas.

In this study, we have categorized and analyzed both the complete set of 62 ideas, as illustrated in Table 1. The selected gestures are classified based on the action parts involved and the types of signs used. The classification of action parts is determined by the primary body parts engaged in executing the motions, including 4 parts—1) *foot*, 2) *hand*, 3) *head* and 4) *mouth*. Several actions integrate multiple movement parts due to gesture combinations; efforts were made to consolidate these into single categories, although some cases with duplication were counted separately (e.g., 'lower head while covering mouth' involves both *head* and *hand*).

Table 1: Statistics of all the 62 brainstormed gestures. One gesture leveraged "hand"+"head", thus was counted twice.

Form \ Part	Foot	Hand	Head	Mouth
Motion	4	46	5	1
Acoustic	4	2	0	1
Total	8	48	5	2

The gestures are segregated into two categories based on their sign type: *motion* and *acoustic*. The *motion* category involves one or more complete actions serving as the trigger sign, as seen in 'sliding on the side of the HMD with your hand'. Conversely, the *acoustic* type employs voice as a trigger, exemplified by 'snap fingers' or 'cough'. According to Table 1, gestures combining *hand* and *motion* are predominant (46 out of 62), signifying that hand movements can efficiently and naturally convey user intent and are the most familiar way for users.

3.4.2 Voted Gestures

Among 62 gestures, 33 received at least one vote, indicating that participants preferred these ideas.

We analyzed the detailed votes and distribution. Among 10 voters, their voting numbers ranged from 7 to 13, with a mean of 10.0 (std: 1.76). The numbers implying participants would appraised the ideas with high quality.

To test the effectivity of the voting, we utilized statistical methods to analyze the agreement and reliability of the result. We represented votings of P_i as a 62 dimension vector in the form of $\vec{V}_i \in \{0,1\}^{62}$. Referring to the similar works [41, 44], we used Equation 1 to calculate the agreement score, which yields $A_{vote} = 0.478$. The score is above 0.4, meaning the voters are of high agreement.

$$A_{vote} = \sum_{p_i \in \{P_i\}} \sum_{p_j \in \{P_i\}} \frac{\vec{V}_i \bullet \vec{V}_j}{\sqrt{\left(\vec{V}_i \bullet \vec{V}_i\right) \times \left(\vec{V}_j \bullet \vec{V}_j\right)}} \quad (1)$$

To evaluate the reliability of voting, we further calculated the intra-class correlation [33] of the result. The corresponding indicator for this experiment was $ICC(3,k)$, meaning "reliability of the mean rating" in Case 3, and the result yielded $ICC(3,k) = 0.811$.

All of the above proved the mean value of the votes was acceptable. Thus we finally took the ideas with a mean rating over a half as the passed ideas, which is listed in Table 2. We did not directly regard the voting number as the actual assessment of every gesture, because on the one hand, the voters are limited to gapping scenarios upon voting, and on the other hand, the voting result still lacks representative experiment conclusions.

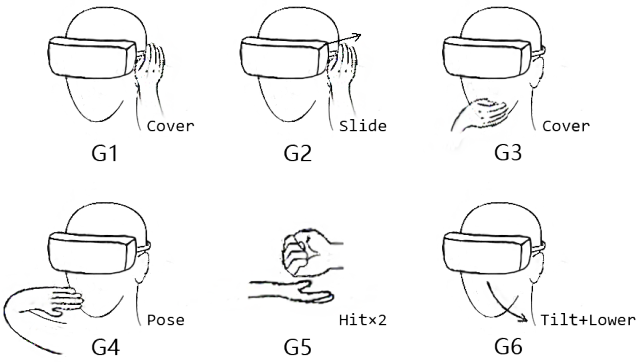


Figure 1: The six gestures with the highest votes.

Table 2: Description of the six gestures with the highest votes.

Part	Votes	Description (Gesture ID)
Hand	8	Cover the ear with one hand. (G1)
Hand	8	Slide the hand on the side of HMD and hold. (G2)
Hand	7	Cover the mouth with one hand. (G3)
Hand	6	Pose the palm facing outward horizontally in front of the face. (G4)
Hand	6	One hand clenches into a fist and hits the other palm twice then stills. (G5)
Head	6	Tilt the head to either side and lower it. (G6)

The finally selected 6 gestures are depicted in Figure 1. Both G1 and G2 involve the head and one hand, showing the intention

of interacting with the HMD. These two designs are most approved by the experts with each of 8 votings. G3 and G4 also involve the head and one hand, but show the intention of blocking out the mouth instead. These two gestures have 7 votes and 6 votes respectively. G5 and G6 also have 6 votes same as G4. G5 utilizes both hands instead of the head, showing the intention to wake up the HMD. G6 only needs the user's head, showing the intention to watch an imagined widget on the collar.

4 Study 2: Selecting the Gesture for CollarMic

We explored further into the user-designed gestures. Based on the outcome from Section 3, we conducted user interviews to evaluate the top gestures by participants. The interviews enabled participants to experience various trigger gestures from both speaker and listener perspectives, and evaluate the gestures based on different experiential dimensions. Finally, the best design is selected based on user evaluations.

4.1 Participants and Apparatus

This study recruits 26 students from the university to test and assess different interactive gestures. The participants consists of 11 males and 15 females, all between the ages of 18 and 35. A Quest 2 is utilized as the HMD for this experiment. Each participant was given $8 as the compensation.

4.2 Study Design

We utilized a one-factor within-subjects study with *gesture* as the only factor. For the factor gesture, we selected the 6 triggering gestures in Figure 1 to evaluate.

Besides, to better simulate real scenarios, we let participants form groups of two. The two participants played different roles: one as the speaker to act triggering gestures and speech voice queries; the other as the listener who observed and listened to the action and voice of the speaker. For each pair of speaker and listener, all of the 6 gestures would be performed in this study. A pair of speaker and listener would exchange their roles after a round of study, in order to avoid the effect of roles.

We provided participants with semi-open scenarios similar to Study 1. Specifically, as a representative, we suggested that participants could imagine themselves in the progress of a discussion in a business setting. We offered some examples for the speakers to simulate inquiries with VA, such as "what's the time", "what's the date", and "help me order my lunch".

The participants were requested to assess each gesture immediately after every task. The evaluation form is a 7-point Likert questionnaire, composed of 4 usability metrics evaluated by the speaker and 2 social effect metrics evaluated by the listener. The usability metrics included:

- Simplicity: "How simple was it to use this gesture?" (difficult - simple)
- Fatigue: "How fatigued did you feel after using this technique?" (fatigued - not fatigued)
- Speed: "Was the interactive speed of this gesture fast or slow?" (slow - fast)
- Embarassment: "How embarassed do you feel using this gesture in a social setting?" (embarassed - not embarassed)

The social effects metrics included:

- Confusion: "How confusing is it to judge the speaker's speech object?" (confusing - not confusing)
- Discomfort: "Is it comfortable to see this gesture in society?" (uncomfortable - comfortable)

We concerns both usability and effects on social in this evaluation design. The significance of our design is to make the evaluation solid in the conversation switching scenarios.

4.3 Procedure

We first grouped the participants in pairs consisting of a speaker and a listener. A study group consisted of two rounds, with the roles participants exchanged by the rounds. Each round consisted of 6 tasks with 6 different triggering gestures as task conditions. A study group had the same condition order derived from a Latin square.

For each task, the speaker wore the HMD and sat face-to-face with the listener. First, we let the participants be familiar with the device and all gestures as well as the scenario settings for 5–10 minutes. We observed that most participants were familiar with these actions after 2–3 repetitions. In subsequent tasks, the speaker performed the triggering gestures strictly following the descriptions while asking identical questions to the HMD. Each gesture was repeated at least twice. During the process of tasks, the listeners faced the speakers directly and observed their behavior. After each task, the participants needed to evaluate their experiences using the questionnaire. The experiment lasted approximately 20–40 minutes.

4.4 Results

We examined the evaluation outcomes of user interviews. During data processing, each participant's ratings of 6 dimensions for 6 gestures were transformed into a 36-dimensional vector. The reliability of these evaluation results was tested using Cronbach's α [6] method, yielding $\alpha = 0.90$. To analyze the evaluation results, we carried out and reported Friedman tests [18] and Wilcoxon signed rank tests [40]. We utilized effect size in the study represented by Hedges's g_s metric as reference [14], which states that g_s can be uniformly used as a benchmark across varied experiments. Visual charts were also used to compare scores for various gestures across different dimensions as depicted in Figure 2. The chart clearly shows that G1, G2, and G6 have higher scores than the rest.

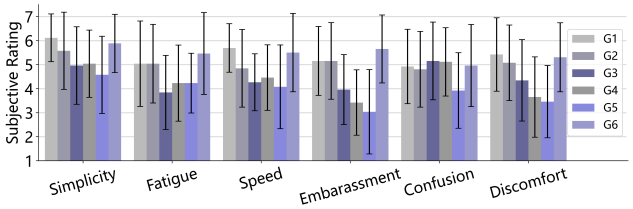


Figure 2: Average rating of the six gestures (7: most positive, 1: most negative). Error bar indicated one standard deviation.

4.4.1 Subjective Ratings

Embarassment & Discomfort The Friedman tests show that the two dimensions related to sociability are the most significant among the evaluation dimensions, which are embarrassment and discomfort.

For embarrassment, the Friedman test shows the most significance of all dimensions ($\chi^2(5) = 50.9, p < .001$). Wilcoxon tests show that G6 is merely over G1 ($W = 49.5, p = .11, g_s = 0.35$) as well as G2 ($W = 36.5, p = .10, g_s = 0.33$). Despite the significance, G6 has the best score in embarrassment.

The discomfort has the second most significance ($\chi^2(5) = 41.4, p < .001$). In this dimension, G1 has a weak advantage of G6 ($W = 61.0, p = .74, g_s = 0.08$), but not significant.

Other Dimensions For dimensions including simplicity, speed, fatigue, and confusion, we found these dimensions have a significant effect on the ratings (simplicity: ($\chi^2(5) = 34.4, p < .001$); speed: ($\chi^2(5) = 34.0, p < .001$); fatigue: ($\chi^2(5) = 25.5, p < .001$); confusion: ($\chi^2(5) = 13.8, p = .05$)).

For pair-wise comparisons, we first inspected in simplicity and found G1 is not better than G6 ($W = 30.0, p = .48, g_s = 0.21$). The

comparison between G1 and G2 is yet not significant ($W = 25.0, p = .08, g_s = 0.40$).

In terms of speed, our analysis indicates that G1 is quite similar to G6 ($W = 110, p = .85, g_s = 0.14$), and both of them are higher than the others.

The fatigue shows G6 is not better than G1 ($W = 66.0, p = .40, g_s = 0.24$) or G2 ($W = 83.0, p = .42, g_s = 0.25$).

With inspection in confusion, all gestures received competitive ratings except G5, with G3 ¿ G4 ¿ G6 ¿ G1 ¿ G2. There are no significant difference among the five gestures, even for G3 and G2 ($W = 101, p = .39, g_s = 0.21$).

4.4.2 Discussion

The findings indicate that G1 and G6 are the top-rated gestures. To further explore the advantages and disadvantages of these two gestures, we reconnected with some participants for casual interviews. Three users pointed out that G1 requires an additional hand compared to G6, which tends to cause fatigue and increase the burden of use. "Touching on HMD shell feels quite unsanitary," one participant criticized. While another participant supported the opposite, "I'm thrilled by this touch-to-speech design, it's truly cool." With the integration of the perspectives of these interviewees, we found that support for G1 is associated with trendiness and maintaining a steady head position, while backing for G6 is linked to not being exposed, cleanliness and free usage of hands. Taking all these factors into account, we propose that G6 is better suited to our technique.

5 Implementation of CollarMic

The design of CollarMic is originated from the brainstorming in Study 1. The idea of CollarMic is to rotate head away to indicate the intention to switch the speech object. Some similar ideas of leverage head based gestures are also shown to be usable, like HeadGesture [44]. However, CollarMic is advanced to the prior work because we used this design to handle the social dilemma in efficiently talking to VAs in a socially acceptable way. Our subsequent usability study also proved CollarMic is functional and better than the keyword spotting techniques.

5.1 Algorithm Design

5.1.1 Algorithm Flow

The structure of our system is outlined as follows. Initially, the HMD sensors capture the user's head movements and voice through its built-in IMU and microphone. Within the HMD, a trigger detection model determines if a triggering gesture has been made by the user, relaying this information back to the system. If triggered, specific speech data from users would be sent to the voice assistant. Our demonstrations utilized a voice command recognition system to determine if users have issued pre-set voice commands. If such command is identified, the corresponding routines will be executed with particular responses returned and displayed for the users.

5.1.2 Triggering Detection Algorithm

We proposed Gaussian Hidden Markov Models (GHMM), a Markov recurrent neural network (RNN) based on Gaussian Mixture Models (GMM) to discriminate head movement. The GMM outputs the logarithmic probability referring to the construction of deep Gaussian Mixture Model (DGMM) [12], then it applied the probability to the time series to construct the RNN recognition model.

The formal description of GHMM is as follows. For continuous motion on a 2D plane, its i^{th} type of two-dimensional features (X_i, Y_i) are represented as continuous sequences $\left(X_{i,t}, Y_{i,t}\right)$, where i signifies the feature category, and t represents the frame. For data $\boldsymbol{d_t} = \left\{\left(X_{i,t}, Y_{i,t}\right), i = 1 \cdots P\right\}$ at frame t, the k^{th} Gaussian model $M^{(k)}$ provides the logarithm of relative probability $M^{(k)}(\boldsymbol{d_t})$ as follows.

$$\exp\left(\sum_{i=1}^{P} b^{(k)} - \frac{\left(\frac{X_i-\mu_x^{(k)}}{\sigma_x^{(k)}}\right)^2 + \left(\frac{Y_i-\mu_y^{(k)}}{\sigma_y^{(k)}}\right)^2 - \frac{2\rho_{xy}^{(k)}\left(X_i-\mu_x^{(k)}\right)\left(Y_i-\mu_y^{(k)}\right)}{\sigma_x^{(k)}\sigma_y^{(k)}}}{2\left(1-{\rho_{xy}^{(k)}}^2\right)}\right) \quad (2)$$

Where $b^{(k)}$ represents the probability normalization coefficient of the Gaussian model, defined as below.

$$b^{(k)} = -\log(2\pi) - \log\left(\sigma_x^{(k)}\sigma_y^{(k)}\right) - \frac{1}{2}\log\left(1-{\rho_{xy}^{(k)}}^2\right) \quad (3)$$

Given relative probability is used, the constant term of $b^{(k)}$ in the above equation can be disregarded.

According to the above, it yields that the relative joint probability distribution of N Gaussian models is $\left\{M^{(k)}, k = 1 \cdots N\right\}$. Assume only K probabilities associated with positive cases are concerned, this distribution can be condensed to $(\exp(p_1), \cdots, \exp(p_K), 1)$ with minimal loss of information. In GHMM, internal states of Markov models are initialized as $(s_1 = p_1, \cdots, s_K = p_K)$. The state or probability transition equation from frame $t-1$ to t is defined as:

$$s_{k,t} = \max\left\{\max_{j=1\cdots K}\left\{s_{j,t-1} + d_{jk}\right\}, 0\right\} + p_{k,t} \quad (4)$$

Further discussions are required to clarify the transition equation. The addition operations are performed on the logarithmic field, which is equivalent to probability multiplication. The GHMM adopts $d_{jk} < 0$ as the decay of state, to avoid infinite growth of any part of the state. Due to the logarithmic nature, the transition equation of Markov chains does not involve typical matrix multiplication. Instead, it directly selects the maximum value and disregards insignificant terms. Moreover, this design results in a smoother gradient during RNN optimization because each coefficient of $s_{k,t}$ is 1, meaning every time frame on the motion sequence has identical coefficients for gradients too. Lastly, this model uses a single-layer linear network and *sigmoid* function to obtain output probabilities for each frame t using $\left\{s_{k,t}, k = 1 \cdots K\right\}$. In this case, $sigmoid(p)$ signifies the left probability of distribution $\exp(p) : 1$.

5.1.3 Recovery of Triggers

To address the issue of interrupting voice commands during real use, we proposed to set the head's returning to the front as the the ending of participants' signal. During the acting of users' gestures in Study 2, we found rotating the head back by $\frac{1}{3}$ was a reliable criterion for determining the completion of a voice command.

5.2 Data Collection

5.2.1 Study of Data Collection

We conducted the simulation study through collecting users' data to access the system's recognition preformance.

Participants and Apparatus We recruited 9 males and 6 females from the campus. We utilized PICO4 as the mixed reality device and used its motion sensor's data. For development, we used Unity Editor 2022.3.4 and PICO Integration 2.3 toolkit.

Experiment Design The participants in this experiment completed three tasks in the order derived from a Latin square. The tasks of looking around and chatting lasted for 5 minutes each. Both tasks asked the participant to keep walking or sitting and have a free look at their will. The task of looking around did not force users to control their head behavior as they kept moving, and the chatting task let the participants freely talk and express their modes with any body parts they wanted. For the task of performing trigger gestures,

users pressed buttons to start tasks. Task variables appeared in front of them only after pressing start button so as to prevent premature action from being taken by users. After completing an action, it needed to press the button again to end the task.

We collected positive and negative samples respectively: positive samples are from users who actively make the triggering gesture, while negative samples are from users who perform natural head behaviours.

For positive samples, we set 10 voice commands and 3 initial orientations: looking up, looking down, and looking forward, resulting in 30 combinations. The order of condition is first selecting the voice command, then applying the three initial orientations in the same order for each command. The reason for constantly changing the initial orientation is to reduce repetition action, resulting in more natural actions performed by users. The process of making triggering gestures will be repeated twice: first as warm-up exercises where all data will be discarded; secondly as formal collection requiring natural actions from users. For each task in data collection, the experiment collects HMD's motion trajectory data.

For negative samples, we let users freely walk and look around in a room, as well as sit down and chat with the experimenter.

Both task asks the participant to keep walking or sitting and have a free look at their will. The task of looking around does not force the user to control their head orientation as they keep moving. The task of chatting needs the participant to freely talk and express their modes with any body parts they will. For both of the tasks, the experimental scene will automatically count the time and interrupt the experiment and generate records when the time is up.

For the task of performing trigger gestures, users press buttons on their hands to start tasks; task variables appear in front of them only after pressing the start button so as to prevent premature action from being taken by users. After completing an action, it is also necessary to press the button again to end the task.

Procedure The participants in this experiment completed the 3 tasks in the order derived from a Latin square. The tasks of looking around and chatting lasted for 5 minutes each. While for the task of triggering, they need to repeatedly do each triggering gesture with each combination in a sequence for 1 time, resulting in 30 triggering gestures.

5.2.2 Data Processing

We collected data including (15 participants $\times$5 minutes=) 75 minutes of walking data, (15 participants $\times$5 minutes=) 75 minutes of chatting data, and (15 participants $\times$30 gestures =) 450 cases of triggering gesture. The motion data is converted into position sequences at a rate of 20 frames per second using linear interpolation. For the data of triggering gestures, any portion before the user starts moving is eliminated by extending the motion direction backwards to the furthest point and discarding earlier data. The frame numbers of all segmented gesture data range from $[22, 136]$, with an average frame number of 61 (about 3.05 seconds). For natural head movement data, we segmented the data into 10 seconds' segmentation (200 frames) using sliding window algorithm. Consequently, we segmented triggering gesture, walking and chatting each into 450 cases, totally 1350 in number.

5.3 Implementation

5.3.1 Training with GHMM

Features In order to achieve trigger action detection, we have attempted different features. Finally, the Euler angles of rotation (X and Y components) of HMD are used as the first feature, and the average angular velocity of these two components in the past 0.75 seconds is used as the second feature.

Environment Settings In Python 3.10.11 environment, we implemented GHMM's supervised learning algorithm using PyTorch 1.12.1 framework. The algorithm uses AdamOptimizer with a learning rate of 10^{-5} as an optimizer, and BCELoss (binary cross entropy) as the optimized loss function. The function BCEWithLogitsLoss is used in the implementation in place of *sigmoid* and BCELoss.

Training Result Using the collected dataset, The model has achieved a stable accuracy after about one thousand training iterations. Continuous training is conducted to minimize loss, eventually surpassing twenty thousand training rounds. The model's final accuracy on its training set is 98.00%.

5.3.2 Deployment

In order to achieve real-time motion data detection, we used C# script to implement a real-time decision system based on GHMM in the Unity development environment, and migrated the parameters of the model. Through testing the transplanted algorithm on the constructed dataset, the results show that the detection results and triggering time of the transplanted algorithm are consistent with the original algorithm for each data case, and the difference in scores between the two algorithms' model outputs is less than 10^{-10}.

5.4 Recognition Performance Simulation

5.4.1 Simulation Design

We compare the proposed GHMM model with the commonly utilized SVM in gesture recognition on the constructed data. Since SVM requires fixed-length feature vectors for input, this experiment standardizes the length of data vectors to 200 frames. For positive cases ranging from $[22, 136]$ frames, the data is randomly inserted into various positions within the vector and padded with first and last frames, which implies that there's no movement before or after a motion sequence. As for negative cases originally composed of 200 frames,we shorten each negative case to align with any positive case's length, ensuring their effective lengths match those of positive cases. Then similar techniques are used to pad them to 200 frames. This is because we find that the SVMs tends to overfit the context—it mainly predicts the labels based on the length of effective information. This also suggests that SVM is not reliable enough for this task. As mentioned above, the provided features include two-dimensional rotation and angular velocities. It means each frame consists of four dimensions, and every case vector has a length of 800.

We further enhance the dataset by data augmentation. The data augmentation method involves converting each initial short data into 10 cases with varying insertion offsets, with each case consisting of 200 frames. Subsequently, to verify that the model's recognition of triggering cases relies exclusively on trigger actions and not recovery actions, we implement variable control on the set. The data from positive trigger cases are truncated at the motion extreme, eliminating any following recovery actions, and replacing the truncated section to initial positions. Only positive cases are modified in the controlled dataset, with their angular velocities recalculated for these modified cases.

The settings of the simulation are as follows. GHMM has only been trained on the original data and has not learned from the controlled data, so the dataset is used to test GHMM's performance. Two types of tests are conducted for SVM: one uses the dataset directly as a training set, while another performs 10-fold cross-validation on this same dataset.

5.4.2 Result

The simulation results are displayed in Table 3. We present the results as confusion matrixes, with rows represent the ground truth (T) and columns indicating the predicted labels (T). Furthermore, we compute accuracy, precision, recall, and F1 score based on the confusion matrix.

Table 3 illustrates the test result on the dataset that includes recovery action control. According to the result, GHMM has an accuracy rate of 98.0%. This indicates that GHMM maintains stable recognition capabilities for triggering actions and its effectiveness is minimally impacted by other conditions. The false triggering rate of GHMM can be estimated on the negative cases, which is 0.13 times per minute. In contrast to GHMM's stability, SVM's training accuracy on the controlled dataset is only 87.9%. Additionally, SVM's 10-fold validation accuracy dropped to 77.5%, representing a decline of 10.4%. The result shows SVM's deficiency of overfitting on training set. When comparing these outcomes, it indicates that SVM's capacity to fit triggering actions significantly lags behind GHMM's ability.

Table 3: Simulation results on the controlled dataset where restoration actions are reset. Neg denoted the negative samples, Pos denoted the positive samples. T denoted the Ground Truth, P denoted the predicted result. All metrics were calculated as macro values (macro-precision, macro-recall, macro-F1).

(a) Confusion Matrix

Model	GHMM (test)		SVM (train)		SVM (10-fold)	
T \ P	Neg	Pos	Neg	Pos	Neg	Pos
Neg	8,810	190	7,776	1,224	6,415	2,585
Pos	81	4,419	408	4,092	450	4,050

(b) Table of Metrics

Metric	GHMM (test)	SVM (train)	SVM (10-fold)
Accuracy	98.0%	87.9%	77.5%
Precision	95.9%	77.0%	61.0%
Recall	98.2%	90.9%	90.0%
F1	97.0%	83.4%	72.7%

6 Study 3: Usability Evaluation in Real Tasks

So far, we have selected speak-to-shoulder (G6) as the final gesture and designed the recognition algorithm based on GHMM. To further explore the usability in real tasks, we conducted this experiment. We aimed to compare with mainstream voice awaking techniques in a semi wizard of oz way.

6.1 Participants and Apparatus

This study involves 24 participants from the university campus, including 11 males and 13 females. The HMD device used in this study is PICO 4. We also utilized a XiaoDu smart speaker for comparison purposes in keyword spotting. This device provides a keyword spotting function ("Xiaodu Xiaodu") that can display a notice on the smartphone. The notice styles are customizable, with adjustable alarm sounds set as a response word. When this response word shows up, the experimenter will manually activate the VA. The smart speaker has a cylindrical form and its dimensions do not exceed 8 cm in any direction. Each participant received $14 as a compensation for their time and participation.

The complete experimental setup includes an HMD and a PC server. Unity Editor 2022.3.6 serves as the development platform for the experimental scenarios, while PICO Integration 2.3.4 is the toolkit utilized for development purposes. The PC server operates in Python's runtime environment version 3.11.4, with offline voice command recognition carried out using pocketsphinx version 5.0.2 behind it. Additionally, a GUI based on tkinter version 8.6 is developed within the server to facilitate real-time playback of user's voice commands, animating user's vocal operations, and manual triggering of operations when necessary.

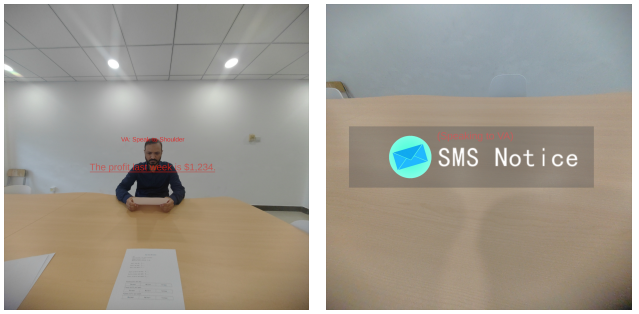
6.2 Study Design

We utilized a two-factor within-subjects design with *technique* and *scenarios* as two within factors. For technique, we compared our head movement based voice switching technology with keyword spotting based voice assistant activation technology. While for the keyword spotting based voice assistant activation technology, the user needed to say "Xiaodu, Xiaodu" as an activation phrase to activate the voice assistant. For scenarios, we adopted business and chatting scenarios as two different common scenarios for usability study.

For business scenario, the role assignment is for the speaker to play the role of an operator using MR for work, and for the real-life role to play the leader of the operator. The task setting in the business scenario requires that the speaker and the leader sit face-to-face across a table, where the speaker needs to use the voice assistant deployed in the MR to inquire about some business information (e.g. ask about company profits). After filling out the forms printed on the table, they needed to report the business data to the leader, which was played by another participant. The leader needed to give instructions based on the information and ask the operator to complete the form through the voice assistant. In this type of scenario, speakers will pay more attention to their words and actions, and need to interact with objects in reality, so they will pay more attention to how their interactions affect reality and society, which simulates the difficult situations within the scope of using MR. Figure 3a illustrates an HMD screenshot in business scenaro.

For chatting scenario, one participant was a speaker and the other participant needed to act as the classmate and friend of the former participant. They needed to sit face-to-face across a table and collaborate on completing a common shopping task, such as query the price and make a purchase. Because of the online shopping context, operations were simplified and the speaker only knew about the names and prices of the items without needing to focus on the physical content. In this type of scenario, the speaker paid more attention to the experience brought by voice interaction.

In order to increase the integrity of interaction contents, the study not only designed operations that require users to subjectively initiation by voice interactions but also included some unexpected situations that must be immediately dealt with by voice interactions. For all seven experimental scenarios, including warm-up scenes, there will be sudden interruptions in the MR scene, including incoming calls, received SMS messages, and received social chat messages (as Figure 3b). In the experimental scenario, a truncated Poisson distribution is used to generate unexpected information with an interval between 30 seconds and 90 seconds, with a mean of 45 seconds.



(a) visual response from VA in the business scenario (b) responding to the SMS in the chatting scenario

Figure 3: The experiment platform in Study 3.

We have created 7 experiment settings, one for warming up and three each for business and chatting experimental scenarios. Each of the 6 experimental scenes comes with a corresponded task guide that

outlines specific procedures. These procedures involve data query and reporting tasks. The voice assistant's backend is powered by voice command recognition, paired with a GUI to animate the voice commands. The GUI includes multiple UI buttons capable of initiating all pre-set commands. There are 13 potential voice commands for every business scenario (for example, staff inquiries, financial inquiries), between 20–21 for each chatting scenario (price inquiries, making purchases, questioning about purchase lists), along with five universal commands for general operations such as interruption handling (dismiss disruptive notification, directive of failure).

6.3 Procedure

The experiment contained 2 stages, differed in the role of the participant. Each participant would act as the speaker and the listener respectively in two stages using a counter-balanced order. A break of 5 minutes was enforced between 2 stages. Before the experiment, we introduced the experiment end and gave the participants 5 minutes to become familiar with the device as well as the triggering techniques. For each stage, the participant needed to take 2 rounds of tasks differed in *technique* in Latin Square counter-balanced order. The order of the tasks was fixed with slightly increasing difficulty. After completing each task, the speaker and the listener needed to evaluate their experience of using the corresponding triggering gestures through a 7-point Likert questionnaire. The experiment took a total of 50 minutes.

6.4 Results

We analyzed the subjective evaluation results of the subjects in Study 3. Study 3 uses the same subjective evaluation method as Study 2. Cronbach's α is tested on the data with 24 subjects and 16 items, with $\alpha = 0.75$. The α value indicates that the reliability of the results is acceptable. The evaluation dimensions in Study 3 are analyzed by Friedman tests and Wilcoxon signed rank tests as same as in Study 2. We also visualized overall ratings for different dimensions and gestures in Figure 4. From the chart, it can be seen that G1 have much higher scores than G2.

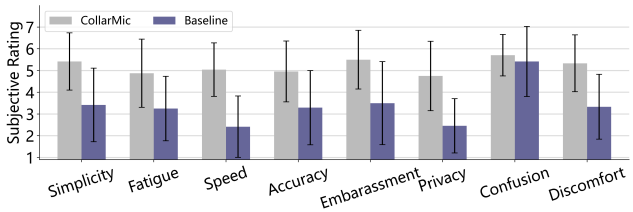


Figure 4: Average rating of the two techniques (7: most positive, 1: most negative). Error bar indicated one standard deviation.

6.4.1 Subjective Ratings

Figure 4 showed the evaluation results concerning different dimensions. In the dimensions of speed, privacy, discomfort, simplicity, embarassment, accuracy and fatigue, significance was found on the ratings (speed: $(\chi^2(1) = 19.2, p < .001)$; privacy: $(\chi^2(1) = 18.0, p < .001)$; discomfort: $(\chi^2(1) = 17.2, p < .001)$; simplicity: $(\chi^2(1) = 12.8, p < .001)$; embarassment: $(\chi^2(1) = 8.89, p = .01)$; accuracy: $(\chi^2(1) = 8.89, p = .01)$; fatigue: $(\chi^2(1) = 7.20, p = .01)$). However, confusion has no significant effect on the ratings $((\chi^2(1) = 0.25, p = .62))$. These showed that our technique was preferred by participants during the usage.

6.4.2 Qualitative Feedback

The majority (20/24) of participants approve of CollarMic, because they were accustomed to specific ways of speaking with VAs on their smart devices. They embrace this unique operational style because it aids them in adopting the preferred speech pattern.

A small group of participants (3/24) maintain that no matter what method of interaction is used, they possess enough skill to distinguish the speech objects of users wearing HMDs. However, this leads them to pay attention to all dialogues, and they find CollarMic beneficial as it assists in screening out irrelevant talkings.

Several users (7/24) have expressed dissatisfaction regarding the low efficiency when uttering trigger keywords during a conversation using the keyword spotting technique. This technique often fails to respond and requires users to distinctly pronounce the trigger words, leading to an disappointing user experience.

7 Discussion

7.1 Feasibility of CollarMic

CollarMic provides a simple function to switch the user's speech object among surrounding scenes and VAs on the HMD. The design of CollarMic originated from the brainstormed idea in Study 1 that head orientation can be utilized to indicate speech intention. Moreover, in our exploration of interaction spaces with VAs, we found most of the functions involved the body parts that are not connected with HMD (e.g. hand or foot). However, we found the design of head initiated interaction is widely approved in user interviews in Study 2, due to its unexposed and hands-free characteristics.

Based on the above findings, we proposed CollarMic as a universal interaction technique with VAs based on head movement. We carried out a simulation experiment on the recognition model, which is named as Gaussian Hidden Markov Model (GHMM). Simulation results showed our model achieved 98.0%, which is 20.5% better than SVM in cross validation. On the data of natural head behaviors, our model had a false triggering rate of 0.13 times per minute.

In Study 3, a usability study in business and chatting scenarios in MR is carried out and CollarMic showed a significant advantage in various dimensions over common keyword spotting (KWS) techniques. Our better dimensions include speed, privacy, simplicity, embarrassment, accuracy, and fatigue from the speakers' perspective. We also evaluated CollarMic from the listeners' perspective. Results showed that CollarMic has a similar capability to keyword spotting in terms of confusion, however, we have better performance in discomfort. It's noticeable that KWS based techniques have expended a great cost to reduce the confusion in terms of intention in the acoustic channel. Instead, CollarMic is more feasible in contrast to KWS at a lower cost based on the reduction of calling out the keyword, which is more preferred by the users. According to the results, CollarMic has dominant advantages in speed and privacy than KWS. These results also proved that CollarMic is a usable tool to support common interactions with VAs on the smart HMD, and could improve user experience as well as social effects in real use scenarios.

7.2 Design Implications

CollarMic demonstrated the feasibility of leveraging head movement with HMD. Compared to other hand-based gestures in VR/MR interaction, the head movement is easier to be captured as most HMD provides a stable head position tracking feature. We also utilized the characteristics of head movement to convey user intentions in social scenarios, allowing easier social interaction.

We have observed the time duration of performing the CollarMic gesture in our simulation and found the average time is 0.954s (std: 0.598s). Therefore, we suggested that head gesture is more efficient than acoustic gestures. We also found the short duration of head actions is comparable to a number of VR/MR gestures that need to be watched for visual feedback with careful control. In a study of text entry for HMD [46], head gestures are proved to outperform head-assisted tap interactions in efficiency, while the former involves a single channel and the latter uses mixed channels. These statements

show that mono-channel gestures are preferred by users in similar tasks such as trigger or response.

As many head gestures are commonly utilized in real scenarios (e.g. nodding or shaking the head), we also recommend that a set of head-based gestures be applied to HMD for usability through high intuition. Though hand interactions are attractive to people, some particular scenarios require minimal gestures by heads in demand of better privacy and social characteristics.

8 Limitation and Future Work

We now summarize the limitations of this work, which we also see as opportunities for future work.

First, the algorithm of CollarMic only leveraged the features of the IMU data from the HMD during head rotation. It is also possible to incorporate features from auxiliary channels (e.g., acoustic features [25, 45]) to further improve its accuracy, and to test the generalizability of the algorithm on other devices (e.g., earbuds), which may further improve the application scenario of the proposed technique.

Second, in different cultural backgrounds, people may have different perceptions of body language (e.g., head rotation). In this paper, we only involved Chinese participants to ensure the internal validity of the results. It would be worthwhile to further validate the feasibility of CollarMic with more diversed participants (e.g., from western countries).

9 Conclusion

Voice assistants (VAs) are becoming increasingly prevalent in Virtual or Augmented Reality. However, with the increasing number of users in VR or AR, speaking to the VA in social scenarios may lead to misunderstanding and embarrassment. Aiming at triggering the VA in a more efficient and natural way, we proposed CollarMic, a technique leveraging speak-to-shoulder gestures for users to switch conversational targets to VA. Through a brainstorming with 10 experts, we first collected a total of 62 voice input triggering gestures in social that leveraged different body parts (e.g., head and hand), reaching 6 candidate gestures through voting. We then asked another 26 participants to act and rate them from different perspectives. The speak-to-shoulder gesture was selected for CollarMic according to the result. To achieve efficient and accurate triggering recognition, we utilized Gaussian Hidden Markov Model (GHMM). Based on the collected data from 15 real users, we reached an accuracy of 98.0%, and a false triggering rate of 0.13 times per minute, which was higher than the SVM-based algorithm. In the usability evaluation study with two different social scenarios (business and chatting), CollarMic was significantly more preferred by the users than keyword spotting, in terms of interaction speed and social acceptance.

References

[1] N. Aburumman, M. Gillies, J. A. Ward, and A. F. d. C. Hamilton. Nonverbal communication in virtual reality: Nodding as a social signal in virtual interactions. *International Journal of Human-Computer Studies*, 164:102819, 2022.

[2] M.-L. Bourguet and A. Ando. Synchronization of speech and hand gestures during multimodal human-computer interaction. In *CHI 98 Conference Summary on Human Factors in Computing Systems*, pp. 241–242, 1998.

[3] S. Carbini, L. Delphin-Poulat, L. Perron, and J.-E. Viallet. From a wizard of oz experiment to a real time speech and gesture multimodal interface. *Signal Processing*, 86(12):3559–3577, 2006.

[4] A. P. Chaves and M. A. Gerosa. Single or multiple conversational agents? an interactional coherence comparison. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2018.

[5] J. Diemer, G. W. Alpers, H. M. Peperkorn, Y. Shiban, and A. Mühlberger. The impact of perception and presence on emotional reactions: a review of research in virtual reality. *Frontiers in psychology*, 6:26, 2015.

[6] L. S. Feldt, D. J. Woodruff, and F. A. Salih. Statistical Inference for Coefficient Alpha. *Applied Psychological Measurement*, 11(1):93–103, 1987. doi: 10.1177/014662168701100107

[7] F. Ge and Y. Yan. Deep neural network based wake-up-word speech recognition with two-stage detection. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2761–2765. IEEE, 2017.

[8] J. Grubert, T. Langlotz, S. Zollmann, and H. Regenbrecht. Towards pervasive augmented reality: Context-awareness in augmented reality. *IEEE transactions on visualization and computer graphics*, 23(6):1706–1724, 2016.

[9] A. G. Hauptmann. Speech and gestures for graphic image manipulation. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 241–245, 1989.

[10] S. Irawati, S. Green, M. Billinghurst, A. Duenser, and H. Ko. An evaluation of an augmented reality multimodal interface using speech and paddle gestures. In *Advances in Artificial Reality and Tele-Existence: 16th International Conference on Artificial Reality and Telexistence, ICAT 2006, Hangzhou, China, November 29-December 1, 2006. Proceedings*, pp. 272–283. Springer, 2006.

[11] S. Khan and B. Tunçer. Gesture and speech elicitation for 3d cad modeling in conceptual design. *Automation in Construction*, 106:102847, 2019.

[12] Y. Kong, B. Xu, B. Zhao, and J. Qi. Deep Gaussian Mixture Model on Multiple Interpretable Features of Fetal Heart Rate for Pregnancy Wellness. In K. Karlapalem, H. Cheng, N. Ramakrishnan, R. K. Agrawal, P. K. Reddy, J. Srivastava, and T. Chakraborty, eds., *Advances in Knowledge Discovery and Data Mining*, pp. 238–250. Springer International Publishing, Cham, 2021.

[13] K. Kumatani, S. Panchapagesan, M. Wu, M. Kim, N. Strom, G. Tiwari, and A. Mandai. Direct modeling of raw audio with dnns for wake word detection. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 252–257. IEEE, 2017.

[14] D. Lakens. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4, 2013. doi: 10.3389/fpsyg.2013.00863

[15] M. Lee and M. Billinghurst. A wizard of oz study for an ar multimodal interface. In *Proceedings of the 10th international conference on Multimodal interfaces*, pp. 249–256, 2008.

[16] M. Lee, M. Billinghurst, W. Baek, R. Green, and W. Woo. A usability study of multimodal input in an augmented reality environment. *Virtual Reality*, 17:293–305, 2013.

[17] D. Maloney, G. Freeman, and D. Y. Wohn. " talking without a voice" understanding non-verbal communication in social virtual reality. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–25, 2020.

[18] M. Marozzi. Testing for concordance between several criteria. *Journal of Statistical Computation and Simulation*, 84(9):1843–1850, 2014. doi: 10.1080/00949655.2013.766189

[19] D. McMillan, B. Brown, I. Kawaguchi, R. Jaber, J. Solsona Belenguer, and H. Kuzuoka. Designing with gaze: Tama–a gaze activated smart-speaker. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–26, 2019.

[20] C. Mignot, C. Valot, and N. Carbonell. An experimental study of future "natural" multimodal human-computer interaction. In *INTERACT'93 and CHI'93 Conference Companion on Human Factors in Computing Systems*, pp. 67–68, 1993.

[21] M. R. Morris. Web on the wall: insights from a multimodal interaction elicitation study. In *Proceedings of the 2012 ACM international conference on Interactive tabletops and surfaces*, pp. 95–104, 2012.

[22] L. Pandey, K. Hasan, and A. S. Arif. Acceptability of speech and silent speech input methods in private and public. *Conference on Human Factors in Computing Systems - Proceedings*, 2021. doi: 10.1145/3411764.3445430

[23] T. Piumsomboon, D. Altimira, H. Kim, A. Clark, G. Lee, and M. Billinghurst. Grasp-shell vs gesture-speech: A comparison of direct and indirect natural interaction techniques in augmented reality. In *2014 IEEE International Symposium on Mixed and Augmented*

Reality (ISMAR), pp. 73–82. IEEE, 2014.

[24] Y. Qin, C. Yu, Z. Li, M. Zhong, Y. Yan, and Y. Shi. Proximic: Convenient voice activation via close-to-mic speech detected by a single microphone. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2021.

[25] Y. Qin, C. Yu, Z. Li, M. Zhong, Y. Yan, and Y. Shi. ProxiMic: Convenient Voice Activation via Close-to-Mic Speech Detected by a Single Microphone. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2021.

[26] J. Ramırez, J. C. Segura, C. Benıtez, A. De La Torre, and A. Rubio. Efficient voice activity detection algorithms using long-term speech information. *Speech communication*, 42(3-4):271–287, 2004.

[27] J. Rasch, V. D. Rusakov, M. Schmitz, and F. Müller. Going, going, gone: Exploring intention communication for multi-user locomotion in virtual reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2023.

[28] S. Robbe. An empirical study of speech and gesture interaction: Toward the definition of ergonomic design guidelines. In *CHI 98 Conference Summary on Human Factors in Computing Systems*, pp. 349–350, 1998.

[29] F. Roider, L. Reisig, and T. Gross. Just look: The benefits of gaze-activated voice input in the car. In *Adjunct Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pp. 210–214, 2018.

[30] D. Roth, C. Klelnbeck, T. Feigl, C. Mutschler, and M. E. Latoschik. Beyond replication: Augmenting social behaviors in multi-user virtual realities. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 215–222. IEEE, 2018.

[31] D. Roth, M. E. Latoschik, K. Vogeley, and G. Bente. Hybrid avatar-agent technology–a conceptual step towards mediated "social" virtual reality and its respective challenges. *i-com*, 14(2):107–114, 2015.

[32] S. Samrose, D. McDuff, R. Sim, J. Suh, K. Rowan, J. Hernandez, S. Rintel, K. Moynihan, and M. Czerwinski. Meetingcoach: An intelligent dashboard for supporting effective & inclusive meetings. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2021.

[33] P. E. Shrout and J. L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological bulletin*, 86(2):420–428, 1979. doi: 10.1037/0033-2909.86.2.420

[34] S. Sigtia, R. Haynes, H. Richards, E. Marchi, and J. Bridle. Efficient voice trigger detection for low resource hardware. In *Interspeech*, pp. 2092–2096, 2018.

[35] M. H. A. Soliman. *Brainstorming for Problems Solving: How Leaders Can Achieve a Successful Brainstorming Session*. Number January. Mohammed Hamed Ahmed Soliman, 2020. doi: 10.5281/zenodo.4270238

[36] S. G. Tanyer and H. Ozer. Voice activity detection in nonstationary noise. *IEEE Transactions on speech and audio processing*, 8(4):478–482, 2000.

[37] B. Tarr, M. Slater, and E. Cohen. Synchrony and social connection in immersive virtual reality. *Scientific reports*, 8(1):3693, 2018.

[38] B. Tarr, M. Slater, and E. Cohen. Synchrony and social connection in immersive virtual reality. *Scientific reports*, 8(1):3693, 2018.

[39] Y. Weng, C. Yu, Y. Shi, Y. Zhao, Y. Yan, and Y. Shi. Facesight: Enabling hand-to-face gesture interaction on ar glasses with a downward-facing camera vision. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2021.

[40] F. Wilcoxon. *Individual Comparisons by Ranking Methods*, pp. 196–202. Springer New York, New York, NY, 1992. doi: 10.1007/978-1-4612-4380-9_16

[41] J. O. Wobbrock, M. R. Morris, and A. D. Wilson. User-defined Gestures for Surface Computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pp. 1083–1092. ACM, New York, NY, USA, 2009. doi: 10.1145/1518701.1518866

[42] Y. Yan, H. Liu, Y. Shi, J. Wang, R. Guo, Z. Li, X. Xu, C. Yu, Y. Wang, and Y. Shi. Conespeech: Exploring directional speech interaction for multi-person remote communication in virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2647–2657, 2023.

[43] Y. Yan, C. Yu, Y. Shi, and M. Xie. PrivateTalk: Activating voice input with hand-on-mouth gesture detected by bluetooth earphones. In *UIST 2019 - Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, pp. 1013–1020, 2019. doi: 10.1145/3332165.3347950

[44] Y. Yan, C. Yu, X. Yi, and Y. Shi. HeadGesture: Hands-Free Input Approach Leveraging Head Movements for HMD Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(4), 2018. doi: 10.1145/3287076

[45] Z. Yang, C. Yu, F. Zheng, and Y. Shi. Proxitalk: Activate speech input by bringing smartphone to the mouth. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):1–25, 2019.

[46] C. Yu, Y. Gu, Z. Yang, X. Yi, H. Luo, and Y. Shi. Tap, dwell or gesture? exploring head-based text entry techniques for hmds. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 4479–4488, 2017.

[47] S. Zhao, B. Westing, S. Scully, H. Nieto, R. Holenstein, M. Jeong, K. Sridhar, B. Newendorp, M. Bastian, S. Raman, et al. Raise to speak: An accurate, low-power detector for activating voice assistants on smartwatches. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2736–2744, 2019.